Quantum Mechanics for Engineers

Leon van Dommelen

08/26/18 Version 5.63 alpha

Copyright and Disclaimer

Copyright © 2004, 2007, 2008, 2010, 2011, and on, Leon van Dommelen. You are allowed to copy and/or print out this work for your personal use. However, do not distribute any parts of this work to others or post it publicly without written permission of the author. Instead please link to this work. I want to be able to correct errors and improve explanations and actually have them corrected and improved. Every attempt is made to make links as permanent as possible.

Readers are advised that text, figures, and data may be inadvertently inaccurate. All responsibility for the use of any of the material in this book rests with the reader. (I took these sentences straight out of a printed commercial book. However, in this web book, I do try to correct "inaccuracies," OK, blunders, pretty quickly if pointed out to me by helpful readers, or if I happen to think twice.)

As far as search engines are concerned, conversions to html of the pdf version of this document are stupid, since there is a much better native html version already available. So try not to do it.

Reading the PDF

Some notes on reading the PDF version of this book:

- Without changing your Preferences/Settings, links to the web might not work on recent pdf readers. That is especially likely if you download the PDF instead of reading it in a web browser. What I had to do with Adobe Acrobat 9 on linux was to go into menu "Edit," "Preferences," "Trust Manager," "Internet Access from PDF files outside the web browser," "Change settings." There I had to enable the opening of all web pages. Specifying "Ask me" did *not* work; it just refused to open the link instead of asking.

- You may want to activate a [back button]. In my linux version of Adobe Acrobat reader, that is done using "Menu," "View," "Toolbars," "More Tools...," "Page Navigation Toolbar." There I had to select the "Previous View" and "Next View" buttons.

- To get/restore a separate table of contents, open "Bookmarks" from menu "View," "Navigation Panels," "Bookmarks."

- Links to some sections and figures may be hard to find. If so, search with the mouse just below the section or figure number.

- Links from the index are activated by clicking on the page number. They put you at the top of the right page; they do not put you at the right place on that page.

- A tip from Bob Sokalski: to read the book on a Nook download in PDF. Then adjust font size otherwise Nook gets absolutely confused when it is not able to fit the lines within the screen width.

Dedication

To my late parents, Piet and Rietje van Dommelen.

Contents

Reading the PDF **iii**

Dedication **v**

Contents **vii**
List of Figures xxv
List of Tables xxxiii

Preface **xxxv**
To the Student xxxv
Acknowledgments xxxvi
Comments and Feedback xxxix

I Special Relativity **1**

1 Special Relativity [Draft] **3**
1.1 Overview of Relativity 3
1.1.1 A note on the history of the theory 3
1.1.2 The mass-energy relation 4
1.1.3 The universal speed of light 5
1.1.4 Disagreements about space and time 7
1.2 The Lorentz Transformation 11
1.2.1 The transformation formulae 11
1.2.2 Proper time and distance 13
1.2.3 Subluminal and superluminal effects 15
1.2.4 Four-vectors 17
1.2.5 Index notation 18
1.2.6 Group property 22
1.3 Relativistic Mechanics 22
1.3.1 Intro to relativistic mechanics 22
1.3.2 Lagrangian mechanics 25

II Basic Quantum Mechanics 29

2 Mathematical Prerequisites 31
2.1 Complex Numbers . . . 31
2.2 Functions as Vectors . . . 34
2.3 The Dot, oops, INNER Product . . . 36
2.4 Operators . . . 40
2.5 Eigenvalue Problems . . . 41
2.6 Hermitian Operators . . . 43
2.7 Additional Points . . . 46
2.7.1 Dirac notation . . . 46
2.7.2 Additional independent variables . . . 47

3 Basic Ideas of Quantum Mechanics 49
3.1 The Revised Picture of Nature . . . 50
3.2 The Heisenberg Uncertainty Principle . . . 53
3.3 The Operators of Quantum Mechanics . . . 54
3.4 The Orthodox Statistical Interpretation . . . 57
3.4.1 Only eigenvalues . . . 57
3.4.2 Statistical selection . . . 58
3.5 A Particle Confined Inside a Pipe . . . 60
3.5.1 The physical system . . . 60
3.5.2 Mathematical notations . . . 61
3.5.3 The Hamiltonian . . . 61
3.5.4 The Hamiltonian eigenvalue problem . . . 62
3.5.5 All solutions of the eigenvalue problem . . . 63
3.5.6 Discussion of the energy values . . . 67
3.5.7 Discussion of the eigenfunctions . . . 68
3.5.8 Three-dimensional solution . . . 70
3.5.9 Quantum confinement . . . 74

4 Single-Particle Systems 77
4.1 The Harmonic Oscillator . . . 78
4.1.1 The Hamiltonian . . . 79
4.1.2 Solution using separation of variables . . . 80
4.1.3 Discussion of the eigenvalues . . . 83
4.1.4 Discussion of the eigenfunctions . . . 85
4.1.5 Degeneracy . . . 88
4.1.6 Noneigenstates . . . 90
4.2 Angular Momentum . . . 92
4.2.1 Definition of angular momentum . . . 93
4.2.2 Angular momentum in an arbitrary direction . . . 93

4.2.3 Square angular momentum . . . 96
4.2.4 Angular momentum uncertainty . . . 99
4.3 The Hydrogen Atom . . . 100
4.3.1 The Hamiltonian . . . 100
4.3.2 Solution using separation of variables . . . 102
4.3.3 Discussion of the eigenvalues . . . 106
4.3.4 Discussion of the eigenfunctions . . . 109
4.4 Expectation Value and Standard Deviation . . . 114
4.4.1 Statistics of a die . . . 115
4.4.2 Statistics of quantum operators . . . 116
4.4.3 Simplified expressions . . . 118
4.4.4 Some examples . . . 119
4.5 The Commutator . . . 122
4.5.1 Commuting operators . . . 122
4.5.2 Noncommuting operators and their commutator . . . 123
4.5.3 The Heisenberg uncertainty relationship . . . 124
4.5.4 Commutator reference . . . 125
4.6 The Hydrogen Molecular Ion . . . 129
4.6.1 The Hamiltonian . . . 129
4.6.2 Energy when fully dissociated . . . 130
4.6.3 Energy when closer together . . . 131
4.6.4 States that share the electron . . . 132
4.6.5 Comparative energies of the states . . . 134
4.6.6 Variational approximation of the ground state . . . 135
4.6.7 Comparison with the exact ground state . . . 136

5 Multiple-Particle Systems 139
5.1 Wave Function for Multiple Particles . . . 140
5.2 The Hydrogen Molecule . . . 142
5.2.1 The Hamiltonian . . . 142
5.2.2 Initial approximation to the lowest energy state . . . 143
5.2.3 The probability density . . . 144
5.2.4 States that share the electrons . . . 145
5.2.5 Variational approximation of the ground state . . . 148
5.2.6 Comparison with the exact ground state . . . 149
5.3 Two-State Systems . . . 150
5.4 Spin . . . 155
5.5 Multiple-Particle Systems Including Spin . . . 157
5.5.1 Wave function for a single particle with spin . . . 157
5.5.2 Inner products including spin . . . 159
5.5.3 Commutators including spin . . . 160
5.5.4 Wave function for multiple particles with spin . . . 161
5.5.5 Example: the hydrogen molecule . . . 164

5.5.6 Triplet and singlet states . . . 164
5.6 Identical Particles . . . 166
5.7 Ways to Symmetrize the Wave Function . . . 168
5.8 Matrix Formulation . . . 174
5.9 Heavier Atoms . . . 178
5.9.1 The Hamiltonian eigenvalue problem . . . 178
5.9.2 Approximate solution using separation of variables . . . 178
5.9.3 Hydrogen and helium . . . 181
5.9.4 Lithium to neon . . . 182
5.9.5 Sodium to argon . . . 187
5.9.6 Potassium to krypton . . . 187
5.9.7 Full periodic table . . . 188
5.10 Pauli Repulsion . . . 192
5.11 Chemical Bonds . . . 193
5.11.1 Covalent sigma bonds . . . 193
5.11.2 Covalent pi bonds . . . 194
5.11.3 Polar covalent bonds and hydrogen bonds . . . 194
5.11.4 Promotion and hybridization . . . 196
5.11.5 Ionic bonds . . . 199
5.11.6 Limitations of valence bond theory . . . 200

6 Macroscopic Systems 203
6.1 Intro to Particles in a Box . . . 204
6.2 The Single-Particle States . . . 206
6.3 Density of States . . . 208
6.4 Ground State of a System of Bosons . . . 211
6.5 About Temperature . . . 212
6.6 Bose-Einstein Condensation . . . 214
6.6.1 Rough explanation of the condensation . . . 218
6.7 Bose-Einstein Distribution . . . 223
6.8 Blackbody Radiation . . . 225
6.9 Ground State of a System of Electrons . . . 228
6.10 Fermi Energy of the Free-Electron Gas . . . 230
6.11 Degeneracy Pressure . . . 232
6.12 Confinement and the DOS . . . 234
6.13 Fermi-Dirac Distribution . . . 237
6.14 Maxwell-Boltzmann Distribution . . . 241
6.15 Thermionic Emission . . . 244
6.16 Chemical Potential and Diffusion . . . 246
6.17 Intro to the Periodic Box . . . 248
6.18 Periodic Single-Particle States . . . 249
6.19 DOS for a Periodic Box . . . 251
6.20 Intro to Electrical Conduction . . . 252

6.21 Intro to Band Structure . . . 256
6.21.1 Metals and insulators . . . 256
6.21.2 Typical metals and insulators . . . 259
6.21.3 Semiconductors . . . 263
6.21.4 Semimetals . . . 264
6.21.5 Electronic heat conduction . . . 265
6.21.6 Ionic conductivity . . . 265
6.22 Electrons in Crystals . . . 267
6.22.1 Bloch waves . . . 268
6.22.2 Example spectra . . . 269
6.22.3 Effective mass . . . 271
6.22.4 Crystal momentum . . . 272
6.22.5 Three-dimensional crystals . . . 277
6.23 Semiconductors . . . 282
6.24 The *P-N* Junction . . . 289
6.25 The Transistor . . . 295
6.26 Zener and Avalanche Diodes . . . 296
6.27 Optical Applications . . . 298
6.27.1 Atomic spectra . . . 298
6.27.2 Spectra of solids . . . 299
6.27.3 Band gap effects . . . 300
6.27.4 Effects of crystal imperfections . . . 301
6.27.5 Photoconductivity . . . 301
6.27.6 Photovoltaic cells . . . 302
6.27.7 Light-emitting diodes . . . 303
6.28 Thermoelectric Applications . . . 304
6.28.1 Peltier effect . . . 304
6.28.2 Seebeck effect . . . 309
6.28.3 Thomson effect . . . 313

7 Time Evolution 315
7.1 The Schrödinger Equation . . . 317
7.1.1 The equation . . . 317
7.1.2 Solution of the equation . . . 318
7.1.3 Energy conservation . . . 319
7.1.4 Stationary states . . . 321
7.1.5 The adiabatic approximation . . . 322
7.2 Time Variation of Expectation Values . . . 323
7.2.1 Newtonian motion . . . 324
7.2.2 Energy-time uncertainty relation . . . 326
7.3 Conservation Laws and Symmetries . . . 327
7.4 Conservation Laws in Emission . . . 333
7.4.1 Conservation of energy . . . 334

7.4.2 Combining angular momenta and parities 336
7.4.3 Transition types and their photons 339
7.4.4 Selection rules 345
7.5 Symmetric Two-State Systems 351
7.5.1 A graphical example 353
7.5.2 Particle exchange and forces 354
7.5.3 Spontaneous emission 360
7.6 Asymmetric Two-State Systems 365
7.6.1 Spontaneous emission revisited 366
7.7 Absorption and Stimulated Emission 372
7.7.1 The Hamiltonian 374
7.7.2 The two-state model 375
7.8 General Interaction with Radiation 379
7.9 Position and Linear Momentum 381
7.9.1 The position eigenfunction 382
7.9.2 The linear momentum eigenfunction 385
7.10 Wave Packets 387
7.10.1 Solution of the Schrödinger equation. 387
7.10.2 Component wave solutions 389
7.10.3 Wave packets 390
7.10.4 Group velocity 392
7.10.5 Electron motion through crystals 395
7.11 Almost Classical Motion 399
7.11.1 Motion through free space 399
7.11.2 Accelerated motion 400
7.11.3 Decelerated motion 401
7.11.4 The harmonic oscillator 401
7.12 Scattering 402
7.12.1 Partial reflection 403
7.12.2 Tunneling 403
7.13 Reflection and Transmission Coefficients 405

8 The Meaning of Quantum Mechanics 409
8.1 Schrödinger's Cat 410
8.2 Instantaneous Interactions 411
8.3 Global Symmetrization 415
8.4 A story by Wheeler 416
8.5 Failure of the Schrödinger Equation? 420
8.6 The Many-Worlds Interpretation 422
8.7 The Arrow of Time 428

III Gateway Topics 431

9 Numerical Procedures 433
9.1 The Variational Method . . . 433
9.1.1 Basic variational statement . . . 433
9.1.2 Differential form of the statement . . . 435
9.1.3 Using Lagrangian multipliers . . . 436
9.2 The Born-Oppenheimer Approximation . . . 439
9.2.1 The Hamiltonian . . . 439
9.2.2 Basic Born-Oppenheimer approximation . . . 441
9.2.3 Going one better . . . 443
9.3 The Hartree-Fock Approximation . . . 445
9.3.1 Wave function approximation . . . 446
9.3.2 The Hamiltonian . . . 453
9.3.3 The expectation value of energy . . . 455
9.3.4 The canonical Hartree-Fock equations . . . 458
9.3.5 Additional points . . . 461

10 Solids 469
10.1 Molecular Solids . . . 469
10.2 Ionic Solids . . . 472
10.3 Metals . . . 476
10.3.1 Lithium . . . 476
10.3.2 One-dimensional crystals . . . 478
10.3.3 Wave functions of one-dimensional crystals . . . 479
10.3.4 Analysis of the wave functions . . . 481
10.3.5 Floquet (Bloch) theory . . . 483
10.3.6 Fourier analysis . . . 484
10.3.7 The reciprocal lattice . . . 484
10.3.8 The energy levels . . . 485
10.3.9 Merging and splitting bands . . . 487
10.3.10 Three-dimensional metals . . . 488
10.4 Covalent Materials . . . 492
10.5 Free-Electron Gas . . . 495
10.5.1 Lattice for the free electrons . . . 495
10.5.2 Occupied states and Brillouin zones . . . 497
10.6 Nearly-Free Electrons . . . 501
10.6.1 Energy changes due to a weak lattice potential . . . 503
10.6.2 Discussion of the energy changes . . . 505
10.7 Additional Points . . . 508
10.7.1 About ferromagnetism . . . 510
10.7.2 X-ray diffraction . . . 512

11 Basic and Quantum Thermodynamics **519**
11.1 Temperature . . . 520
11.2 Single-Particle versus System States . . . 521
11.3 How Many System Eigenfunctions? . . . 526
11.4 Particle-Energy Distribution Functions . . . 531
11.5 The Canonical Probability Distribution . . . 532
11.6 Low Temperature Behavior . . . 535
11.7 The Basic Thermodynamic Variables . . . 537
11.8 Intro to the Second Law . . . 541
11.9 The Reversible Ideal . . . 542
11.10 Entropy . . . 548
11.11 The Big Lie of Distinguishable Particles . . . 555
11.12 The New Variables . . . 555
11.13 Microscopic Meaning of the Variables . . . 562
11.14 Application to Particles in a Box . . . 563
11.14.1 Bose-Einstein condensation . . . 565
11.14.2 Fermions at low temperatures . . . 566
11.14.3 A generalized ideal gas law . . . 567
11.14.4 The ideal gas . . . 568
11.14.5 Blackbody radiation . . . 569
11.14.6 The Debye model . . . 571
11.15 Specific Heats . . . 573

12 Angular momentum **579**
12.1 Introduction . . . 579
12.2 The fundamental commutation relations . . . 580
12.3 Ladders . . . 581
12.4 Possible values of angular momentum . . . 584
12.5 A warning about angular momentum . . . 586
12.6 Triplet and singlet states . . . 587
12.7 Clebsch-Gordan coefficients . . . 590
12.8 Some important results . . . 593
12.9 Momentum of partially filled shells . . . 595
12.10 Pauli spin matrices . . . 598
12.11 General spin matrices . . . 601
12.12 The Relativistic Dirac Equation . . . 602

13 Electromagnetism **605**
13.1 The Electromagnetic Hamiltonian . . . 605
13.2 Maxwell's Equations . . . 607
13.3 Example Static Electromagnetic Fields . . . 616
13.3.1 Point charge at the origin . . . 616
13.3.2 Dipoles . . . 621

13.3.3 Arbitrary charge distributions . . . 625
13.3.4 Solution of the Poisson equation . . . 627
13.3.5 Currents . . . 628
13.3.6 Principle of the electric motor . . . 630
13.4 Particles in Magnetic Fields . . . 632
13.5 Stern-Gerlach Apparatus . . . 636
13.6 Nuclear Magnetic Resonance . . . 637
13.6.1 Description of the method . . . 637
13.6.2 The Hamiltonian . . . 638
13.6.3 The unperturbed system . . . 639
13.6.4 Effect of the perturbation . . . 642

14 Nuclei [Unfinished Draft] 645
14.1 Fundamental Concepts . . . 646
14.2 Draft: The Simplest Nuclei . . . 650
14.2.1 Draft: The proton . . . 650
14.2.2 Draft: The neutron . . . 651
14.2.3 Draft: The deuteron . . . 652
14.2.4 Draft: Property summary . . . 654
14.3 Draft: Overview of Nuclei . . . 656
14.4 Draft: Magic numbers . . . 666
14.5 Draft: Radioactivity . . . 666
14.5.1 Draft: Half-life and decay rate . . . 667
14.5.2 Draft: More than one decay process . . . 671
14.5.3 Draft: Other definitions . . . 671
14.6 Draft: Mass and energy . . . 672
14.7 Draft: Binding energy . . . 674
14.8 Draft: Nucleon separation energies . . . 676
14.9 Draft: Modeling the Deuteron . . . 676
14.10 Draft: Liquid drop model . . . 686
14.10.1 Draft: Nuclear radius . . . 686
14.10.2 Draft: von Weizsäcker formula . . . 687
14.10.3 Draft: Explanation of the formula . . . 687
14.10.4 Draft: Accuracy of the formula . . . 688
14.11 Draft: Alpha Decay . . . 688
14.11.1 Draft: Decay mechanism . . . 690
14.11.2 Draft: Comparison with data . . . 692
14.11.3 Draft: Forbidden decays . . . 696
14.11.4 Draft: Why alpha decay? . . . 698
14.12 Draft: Shell model . . . 701
14.12.1 Draft: Average potential . . . 702
14.12.2 Draft: Spin-orbit interaction . . . 707
14.12.3 Draft: Example occupation levels . . . 711

14.12.4 Draft: Shell model with pairing . . . 715
14.12.5 Draft: Configuration mixing . . . 722
14.12.6 Draft: Shell model failures . . . 727
14.13 Draft: Collective Structure . . . 731
14.13.1 Draft: Classical liquid drop . . . 732
14.13.2 Draft: Nuclear vibrations . . . 734
14.13.3 Draft: Nonspherical nuclei . . . 738
14.13.4 Draft: Rotational bands . . . 738
14.14 Draft: Fission . . . 751
14.14.1 Draft: Basic concepts . . . 751
14.14.2 Draft: Some basic features . . . 752
14.15 Draft: Spin Data . . . 755
14.15.1 Draft: Even-even nuclei . . . 755
14.15.2 Draft: Odd mass number nuclei . . . 755
14.15.3 Draft: Odd-odd nuclei . . . 760
14.16 Draft: Parity Data . . . 764
14.16.1 Draft: Even-even nuclei . . . 766
14.16.2 Draft: Odd mass number nuclei . . . 766
14.16.3 Draft: Odd-odd nuclei . . . 766
14.16.4 Draft: Parity Summary . . . 766
14.17 Draft: Electromagnetic Moments . . . 771
14.17.1 Draft: Classical description . . . 771
14.17.2 Draft: Quantum description . . . 775
14.17.3 Draft: Magnetic moment data . . . 781
14.17.4 Draft: Quadrupole moment data . . . 785
14.18 Draft: Isospin . . . 787
14.18.1 Draft: Basic ideas . . . 789
14.18.2 Draft: Heavier nuclei . . . 793
14.18.3 Draft: Additional points . . . 798
14.18.4 Draft: Why does this work? . . . 799
14.19 Draft: Beta decay . . . 801
14.19.1 Draft: Introduction . . . 801
14.19.2 Draft: Energetics Data . . . 803
14.19.3 Draft: Beta decay and magic numbers . . . 809
14.19.4 Draft: Von Weizsäcker approximation . . . 810
14.19.5 Draft: Kinetic Energies . . . 813
14.19.6 Draft: Forbidden decays . . . 816
14.19.7 Draft: Data and Fermi theory . . . 821
14.19.8 Draft: Parity violation . . . 827
14.20 Draft: Gamma Decay . . . 829
14.20.1 Draft: Energetics . . . 830
14.20.2 Draft: Forbidden decays . . . 830
14.20.3 Draft: Isomers . . . 835

14.20.4 Draft: Weisskopf estimates . . . 836
14.20.5 Draft: Comparison with data . . . 844
14.20.6 Draft: Internal conversion . . . 850

IV Supplementary Information 855

A Addenda 857
A.1 Classical Lagrangian mechanics . . . 857
A.1.1 Introduction . . . 857
A.1.2 Generalized coordinates . . . 858
A.1.3 Lagrangian equations of motion . . . 859
A.1.4 Hamiltonian dynamics . . . 861
A.1.5 Fields . . . 862
A.2 An example of variational calculus . . . 864
A.3 Galilean transformation . . . 868
A.4 More on index notation . . . 870
A.5 The reduced mass . . . 875
A.6 Constant spherical potentials . . . 878
A.6.1 The eigenvalue problem . . . 879
A.6.2 The eigenfunctions . . . 879
A.6.3 About free space solutions . . . 881
A.7 Accuracy of the variational method . . . 882
A.8 Positive ground state wave function . . . 883
A.9 Wave function symmetries . . . 885
A.10 Spin inner product . . . 889
A.11 Thermoelectric effects . . . 890
A.11.1 Peltier and Seebeck coefficient ballparks . . . 890
A.11.2 Figure of merit . . . 891
A.11.3 Physical Seebeck mechanism . . . 893
A.11.4 Full thermoelectric equations . . . 893
A.11.5 Charge locations in thermoelectrics . . . 896
A.11.6 Kelvin relationships . . . 897
A.12 Heisenberg picture . . . 901
A.13 Integral Schrödinger equation . . . 904
A.14 The Klein-Gordon equation . . . 905
A.15 Quantum Field Theory in a Nanoshell . . . 908
A.15.1 Occupation numbers . . . 910
A.15.2 Creation and annihilation operators . . . 917
A.15.3 The caHermitians . . . 920
A.15.4 Recasting a Hamiltonian as a quantum field one . . . 921
A.15.5 The harmonic oscillator as a boson system . . . 923

A.15.6 Canonical (second) quantization 926
A.15.7 Spin as a fermion system 927
A.15.8 More single particle states 929
A.15.9 Field operators 931
A.15.10 Nonrelativistic quantum field theory 937
A.16 The adiabatic theorem 941
A.17 The virial theorem 943
A.18 The energy-time uncertainty relationship 946
A.19 Conservation Laws and Symmetries 947
A.19.1 An example symmetry transformation 947
A.19.2 Physical description of a symmetry 949
A.19.3 Derivation of the conservation law 952
A.19.4 Other symmetries 957
A.19.5 A gauge symmetry and conservation of charge 960
A.19.6 Reservations about time shift symmetry 963
A.20 Angular momentum of vector particles 964
A.21 Photon type 2 wave function 970
A.21.1 The wave function 970
A.21.2 Simplifying the wave function 972
A.21.3 Photon spin 974
A.21.4 Energy eigenstates 975
A.21.5 Normalization of the wave function 975
A.21.6 States of definite linear momentum 976
A.21.7 States of definite angular momentum 978
A.22 Forces by particle exchange 981
A.22.1 Classical selectostatics 982
A.22.2 Classical selectodynamics 991
A.22.3 Quantum selectostatics 997
A.22.4 Poincaré and Einstein try to save the universe 1008
A.22.5 Lorenz saves the universe 1016
A.22.6 Gupta-Bleuler condition 1018
A.22.7 The conventional Lagrangian 1021
A.22.8 Quantization following Fermi 1024
A.22.9 The Coulomb potential and the speed of light 1032
A.23 Quantization of radiation 1032
A.23.1 Properties of classical electromagnetic fields 1033
A.23.2 Photon wave functions 1035
A.23.3 The electromagnetic operators 1036
A.23.4 Properties of the observable electromagnetic field 1039
A.24 Quantum spontaneous emission 1044
A.25 Multipole transitions 1050
A.25.1 Approximate Hamiltonian 1051
A.25.2 Approximate multipole matrix elements 1054

A.25.3 Corrected multipole matrix elements . . . 1055
A.25.4 Matrix element ballparks . . . 1058
A.25.5 Selection rules . . . 1060
A.25.6 Ballpark decay rates . . . 1064
A.25.7 Wave functions of definite angular momentum . . . 1066
A.25.8 Weisskopf and Moszkowski estimates . . . 1068
A.25.9 Errors in other sources . . . 1078
A.26 Fourier inversion theorem and Parseval . . . 1079
A.27 Details of the animations . . . 1084
A.28 WKB Theory of Nearly Classical Motion . . . 1092
A.29 WKB solution near the turning points . . . 1096
A.30 Three-dimensional scattering . . . 1100
A.30.1 Partial wave analysis . . . 1103
A.30.2 Partial wave amplitude . . . 1105
A.30.3 The Born approximation . . . 1107
A.31 The Born series . . . 1108
A.32 The evolution of probability . . . 1110
A.33 Explanation of the London forces . . . 1114
A.34 Explanation of Hund's first rule . . . 1118
A.35 The third law . . . 1120
A.36 Alternate Dirac equations . . . 1122
A.37 Maxwell's wave equations . . . 1123
A.38 Perturbation Theory . . . 1126
A.38.1 Basic perturbation theory . . . 1126
A.38.2 Ionization energy of helium . . . 1128
A.38.3 Degenerate perturbation theory . . . 1131
A.38.4 The Zeeman effect . . . 1133
A.38.5 The Stark effect . . . 1134
A.39 The relativistic hydrogen atom . . . 1138
A.39.1 Introduction . . . 1138
A.39.2 Fine structure . . . 1140
A.39.3 Weak and intermediate Zeeman effect . . . 1146
A.39.4 Lamb shift . . . 1147
A.39.5 Hyperfine splitting . . . 1148
A.40 Deuteron wave function . . . 1150
A.41 Deuteron model . . . 1153
A.41.1 The model . . . 1154
A.41.2 The repulsive core . . . 1156
A.41.3 Spin dependence . . . 1158
A.41.4 Noncentral force . . . 1158
A.41.5 Spin-orbit interaction . . . 1161
A.42 Nuclear forces . . . 1162
A.42.1 Basic Yukawa potential . . . 1162

A.42.2 OPEP potential 1166
A.42.3 Explanation of the OPEP potential 1167
A.42.4 Multiple pion exchange and such 1175
A.43 Classical vibrating drop 1177
A.43.1 Basic definitions 1177
A.43.2 Kinetic energy 1178
A.43.3 Energy due to surface tension 1181
A.43.4 Energy due to Coulomb repulsion 1183
A.43.5 Frequency of vibration 1185
A.44 Relativistic neutrinos 1186
A.45 Fermi theory 1191
A.45.1 Form of the wave function 1192
A.45.2 Source of the decay 1194
A.45.3 Allowed or forbidden 1198
A.45.4 The nuclear operator 1200
A.45.5 Fermi's golden rule 1203
A.45.6 Mopping up 1207
A.45.7 Electron capture 1211

D Derivations **1213**
D.1 Generic vector identities 1213
D.2 Some Green's functions 1214
D.2.1 The Poisson equation 1214
D.2.2 The screened Poisson equation 1217
D.3 Lagrangian mechanics 1218
D.3.1 Lagrangian equations of motion 1218
D.3.2 Hamiltonian dynamics 1219
D.3.3 Fields 1220
D.4 Lorentz transformation derivation 1224
D.5 Lorentz group property derivation 1225
D.6 Lorentz force derivation 1227
D.7 Derivation of the Euler formula 1227
D.8 Completeness of Fourier modes 1228
D.9 Momentum operators are Hermitian 1232
D.10 The curl is Hermitian 1233
D.11 Extension to three-dimensional solutions 1234
D.12 The harmonic oscillator solution 1236
D.13 The harmonic oscillator and uncertainty 1239
D.14 The spherical harmonics 1240
D.14.1 Derivation from the eigenvalue problem 1240
D.14.2 Parity 1242
D.14.3 Solutions of the Laplace equation 1243
D.14.4 Orthogonal integrals 1243

D.14.5 Another way to find the spherical harmonics 1244
D.14.6 Still another way to find them 1245
D.15 The hydrogen radial wave functions 1245
D.16 Constant spherical potentials derivations 1247
D.16.1 The eigenfunctions 1248
D.16.2 The Rayleigh formula 1248
D.17 Inner product for the expectation value 1249
D.18 Eigenfunctions of commuting operators 1250
D.19 The generalized uncertainty relationship 1251
D.20 Derivation of the commutator rules 1252
D.21 Solution of the hydrogen molecular ion 1254
D.22 Unique ground state wave function 1256
D.23 Solution of the hydrogen molecule 1265
D.24 Hydrogen molecule ground state and spin 1266
D.25 Number of boson states 1267
D.26 Density of states 1268
D.27 Radiation from a hole 1270
D.28 Kirchhoff's law 1271
D.29 The thermionic emission equation 1272
D.30 Number of conduction band electrons 1275
D.31 Integral Schrödinger equation 1275
D.32 Integral conservation laws 1277
D.33 Quantum field derivations 1280
D.34 The adiabatic theorem 1283
D.35 The evolution of expectation values 1287
D.36 Photon wave function derivations 1287
D.36.1 Rewriting the energy integral 1288
D.36.2 Angular momentum states 1289
D.37 Forces by particle exchange derivations 1296
D.37.1 Classical energy minimization 1296
D.37.2 Quantum energy minimization 1297
D.37.3 Rewriting the Lagrangian 1298
D.37.4 Coulomb potential energy 1299
D.38 Time-dependent perturbation theory 1300
D.39 Selection rules 1302
D.40 Quantization of radiation derivations 1307
D.41 Derivation of the Einstein B coefficients 1310
D.42 Derivation of the Einstein A coefficients 1314
D.43 Multipole derivations 1315
D.43.1 Matrix element for linear momentum modes 1317
D.43.2 Matrix element for angular momentum modes 1319
D.43.3 Weisskopf and Moszkowski estimates 1322
D.44 Derivation of group velocity 1326

D.45 Motion through crystals . . . 1328
D.45.1 Propagation speed . . . 1328
D.45.2 Motion under an external force . . . 1329
D.45.3 Free-electron gas with constant electric field . . . 1330
D.46 Derivation of the WKB approximation . . . 1331
D.47 Born differential cross section . . . 1333
D.48 About Lagrangian multipliers . . . 1334
D.49 The generalized variational principle . . . 1335
D.50 Spin degeneracy . . . 1337
D.51 Born-Oppenheimer nuclear motion . . . 1337
D.52 Simplification of the Hartree-Fock energy . . . 1342
D.53 Integral constraints . . . 1346
D.54 Derivation of the Hartree-Fock equations . . . 1347
D.55 Why the Fock operator is Hermitian . . . 1356
D.56 Number of system eigenfunctions . . . 1356
D.57 The particle energy distributions . . . 1360
D.58 The canonical probability distribution . . . 1366
D.59 Analysis of the ideal gas Carnot cycle . . . 1368
D.60 Checks on the expression for entropy . . . 1369
D.61 Chemical potential in the distributions . . . 1372
D.62 Fermi-Dirac integrals at low temperature . . . 1375
D.63 Angular momentum uncertainty . . . 1377
D.64 Spherical harmonics by ladder operators . . . 1377
D.65 How to make Clebsch-Gordan tables . . . 1378
D.66 The triangle inequality . . . 1379
D.67 Momentum of shells . . . 1380
D.68 Awkward questions about spin . . . 1383
D.69 More awkwardness about spin . . . 1384
D.70 Emergence of spin from relativity . . . 1385
D.71 Electromagnetic commutators . . . 1387
D.72 Various electrostatic derivations. . . . 1389
D.72.1 Existence of a potential . . . 1389
D.72.2 The Laplace equation . . . 1390
D.72.3 Egg-shaped dipole field lines . . . 1391
D.72.4 Ideal charge dipole delta function . . . 1392
D.72.5 Integrals of the current density . . . 1392
D.72.6 Lorentz forces on a current distribution . . . 1393
D.72.7 Field of a current dipole . . . 1395
D.72.8 Biot-Savart law . . . 1397
D.73 Orbital motion in a magnetic field . . . 1397
D.74 Electron spin in a magnetic field . . . 1399
D.75 Solving the NMR equations . . . 1400
D.76 Harmonic oscillator revisited . . . 1400

D.77 Impenetrable spherical shell . . . 1402
D.78 Shell model quadrupole moment . . . 1402
D.79 Derivation of perturbation theory . . . 1403
D.80 Hydrogen ground state Stark effect . . . 1408
D.81 Dirac fine structure Hamiltonian . . . 1410
D.82 Classical spin-orbit derivation . . . 1417
D.83 Expectation powers of r for hydrogen . . . 1419
D.84 Band gap explanation derivations . . . 1424

N Notes **1427**
N.1 Why this book? . . . 1427
N.2 History and wish list . . . 1431
N.3 Nature and real eigenvalues . . . 1443
N.4 Are Hermitian operators really like that? . . . 1443
N.5 Why boundary conditions are tricky . . . 1444
N.6 Is the variational approximation best? . . . 1445
N.7 Shielding approximation limitations . . . 1445
N.8 Why the s states have the least energy . . . 1446
N.9 Explanation of the band gaps . . . 1446
N.10 A less fishy story . . . 1452
N.11 Better description of two-state systems . . . 1454
N.12 Second quantization in other books . . . 1454
N.13 Combining angular momentum factors . . . 1455
N.14 The electric multipole problem . . . 1458
N.15 A tenth of a googol in universes . . . 1461
N.16 A single Slater determinant is not exact . . . 1461
N.17 Generalized orbitals . . . 1462
N.18 "Correlation energy" . . . 1464
N.19 Ambiguities in electron affinity . . . 1467
N.20 Why Floquet theory should be called so . . . 1469
N.21 Superfluidity versus BEC . . . 1469
N.22 The mechanism of ferromagnetism . . . 1471
N.23 Fundamental assumption of statistics . . . 1472
N.24 A problem if the energy is given . . . 1474
N.25 The recipe of life . . . 1475
N.26 Physics of the fundamental commutators . . . 1476
N.27 Magnitude of components of vectors . . . 1477
N.28 Adding angular momentum components . . . 1478
N.29 Clebsch-Gordan tables are bidirectional . . . 1478
N.30 Machine language Clebsch-Gordan tables . . . 1478
N.31 Existence of magnetic monopoles . . . 1478
N.32 More on Maxwell's third law . . . 1479
N.33 Setting the record straight on alignment . . . 1479

N.34 NuDat 2 data selection 1479
N.35 Auger discovery 1481
N.36 Draft: Cage-of-Faraday proposal 1482

Web Pages **1485**

References **1489**

Notations **1493**

Index **1539**

List of Figures

1.1 Different views of the same experiment. 8
1.2 Coordinate systems for the Lorentz transformation. 11
1.3 Example elastic collision seen by different observers. 23
1.4 A completely inelastic collision. 25

2.1 The classical picture of a vector. 34
2.2 Spike diagram of a vector. 35
2.3 More dimensions. 35
2.4 Infinite dimensions. 35
2.5 The classical picture of a function. 35
2.6 Forming the dot product of two vectors. 37
2.7 Forming the inner product of two functions. 38
2.8 Illustration of the eigenfunction concept. 42

3.1 The old incorrect Newtonian physics. 51
3.2 The correct quantum physics. 51
3.3 Illustration of the Heisenberg uncertainty principle. 54
3.4 Classical picture of a particle in a closed pipe. 60
3.5 Quantum mechanics picture of a particle in a closed pipe. 60
3.6 Definitions for one-dimensional motion in a pipe. 61
3.7 One-dimensional energy spectrum for a particle in a pipe. 67
3.8 One-dimensional ground state of a particle in a pipe. 69
3.9 Second and third lowest one-dimensional energy states. 69
3.10 Definition of all variables for motion in a pipe. 71
3.11 True ground state of a particle in a pipe. 73
3.12 True second and third lowest energy states. 73
3.13 A combination of ψ_{111} and ψ_{211} seen at some typical times. . . . 75

4.1 Classical picture of an harmonic oscillator. 78
4.2 The energy spectrum of the harmonic oscillator. 84
4.3 Ground state of the harmonic oscillator 86
4.4 Wave functions ψ_{100} and ψ_{010}. 87
4.5 Energy eigenfunction ψ_{213}. 88
4.6 Arbitrary wave function (not an energy eigenfunction). 91

4.7 Spherical coordinates of an arbitrary point P. 94
4.8 Spectrum of the hydrogen atom. 107
4.9 Ground state wave function of the hydrogen atom. 110
4.10 Eigenfunction ψ_{200}. 111
4.11 Eigenfunction ψ_{210}, or $2p_z$. 111
4.12 Eigenfunction ψ_{211} (and ψ_{21-1}). 112
4.13 Eigenfunctions $2p_x$, left, and $2p_y$, right. 112
4.14 Hydrogen atom plus free proton far apart. 130
4.15 Hydrogen atom plus free proton closer together. 131
4.16 The electron being antisymmetrically shared. 132
4.17 The electron being symmetrically shared. 133

5.1 State with two neutral atoms. 145
5.2 Symmetric sharing of the electrons. 147
5.3 Antisymmetric sharing of the electrons. 147
5.4 Approximate solutions for hydrogen and helium. 182
5.5 Abbreviated periodic table of the elements. 183
5.6 Approximate solutions for lithium (left) and beryllium (right). . 184
5.7 Example approximate solution for boron. 185
5.8 Periodic table of the elements. 189
5.9 Covalent sigma bond consisting of two $2p_z$ states. 194
5.10 Covalent pi bond consisting of two $2p_x$ states. 195
5.11 Covalent sigma bond consisting of a $2p_z$ and a 1s state. 195
5.12 Shape of an sp^3 hybrid state. 198
5.13 Shapes of the sp^2 and sp hybrids. 198

6.1 Allowed wave number vectors, left, and energy spectrum, right. . 207
6.2 Ground state of a system of noninteracting bosons in a box. . . 211
6.3 The system of bosons at a very low temperature. 215
6.4 The system of bosons at a relatively low temperature. 215
6.5 Ground state energy eigenfunction for a simple system. 218
6.6 State with 5 times the single-particle ground state energy. . . 218
6.7 Distinguishable particles: eigenfunctions for distribution A. . . . 219
6.8 Distinguishable particles: eigenfunctions for distribution B. . . . 220
6.9 Bosons: only 3 energy eigenfunctions for distribution A. 221
6.10 Bosons: also only 3 energy eigenfunctions for distribution B. . . 222
6.11 Ground state of noninteracting electrons (fermions) in a box. . . 229
6.12 Severe confinement in the y-direction. 235
6.13 Severe confinement in both the y and z directions. 236
6.14 Severe confinement in all three directions. 237
6.15 A system of fermions at a nonzero temperature. 239
6.16 Particles at high-enough temperature and volume. 242
6.17 Ground state of noninteracting electrons in a periodic box. . . . 251

6.18 Conduction in the free-electron gas model. 253
6.19 Sketch of electron spectra in solids at zero temperature. 257
6.20 Sketch of electron spectra in solids at nonzero temperature. . . . 263
6.21 Potential energy seen by an electron along a line of nuclei. . . . 267
6.22 Potential energy in the one-dimensional Kronig & Penney model. 267
6.23 Example Kronig & Penney spectra. 269
6.24 Spectrum against wave number in the extended zone scheme. . . 273
6.25 Spectrum against wave number in the reduced zone scheme. . . 274
6.26 One-dimensional energy bands for basic semiconductors. 274
6.27 Spectrum against wave number in the periodic zone scheme. . . 276
6.28 Schematic of the zinc blende (ZnS) crystal. 278
6.29 First Brillouin zone of the FCC crystal. 280
6.30 Sketch of a more complete spectrum of germanium. 281
6.31 Vicinity of the band gap of intrinsic and doped semiconductors. 283
6.32 Relationship between conduction electron density and hole density. 287
6.33 The *p*-*n* junction in thermal equilibrium. 289
6.34 Schematic of the operation of an *p*-*n* junction. 292
6.35 Schematic of the operation of an *n*-*p*-*n* transistor. 295
6.36 Vicinity of the band gap of an insulator. 300
6.37 Peltier cooling. 305
6.38 An example Seebeck voltage generator. 309
6.39 The Galvani potential does not produce a usable voltage. 311
6.40 The Seebeck effect is not directly measurable. 312

7.1 The ground state wave function looks the same at all times. . . 321
7.2 The first excited state at all times. 322
7.3 Concept sketch of the emission of a photon by an atom. 333
7.4 Addition of angular momenta in classical physics. 336
7.5 Longest and shortest possible final momenta in classical physics. 337
7.6 A combination of two eigenfunctions at some typical times. . . . 354
7.7 Energy slop diagram. 368
7.8 Schematized energy slop diagram. 368
7.9 Emission and absorption of radiation by an atom. 372
7.10 Dirac delta function. 383
7.11 The real part (red) and envelope (black) of an example wave. . . 389
7.12 The wave moves with the phase speed. 390
7.13 The real part and magnitude or envelope of a wave packet. . . . 390
7.14 The velocities of wave and envelope are not equal. 391
7.15 A particle in free space. 400
7.16 An accelerating particle. 400
7.17 A decelerating particle. 401
7.18 Unsteady solution for the harmonic oscillator. 402
7.19 A partial reflection. 403

7.20 An tunneling particle. 404
7.21 Penetration of an infinitely high potential energy barrier. 404
7.22 Schematic of a scattering wave packet. 405

8.1 Separating the hydrogen ion. 411
8.2 Before the Venus measurement and immediately after it. 412
8.3 Spin measurement directions. 413
8.4 Earth's view of events and that of a moving observer. 414
8.5 The space-time diagram of Wheeler's single electron. 417
8.6 Bohm's version of the Einstein, Podolski, Rosen Paradox. 422
8.7 Nonentangled positron and electron spins; up and down. 423
8.8 Nonentangled positron and electron spins; down and up. 423
8.9 The wave functions of two universes combined 424
8.10 The Bohm experiment repeated. 426
8.11 Repeated experiments on the same electron. 427

10.1 Billiard-ball model of the salt molecule. 472
10.2 Billiard-ball model of a salt crystal. 473
10.3 The salt crystal disassembled to show its structure. 475
10.4 The lithium atom, scaled more correctly than before. 476
10.5 Body-centered-cubic (BCC) structure of lithium. 477
10.6 Fully periodic wave function of a two-atom lithium "crystal." . . 478
10.7 Flip-flop wave function of a two-atom lithium "crystal." 480
10.8 Wave functions of a four-atom lithium "crystal." 481
10.9 Reciprocal lattice of a one-dimensional crystal. 485
10.10 Schematic of energy bands. 486
10.11 Schematic of merging bands. 487
10.12 A primitive cell and primitive translation vectors of lithium. . . 488
10.13 Wigner-Seitz cell of the BCC lattice. 489
10.14 Schematic of crossing bands. 493
10.15 Ball and stick schematic of the diamond crystal. 494
10.16 Assumed simple cubic reciprocal lattice in cross-section. 496
10.17 Occupied states for one, two, and three electrons per lattice cell. 499
10.18 Redefining the occupied wave numbers into Brillouin zones. . . . 500
10.19 Second, third, and fourth zones in the periodic zone scheme. . . 501
10.20 The wavenumber vector of a sample free electron wave function. 502
10.21 The grid of nonzero Hamiltonian perturbation coefficients. . . . 504
10.22 Tearing apart of the wave number space energies. 505
10.23 Effect of a lattice potential on the energy. 506
10.24 Bragg planes seen in wave number space cross section. 507
10.25 Occupied states if there are two valence electrons per lattice cell. 508
10.26 Smaller lattice potential. 509
10.27 Depiction of an electromagnetic ray. 513

10.28 Law of reflection in elastic scattering from a plane. 514
10.29 Scattering from multiple "planes of atoms." 515
10.30 Difference in distance when scattered from P rather than O. . . 516

11.1 An arbitrary system eigenfunction for 36 distinguishable particles. 523
11.2 An arbitrary system eigenfunction for 36 identical bosons. . . . 524
11.3 An arbitrary system eigenfunction for 33 identical fermions. . . 525
11.4 Illustrative small model system having 4 distinguishable particles. 527
11.5 The number of system eigenfunctions for a model system. 528
11.6 Number of energy eigenfunctions on the oblique energy line. . . 530
11.7 Probabilities if there is uncertainty in energy. 534
11.8 Probabilities if shelf 1 is a nondegenerate ground state. 535
11.9 Like the previous figure, but at a lower temperature. 536
11.10 Like the previous figures, but at a still lower temperature. . . . 536
11.11 Schematic of the Carnot refrigeration cycle. 543
11.12 Schematic of the Carnot heat engine. 546
11.13 A generic heat pump next to a reversed Carnot one. 547
11.14 Comparison of integration paths for finding the entropy. 549
11.15 Specific heat at constant volume of gases. 574
11.16 Specific heat at constant pressure of solids. 576

12.1 Example bosonic ladders. 583
12.2 Example fermionic ladders. 584
12.3 Triplet and singlet states in terms of ladders 589
12.4 Clebsch-Gordan coefficients of two spin one half particles. 590
12.5 Clebsch-Gordan coefficients for second momentum one-half. . . . 592
12.6 Clebsch-Gordan coefficients for second angular momentum one. 594

13.1 Relationship of Maxwell's first equation to Coulomb's law. . . . 608
13.2 Maxwell's first equation for a more arbitrary region. 609
13.3 The net number of outgoing field lines indicates net charge. . . . 610
13.4 The net number of outgoing magnetic field lines is zero. 611
13.5 Electric power generation. 612
13.6 Two ways to generate a magnetic field. 613
13.7 Electric field and potential of a uniform spherical charge. 620
13.8 Electric field of a two-dimensional line charge. 620
13.9 Field lines of a vertical electric dipole. 621
13.10 Electric field of a two-dimensional dipole. 622
13.11 Field of an ideal magnetic dipole. 623
13.12 Electric field of an almost ideal two-dimensional dipole. 624
13.13 Magnetic field lines around an infinite straight electric wire. . . 628
13.14 An electromagnet consisting of a single wire loop. 629
13.15 A current dipole. 630
13.16 Electric motor using a single wire loop. 631

13.17 Computation of the moment on a wire loop in a magnetic field. 631
13.18 Larmor precession of the expectation spin. 641
13.19 Probability of being able to find the nuclei at elevated energy. . 643
13.20 Maximum probability of finding the nuclei at elevated energy. . 643
13.21 Effect of a magnetic field rotating at the Larmor frequency. . . . 644

14.1 Chart of the nuclides. 659
14.2 Nuclear decay modes. 660
14.3 Nuclear half-lifes. 669
14.4 Binding energy per nucleon. 675
14.5 Proton separation energy. 677
14.6 Neutron separation energy. 678
14.7 Proton pair separation energy. 679
14.8 Neutron pair separation energy. 680
14.9 Error in the von Weizsäcker formula. 689
14.10 Half-life versus energy release in alpha decay. 690
14.11 Schematic potential for a tunneling alpha particle. 691
14.12 Half-life predicted by the Gamow / Gurney & Condon theory. . 695
14.13 Example average nuclear potentials. 702
14.14 Nuclear energy levels for various average nuclear potentials. . . . 705
14.15 Schematic effect of spin-orbit interaction on the energy levels. . 710
14.16 Energy levels for doubly-magic oxygen-16 and neighbors. 712
14.17 Nucleon pairing effect. 716
14.18 Energy levels for neighbors of doubly-magic calcium-40. 721
14.19 2^+ excitation energy of even-even nuclei. 724
14.20 Collective motion effects. 726
14.21 Failures of the shell model. 728
14.22 An excitation energy ratio for even-even nuclei. 736
14.23 Textbook vibrating nucleus tellurium-120. 737
14.24 Rotational bands of hafnium-177. 740
14.25 Ground state rotational band of tungsten-183. 745
14.26 Rotational bands of aluminum-25. 746
14.27 Rotational bands of erbium-164. 747
14.28 Ground state rotational band of magnesium-24. 748
14.29 Rotational bands of osmium-190. 750
14.30 Simplified energetics for fission of fermium-256. 754
14.31 Spin of even-even nuclei. 756
14.32 Spin of even-odd nuclei. 758
14.33 Spin of odd-even nuclei. 759
14.34 Spin of odd-odd nuclei. 762
14.35 Odd-odd spins predicted using the neighbors. 763
14.36 Odd-odd spins predicted from theory. 765
14.37 Parity of even-even nuclei. 767

14.38 Parity of even-odd nuclei. 768
14.39 Parity of odd-even nuclei. 769
14.40 Parity of odd-odd nuclei. 770
14.41 Parity versus the shell model. 772
14.42 Magnetic dipole moments of the ground-state nuclei. 783
14.43 2^+ magnetic moment of even-even nuclei. 784
14.44 Electric quadrupole moment. 786
14.45 Electric quadrupole moment corrected for spin. 788
14.46 Isobaric analog states. 795
14.47 Energy release in beta decay of even-odd nuclei. 804
14.48 Energy release in beta decay of odd-even nuclei. 805
14.49 Energy release in beta decay of odd-odd nuclei. 806
14.50 Energy release in beta decay of even-even nuclei. 807
14.51 Examples of beta decay. 811
14.52 The Fermi integral. 820
14.53 Beta decay rates. 822
14.54 Beta decay rates as fraction of a ballparked value. 823
14.55 Parity violation. 828
14.56 Energy levels of tantalum-180. 832
14.57 Half-life of the longest-lived even-odd isomers. 837
14.58 Half-life of the longest-lived odd-even isomers. 838
14.59 Half-life of the longest-lived odd-odd isomers. 839
14.60 Half-life of the longest-lived even-even isomers. 840
14.61 Weisskopf ballpark half-lifes for electromagnetic transitions. . . 842
14.62 Moszkowski ballpark half-lifes for magnetic transitions. 843
14.63 Comparison of electric gamma decay rates with theory. 845
14.64 Comparison of magnetic gamma decay rates with theory. 846
14.65 Decay rates between the same initial and final states. 849

A.1 Analysis of conduction. 897
A.2 A system eigenfunction for 36 distinguishable particles. 911
A.3 A system energy eigenfunction for 36 identical bosons. 912
A.4 A system energy eigenfunction for 33 identical fermions. 913
A.5 Wave functions with just one type of single particle state. 915
A.6 Creation and annihilation operators for one single particle state. 917
A.7 Effect of coordinate system rotation on spherical coordinates . . 948
A.8 Effect of coordinate system rotation on a vector. 965
A.9 Example energy eigenfunction for the particle in free space. . . . 1084
A.10 Example energy eigenfunction for an accelerating force. 1085
A.11 Example energy eigenfunction for a decelerating force. 1087
A.12 Example energy eigenfunction for the harmonic oscillator. 1087
A.13 Example energy eigenfunction for a brief accelerating force. . . . 1088
A.14 Example energy eigenfunction for tunneling through a barrier. . 1089

A.15 Tunneling through a delta function barrier. 1089
A.16 Harmonic oscillator potential energy and example eigenfunction. 1092
A.17 The Airy Ai and Bi functions. 1096
A.18 Connection formulae for going from normal to tunneling. 1098
A.19 Connection formulae for going from tunneling to normal. 1098
A.20 WKB approximation of tunneling. 1099
A.21 Scattering of a beam off a target. 1100
A.22 Graphical interpretation of the Born series. 1109
A.23 Possible polarizations of a pair of hydrogen atoms. 1115
A.24 Crude deuteron model. 1154
A.25 Crude deuteron model with a 0.5 fm repulsive core. 1157
A.26 Effects of uncertainty in orbital angular momentum. 1161
A.27 Possible momentum states for a particle in a periodic box. . . . 1207

D.1 Blunting of the absolute potential. 1256
D.2 Bosons in single-particle-state boxes. 1267
D.3 Schematic of an example boson distribution on a shelf. 1359
D.4 Schematic of the Carnot refrigeration cycle. 1368

N.1 Spectrum for a weak potential. 1447
N.2 The 17 real wave functions of lowest energy. 1448
N.3 Spherical coordinates of an arbitrary point P. 1532

List of Tables

4.1 One-dimensional eigenfunctions of the harmonic oscillator. . . . 82
4.2 The first few spherical harmonics. 97
4.3 The first few spherical harmonics rewritten. 98
4.4 The first few radial wave functions for hydrogen. 104

6.1 Lowest single-particle energy in a cube with 1 cm sides. 208

7.1 Properties of photons emitted in multipole transitions. 343

12.1 Possible combined angular momentum of identical fermions. . . 596
12.2 Possible combined angular momentum of identical bosons. . . . 599

13.1 Electromagnetics I: Fundamental equations and basic solutions. 617
13.2 Electromagnetics II: Electromagnetostatic solutions. 618

14.1 Properties of the electron and of the simplest nuclei. 655
14.2 Alternate names for nuclei. 664
14.3 Candidates for nuclei ejected by fermium-256 and others. 700
14.4 Spin and parity changes in electromagnetic transitions. 833
14.5 Half lifes for E0 transitions. 852

A.1 Radial correction factors for hydrogen atom wave functions. . . 1071
A.2 More realistic radial integral correction factors for nuclei. . . . 1073
A.3 Gamma-decay angular integral correction factors. 1075
A.4 Deuteron model data. 1155
A.5 Deuteron model data with a repulsive core of 0.5 fm. 1157

D.1 Additional combined angular momentum values. 1381

Preface

To the Student

This is a book on the real quantum mechanics. On quantum scales it becomes clear that classical physics is simply wrong. It is quantum mechanics that describes how nature truly behaves; classical physics is just a simplistic approximation of it that can be used for some computations describing macroscopic systems. And not too many of those, either.

Here you will find the same story as physicists tell their own students. The difference is that this book is designed to be much easier to read and understand than comparable texts. Quantum mechanics is inherently mathematical, and this book explains it fully. But the mathematics is only covered to the extent that it provides insight in quantum mechanics. This is not a book for developing your skills in clever mathematical manipulations that have absolutely nothing to do with physical understanding. You can find many other texts like that already, if that is your goal.

The book was primarily written for engineering graduate students who find themselves caught up in nano technology. It is a simple fact that the typical engineering education does not provide anywhere close to the amount of physics you will need to make sense out of the literature of your field. You can start from scratch as an undergraduate in the physics department, or you can read this book.

The first part of this book provides a solid introduction to classical (i.e. non-relativistic) quantum mechanics. It is intended to explain the ideas both rigorously and clearly. It follows a "just-in-time" learning approach. The mathematics is fully explained, but not emphasized. The intention is not to practice clever mathematics, but to understand quantum mechanics. The coverage is at the normal calculus and physics level of undergraduate engineering students. If you did well in these courses, you should be able to understand the discussion, assuming that you start reading from the beginning. In particular, you simply cannot skip the short first chapter. There are some hints in the notations section, if you forgot some calculus. If you forgot some physics, just don't worry too much about it: quantum physics is so much different that even the most basic concepts need to be covered from scratch.

Whatever you do, read all of chapters 2 and 3. That is the very language of quantum mechanics. It will be hard to read the rest of the book if you do not know the language.

Derivations are usually "banned" to notes at the end of this book, in case you need them for one reason or the other. They correct a considerable number of mistakes that you will find in other books. No doubt they add a few new ones. Let me know and I will correct them quickly; that is the advantage of a web book.

The second part of this book discusses more advanced topics. It starts with numerical methods, since engineering graduate students are typically supported by a research grant, and the quicker you can produce some results, the better. A description of density functional theory is still missing, unfortunately.

The remaining chapters of the second part are intended to provide a crash course on many topics that nano literature would consider elementary physics, but that nobody has ever told you about. Most of it is not really part of what is normally understood to be a quantum mechanics course. Reading, rereading, and understanding it is highly recommended anyway.

The purpose is not just to provide basic literacy in those topics, although that is very important. But the purpose is also explain enough of their fundamentals, in terms that an engineer can understand, so that you can make sense of the literature in those fields if you do need to know more than can be covered here. Consider these chapters gateways into their topic areas.

There is a final chapter in part II on how to interpret quantum mechanics philosophically. Read it if you are interested; it will probably not help you do quantum mechanics any better. But as a matter of basic literacy, it is good to know how truly weird quantum mechanics really is.

The usual "Why this book?" blah-blah can be found in a note at the back of this book, {N.1} A version history is in note {N.2}.

Acknowledgments

This book is for a large part based on my reading of the excellent book by Griffiths, [25]. It now contains essentially all material in that book in one way or the other. (But you may need to look in the notes for some of it.) This book also evolved to include a lot of additional material that I thought would be appropriate for a physically-literate engineer. There are chapters on relativity, numerical methods, thermodynamics, solid mechanics, electromagnetism, and nuclei.

Somewhat to my surprise, I find that my coverage actually tends to be closer to Yariv's book, [52]. I still think Griffiths is more readable for an engineer, though Yariv has some very good items that Griffiths does not.

Matthew Leung pointed out that I had "left" and "right" mixed up in my discussion of the relativistic Doppler effect. I am dyslexic that way.

The idea of using the Lagrangian for the derivations of relativistic mechanics is from A. Kompanayets, *theoretical physics*, an excellent book.

I rewrote the section on functions as vectors to some extent based on comments of Germano Galasso.

I thank Rob Vossen for pointing out some rather horrible typos in the section on Dirac notation.

I thank Chris Cline for pointing out a bad label on the dot product figure in the discussion of functions as vectors. I thank Richard Mertz and Mike Day for pointing out typos and poor phrasing in the same sections.

The nanomaterials lectures of colleague Anter El-Azab that I audited inspired me to add a bit on simple quantum confinement to the first system studied, the particle in the box. That does add a bit to a section that I wanted to keep as simple as possible, but then I figure it also adds a sense that this is really relevant stuff for future engineers. I also added a discussion of the effects of confinement on the density of states to the section on the free-electron gas.

I thank Swapnil Jain for pointing out that the initial subsection on quantum confinement in the pipe was definitely unclear and is hopefully better now.

I thank Ed Williams for pointing out a mistake in the formula for the combination probabilities of the hydrogen atom electrons and Johann Joss for one in the formula for the averaged energy of two-state systems.

Thomas Pak noted some poor phrasing in the section on metals and insulators.

The discussions on two-state systems are mainly based on Feynman's notes, [22, chapters 8-11]. Since it is hard to determine the precise statements being made, much of that has been augmented by data from web sources, mainly those referenced.

I thank Murat Ozer for pointing out that the two highest wave functions in N.2 were $Z = 14$ instead of 16.

The discussion of the Onsager theorem comes from Desloge, [12], an emeritus professor of physics at the Florida State University.

The section on conservation laws and symmetries is almost completely based on Feynman, [22] and [20].

Harald Kirsch reported various problems in the sections on conservation laws and on position eigenfunctions.

Bob Sokalski reported an error in the section on the two-state model.

The note on the derivation of the selection rules is from [25] and lecture notes from a University of Tennessee quantum course taught by Marianne Breinig. The subsection on conservation laws and selection rules was inspired by Ellis, [15].

The many-worlds discussion is based on Everett's exposition, [17]. It is brilliant but quite impenetrable.

The section on the Born-Oppenheimer approximation comes from Wikipedia, [[21]], with modifications including the inclusion of spin.

The section on the Hartree-Fock method is mainly based on Szabo and Ostlund [46], a well-written book, with some Parr and Yang [34] thrown in.

The section on solids is mainly based on Sproull, [42], a good source for practical knowledge about application of the concepts. It is surprisingly up to date, considering it was written half a century ago. Various items, however, come from Kittel [29]. The discussion of ionic solids really comes straight from hyperphysics [[6]]. I prefer hyperphysics' example of NaCl, instead of Sproull's equivalent discussion of KCl. My colleague Steve Van Sciver helped me get some handle on what to say about helium and Bose-Einstein condensation.

The thermodynamics section started from Griffiths' discussion, [25], which follows Yariv's, [52]. However, it expanded greatly during writing. It now comes mostly from Baierlein [4], with some help from Feynman, [18], and some of the books I use in undergraduate thermo.

Mark Troll noted that the discussion of the specific heat of gases was pretty poorly written. I have rewritten it pretty much along the lines he suggested.

The derivation of the classical energy of a spinning particle in a magnetic field is from Yariv, [52].

The initial inspiration for the chapter on nuclear physics was the Nobel Prize acceptance lecture of Goeppert Mayer [[10]]. This is an excellent introduction to nuclear physics for a nonspecialist audience. It is freely available on the web. As the chapter expanded, the main reference became the popular book by Krane [31]. That book is particularly recommended if you want an understandable description of how the experimental evidence led physicists to formulate the theoretical models for nuclei. Other primary references were [36] and [40]. The Handbook of Physics, Hyperphysics, and various other web sources were also helpful. Much of the experimental data are from NUBASE 2003, an official database of nuclei, [3]. Updates after 2003 are not included. Data on magnetic moments derive mostly from a 2001 preprint by Stone; see [45]. Nu-Dat 2 [[12]] provided the the excited energy levels and additional reference data to validate various data in [45].

Lynn Bowen corrected a bad number on the life time of helium with another proton or neutron added, because I stupidly misread ys (yoctosecond) in a reference as y (year). Very embarrassing, especially as I was amazed by the number.

The discussion of the Born series follows [25].

The brief description of quantum field theory and the quantization of the electromagnetic field is mostly from Wikipedia, [[21]], with a bit of fill-in from Yariv [52], Feynman [18], Kittel [29], and citizendium [[2]]. The example on field operators is an exercise from Srednicki [43], whose solution was posted online by a TA of Joe Polchinski from UCSB.

Acknowledgments for specific items are not listed here if a citation is given in the text, or if, as far as I know, the argument is standard theory. This is a text book, not a research paper or historical note. But if a reference is appropriate somewhere, let me know.

Grammatical and spelling errors have been pointed out by Ernesto Bosque, Eric Eros, Tag Jong Lee, Alastair McDonald, Samuel Rustan, Dan Schmidt, Mark Vanderlaan, Ramaswami Sastry Vedamm, Mikas Vengris, Rob Vossen, and Ed Williams. I will try to keep changing "therefor" into "therefore," and "send" into "sent", but they do keep sneaking in.

Thank you all.

Comments and Feedback

If you find an error, please let me know. There seems to be an unending supply of them. As one author described it brilliantly, "the hand is still writing though the brain has long since disengaged."

Also let me know if you find points that are unclear to the intended readership, mechanical engineering graduate students with a typical exposure to mathematics and physics, or equivalent. Every section, except a few explicitly marked as requiring advanced linear algebra, should be understandable by anyone with a good knowledge of calculus and undergraduate physics.

The same for sections that cannot be understood without delving back into earlier material. All within reason of course. If you pick a random starting word somewhere in the book and start reading from there, you most likely will be completely lost. But sections are intended to be fairly self-contained, and you should be able read one without backing up through all of the text.

General editorial comments are also welcome. I'll skip the philosophical discussions. I am an engineer.

Feedback can be e-mailed to me at quantum@dommelen.net.

This is a living document. I am still adding things here and there, and fixing various mistakes and doubtful phrasing. Even before every comma is perfect, I think the document can be of value to people looking for an easy-to-read introduction to quantum mechanics at a calculus level. So I am treating it as software, with version numbers indicating the level of confidence I have in it all.

Part I

Special Relativity

Chapter 1

Special Relativity [Draft]

Abstract

This first chapter reviews the theory of special relativity. It can be skipped if desired. Special relativity is not needed to understand the discussion of quantum mechanics in the remainder of this book. However, some parts of this chapter might provide a deeper understanding and justification for some of the issues that will come up in quantum mechanics.

The main reason for this chapter is that the book can be used as a review and expansion of typical courses on "Modern Physics." Such classes always cover relativity. While relativity is nowhere as important as basic quantum mechanics, it does have that "Einstein mystique" that is great at parties.

The chapter starts with an overview of the key ideas of relativity. This is material that is typically covered in modern physics classes. Subsequent sections provide more advanced explanations of the various ideas of special relativity.

1.1 Overview of Relativity

1.1.1 A note on the history of the theory

Special relativity is commonly attributed to Albert Einstein's 1905 papers. That is certainly justifiable. However, Einstein swiped the big ideas of relativity from Henri Poincaré, (who developed and named the principle of relativity in 1895 and a mass-energy relation in 1900), without giving him any credit or even mentioning his name.

He may also have swiped the underlying mathematics he used from Lorentz, (who is mentioned, but not in connection with the Lorentz transformation.)

However, in case of Lorentz, it is possible to believe that Einstein was unaware of his earlier work, if you are so trusting. Before you do, it must be pointed out that a review of Lorentz' 1904 work appeared in the second half of February 1905 in *Beiblätter zu den Annalen der Physik*. Einstein was well aware of that journal, since he wrote 21 reviews for it himself in 1905. Several of these were in the very next issue after the one with the Lorentz review, in the first half of March. Einstein's first paper on relativity was received June 30 1905 and published September 26 in *Annalen der Physik*. Einstein had been regularly writing papers for *Annalen der Physik* since 1901. You do the math.

In case of Poincaré, it is known that Einstein and a friend pored over Poincaré's 1902 book "Science and Hypothesis." In fact the friend noted that it kept them "breathless for weeks on end." So Einstein cannot possibly have been unaware of Poincaré's work.

However, Einstein should not just be blamed for his boldness in swiping most of the ideas in his paper from then more famous authors, but also be commended for his boldness in completely abandoning the basic premises of Newtonian mechanics, where earlier authors wavered.

It should also be noted that *general* relativity can surely be credited to Einstein fair and square. But he was a lot less hungry then. And had a lot more false starts. (There is a possibility that the mathematician Hilbert may have some partial claim on completing general relativity, but it is clearly Einstein who developed it. In fact, Hilbert wrote in one paper that his differential equations seemed to agree with the "magnificent theory of general relativity established by Einstein in his later papers." Clearly then, Hilbert himself agreed that Einstein established general relativity.)

1.1.2 The mass-energy relation

The most important result of relativity for the rest of this book is without doubt Einstein's famous relation $E = mc^2$. Here E is energy, m mass, and c the speed of light. (A very limited version of this relation was given before Einstein by Poincaré.)

The relation implies that the kinetic energy of a particle is not $\frac{1}{2}mv^2$, with m the mass and v the velocity, as Newtonian physics would have it. Instead the kinetic energy is the difference between the energy $m_v c^2$ based on the mass m_v of the particle in motion and the energy mc^2 based on the mass m of the same particle at rest. According to special relativity the mass in motion is related to the mass at rest as

$$m_v = \frac{m}{\sqrt{1-(v/c)^2}} \tag{1.1}$$

Therefore the true kinetic energy can be written as

$$T = \frac{m}{\sqrt{1-(v/c)^2}}c^2 - mc^2$$

For velocities small compared to the tremendous speed of light, this is equivalent to the classical $\frac{1}{2}mv^2$. That can be seen from Taylor series expansion of the square root. But when the particle speed approaches the speed of light, the above expression implies that the kinetic energy approaches infinity. Since there is no infinite supply of energy, the velocity of a material object must always remain less than the speed of light.

The photons of electromagnetic radiation, (which includes radio waves, microwaves, light, x-rays, gamma rays, etcetera), do travel at the speed of light through a vacuum. However, the only reason that they can do so is because they have zero rest mass m. There is no way that photons in vacuum can be brought to a halt, or even slowed down, because there would be nothing left.

If the kinetic energy is the difference between $m_v c^2$ and mc^2, then both of these terms must have units of energy. That does of course not prove that each term is a physically meaningful energy by itself. But it does look plausible. It suggests that a particle at rest still has a "rest mass energy" mc^2 left. And so it turns out to be. For example, an electron and a positron can completely annihilate each other, releasing their rest mass energies as two photons that fly apart in opposite directions. Similarly, a photon of electromagnetic radiation with enough energy can create an electron-positron pair out of nothing. (This does require that a heavy nucleus is around to absorb the photon's linear momentum without absorbing too much of its energy; otherwise it would violate momentum conservation.) Perhaps more importantly for engineering applications, the energy released in nuclear reactions is produced by a reduction in the rest masses of the nuclei involved.

Quantum mechanics does not use the speed v of a particle, but its momentum $p = m_v v$. In those terms the total energy, kinetic plus rest mass energy, can be rewritten as

$$\boxed{E = T + mc^2 = \sqrt{(mc^2)^2 + p^2c^2}} \tag{1.2}$$

This expression is readily checked by substituting in for p, then for m_v, and cleaning up.

1.1.3 The universal speed of light

The key discovery of relativity is that the observed speed of light through vacuum is the same regardless of how fast you are traveling. One historical step that led to this discovery was a famous experiment by Michelson & Morley. In simplified terms, Michelson & Morley tried to determine the absolute speed of the earth through space by "horse-racing" it against light. If a passenger jet airplane flies at three quarters of the speed of sound, then sound waves going in the same direction as the plane only have a speed advantage of one quarter of the speed of sound over the plane. Seen from inside the plane, that sound seems to move away from it at only a quarter of the normal speed of sound.

Essentially, Michelson & Morley reasoned that the speed of the earth could similarly be observed by measuring how much it reduces the apparent speed of light moving in the same direction through a vacuum. But it proved that the motion of the earth produced no reduction in the apparent speed of light whatsoever. It is as if you are racing a fast horse, but regardless of how fast you are going, you do not reduce the velocity difference any more than if you would just stop your horse and have a drink.

The simplest explanation would be that earth is at rest compared to the universe. But that cannot possibly be true. Earth is a minor planet in an outer arm of the galaxy. And earth moves around the sun once a year. Obviously, the entire universe could not possibly follow that noninertial motion.

So how come that earth still seems to be at rest compared to light waves moving through vacuum? You can think up a hundred excuses. In particular, the sound *inside* a plane does not seem to move any slower in the direction of motion. But of course, sound is transmitted by real air molecules that can be trapped inside a plane by well established mechanisms. It is not transmitted through empty space like light.

But then, at the time of the Michelson & Morley experiment the prevailing theory was that light *did* move through some hypothetical medium. This made-up medium was called the "ether." It was supposedly maybe dragged along by the earth, or maybe dragged along a bit, or maybe not after all. In fact, Michelson & Morley were really trying to decide how much it was being dragged along. Looking back, it seems self-evident that this ether was an unsubstantiated theory full of holes. But at the time most scientists took it very seriously.

The results of the Michelson & Morley experiment and others upped the ante. To what could reasonably be taken to be experimental error, the earth did not seem to move relative to light waves in vacuum. So in 1895 Poincaré reasoned that experiments like the one of Michelson & Morley suggested that it is impossible to detect absolute motion. In 1900 he proposed the "Principle of Relative Motion." It proposed that the laws of movement would be the same in all coordinate systems regardless of their velocity, as long as they are not accelerating. In 1902, in the book read by Einstein, he discussed philosophical assessments on the relativity of space, time, and simultaneity, and the idea that a violation of the relativity principle can never be detected. In 1904 he called it

> "The principle of relativity, according to which the laws of physical phenomena must be the same for a stationary observer as for one carried along in a uniform motion of translation, so that we have no means, and can have none, of determining whether or not we are being carried along in such a motion."

In short, if two observers are moving at different, but constant speeds, it is impossible to say which one, if any, is at rest. The laws of physics observed by the two observers are exactly the same. In particular, the

moving observers see the same speed of light regardless of their different physical motion.

(Do note however that if an observer is accelerating or spinning around, that can be determined through the generated inertia forces. Not all motion is relative. Just an important subset of it.)

A couple of additional historical notes may be appropriate. Quite a number of historians of science argue that Poincaré did not "really" propose relativity, because he continued to use the ether in various computations afterwards. This argument is unjustified. To this very day, the overwhelming majority of physicists and engineers still use Newtonian physics in their computations. That does not mean that these physicists and engineers do not believe in special relativity. It means that they find doing the Newtonian computation a lot easier, and it gives the right answer for their applications. Similarly, Poincaré himself clearly stated that he still considered the ether a "convenient hypothesis." There were well established procedures for computing such things as the propagation of light in moving media using an assumed ether that had been well verified by experiment.

A more interesting hypothesis advanced by historians is that Einstein may have been more inclined to do away with the ether from the start than other physicists. The concept of the ether was without doubt significantly motivated by the propagation of other types of waves like sound waves and water waves. In such waves, there is some material substance that performs a wave motion. Unlike waves, however, particles of all kinds readily propagate through empty space; they do not depend on a separate medium that waves. That did not seem relevant to light, because its wave-like nature had been well established. But in quantum mechanics, the complementary nature of light as particles called photons was emerging. And Einstein may have been more comfortable with the quantum mechanical concept of light than most at the time. He was a major developer of it.

1.1.4 Disagreements about space and time

At first, it may not seem such a big deal that the speed of light is the same regardless of the motion of the observer. But when this notion is examined in some more detail, it leads to some very counter-intuitive conclusions.

It turns out that if observers are in motion compared to each other, they will unavoidably disagree about such things as spatial distances and the time that things take. Often, different observers cannot even agree on which of two physical events takes place earlier than the other. Assuming that they determine the times correctly in their own coordinate system, they will come up with different answers.

Self-evidently, if observers cannot even agree on which of two events happened first, then an absolute time scale that everyone can agree on is not possible

either. And neither is a system of absolute spatial coordinates that everybody can agree upon.



Figure 1.1: Different views of the same experiment. Left is the view of observers on the planets. Right is the view of an alien space ship.

Consider a basic thought experiment. A thought experiment is an experiment that should in principle be possible, but you do not want to be in charge of actually doing it. Suppose that the planets Venus and Mars happen to be at opposite sides of earth, and at roughly the same distance from it. The left side of figure 1.1 shows the basic idea. Experimenters on earth flash simultaneous light waves at each planet. Since Venus happens to be a bit closer than Mars, the light hits Venus first. All very straightforward. Observers on Venus and Mars would agree completely with observers on earth that Venus got hit first. They also agree with earth about how many minutes it took for the light to hit Venus and Mars.

To be sure, the planets move with speeds of the order of 100,000 mph relative to one another. But that speed, while very large in human terms, is so small compared to the tremendous speed of light that it can be ignored. For the purposes of this discussion, it can be assumed that the planets are at rest relative to earth.

Next assume that a space ship with aliens was just passing by and watched the whole thing, like in the right half of figure 1.1. As seen by observers on earth, the aliens are moving to the right with half the speed of light. However, the aliens can argue that it is they that are at rest, and that the three planets are moving towards the left with half the speed of light. According to the principle of relativity, both points of view are equally valid. There is nothing that can show whether the space ship or the planets are at rest, or neither one.

In particular, the speed of the light waves that the aliens observe is identical to the speed that earth sees. But now note that as far as the aliens are concerned, Venus moves with half the speed of light *away* from its incoming light wave. Of course, that significantly increases the time that the light needs to reach Venus. On the other hand, the aliens see Mars moving at half the speed of light *towards* its incoming light wave. That roughly halves the time needed for the light wave to hit Mars. In short, unlike earth, the aliens observe that the light hits Mars a lot earlier than it hits Venus.

That example demonstrates that observers in relative motion disagree about the time difference between events occurring at different locations. Worse, even if two events happen right in the hands of one of the observers, the observers

will disagree about how long the entire thing takes. In that case, the observer compared to which the location of the events is in motion will think that it takes longer. This is called "time-dilation." The time difference between two events slows down according to

$$\boxed{\Delta t_v = \frac{\Delta t_0}{\sqrt{1-(v/c)^2}}} \tag{1.3}$$

Here Δt_0 is shorthand for the time difference between the two events as seen by an observer compared to whom the two events occur at the same location. Similarly Δt_v is the time difference between the two events as perceived by an observer compared to whom the location of the events is moving at speed v.

An "event" can be anything with an unambiguous physical meaning, like when the hands of a clock reach a certain position. So clocks are found to run slow when they are in motion compared to the observer. The best current clocks are accurate enough to directly measure this effect at human-scale speeds, as low as 20 mph. But relativity has already been verified in myriad other ways. The time is long gone that serious scientists still doubted the conclusions of relativity.

As a more practical example, cosmic rays can create radioactive particles in the upper atmosphere that survive long enough to reach the surface of the earth. The surprising thing is that at rest in a laboratory these same particles would not survive that long by far. The particles created by cosmic rays have extremely high speed when seen by observers standing on earth. That slows down the decay process due to time dilation.

Which of course raises the question: should then not an observer moving along with one such particle observe that the particle does not reach the earth? The answer is no; relativity maintains a single reality; a particle either reaches the earth or not, regardless of who is doing the observing. It is quantum mechanics, not relativity, that does away with a single reality. The observer moving with the particle observes that the particle reaches the earth, not because the particle seems to last longer than usual, but because the distance to travel to the surface of the earth has become much shorter! This is called "Lorentz-Fitzgerald contraction."

For the observer moving with the particle, it seems that the entire earth system, including the atmosphere, is in motion with almost the speed of light. The size of objects in motion seems to contract in the direction of the motion according to

$$\boxed{\Delta x_v = \Delta x_0 \sqrt{1-(v/c)^2}} \tag{1.4}$$

Here the x-axis is taken to be in the direction of motion. Also Δx_0 is the distance in the x-direction between any two points as seen by an observer compared to whom the points are at rest. Similarly, Δx_v is the distance as seen by an observer compared to whom the points are moving with speed v in the x-direction.

In short, for the observer standing on earth, the particle reaches earth because its motion slows down the decay process by a factor $1/\sqrt{1-(v/c)^2}$. For the observer moving along with the particle, the particle reaches earth because the distance to travel to the surface of the earth has become shorter by exactly that same factor. The reciprocal square root is called the "Lorentz factor."

Lorentz-Fitzgerald contraction is also evident in how the aliens see the planets in figure 1.1. But note that the difference in the wave lengths of the light waves is not a simple matter of Lorentz-Fitzgerald contraction. The light waves are in motion compared to both observers, so Lorentz-Fitzgerald contraction simply does not apply.

The correct equation that governs the difference in observed wave length λ of the light, and the corresponding difference in observed frequency ω, is

$$\boxed{\lambda_v = \lambda_0 \sqrt{\frac{1+(v/c)}{1-(v/c)}} \qquad \omega_v = \omega_0 \sqrt{\frac{1-(v/c)}{1+(v/c)}}} \tag{1.5}$$

Here the subscript 0 stands for the emitter of the light, and subscript v for an observer moving with speed v away from the emitter. If the observer moves towards the emitter, v is negative. (To be true, the formulae above apply whether the observer 0 is emitting the light or not. But in most practical applications, observer 0 is indeed the emitter.)

In terms of the example figure 1.1, 0 indicates the emitter earth, and v indicates the aliens observing the radiation. If the aliens are still to the left of earth, they are still closing in on it and v is negative. Then the formulae above say that the wave length seen by the aliens is shorter than the one seen by earth. Also, the frequency seen by the aliens is higher than the one seen by earth, and so is the energy of the light. When the aliens get to the right of earth, they are moving away from it. That makes v positive, and the light from earth that is reaching them now seems to be of longer wave length, of lower frequency, and less energetic. These changes are referred to as "Doppler shifts."

One related effect is cosmic redshift. The entire universe is expanding. As a result, far away galaxies move away from us at extremely high speeds. That causes wave length shifts; the radiation emitted or absorbed by various excited atoms in these galaxies appears to us to have wave lengths that are too long. The received wave lengths are longer than those that these same atoms would emit or absorb on earth. In particular, the colors of visible light are shifted towards the red side of the spectrum. To observers in the galaxies themselves, however, the colors would look perfectly fine.

Note that the cosmic redshift can only qualitatively be understood from the formulae above. It is more accurate to say that the photons traveling to us from remote galaxies get stretched due to the expansion of the universe. The cosmic redshift is not due to the motion of the galaxies *through* space, but due to the motion *of* space itself. If the expansion of space is rephrased in terms of

a relative velocity of the galaxies compared to us, that velocity can exceed the speed of light. That would produce nonsense in the formulae above. Objects cannot move faster than the speed of light through space, but the velocity of different regions of space compared to each other can exceed the speed of light.

Returning to the normal Doppler shift, the changes in wave length are not directly due to Lorentz-Fitzgerald contraction. Instead, they can in part be attributed to time dilation. In figure 1.1 both the aliens and earth can deduce the wave length from how frequently the peaks of the wave leave the emitter earth. But in doing so, one source of disagreement is time dilation. Since earth is in motion compared to the aliens, the aliens think that the peaks leave earth less frequently than earth does. In addition, the aliens and earth disagree about the relative velocity between the light waves and earth. Earth thinks that the light waves leave with the speed of light relative to earth. The aliens also think that the light waves travel with the speed of light, but in addition they see earth moving towards the left with half the speed of light. Combine the two effects, for arbitrary velocity of the aliens, and the relation between the wave lengths is as given above. Further, since the speed of light is the same for both earth and aliens, the observed frequency of the light is inversely proportional to the observed wave length.

1.2 The Lorentz Transformation

The "Lorentz transformation" describes how measurements of the position and time of events change from one observer to the next. It includes Lorentz-Fitzgerald contraction and time dilation as special cases.

1.2.1 The transformation formulae

This subsection explains how the position and time coordinates of events differ from one observer to the next.

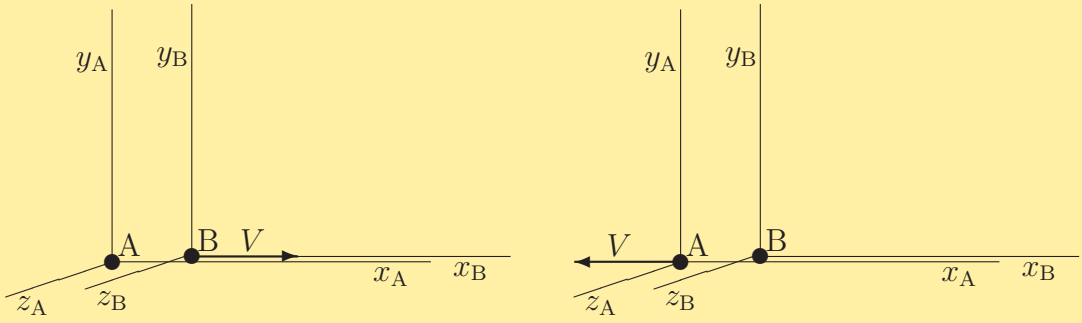


Figure 1.2: Coordinate systems for the Lorentz transformation.

Consider two observers A and B that are in motion compared to each other with a relative speed V. To make things as simple as possible, it will be assumed that the relative motion is along the line through the two observers,

As the left side of figure 1.2 shows, observer A can believe herself to be at rest and see observer B moving away from her at speed V; similarly, observer B can believe himself to be at rest and see observer A moving away from him at speed V, as in the right side of the figure. The principle of relativity says that both views are equally valid; there is no physical measurement that can find a fundamental difference between the two observers. That implies that both observers must agree on the same magnitude of the relative velocity V between them. And it implies that they need to agree on the speed c that light moves.

It will further be assumed that both observers use coordinate systems with themselves at the origin to describe the locations and times of events. In addition, they both take their x axes along the line of their relative motion. They also align their y and z axes. And they both define time to be zero at the instant that they meet.

In that case the Lorentz transformation says that the relation between positions and times of events as perceived by the two observers is, {D.4}:

$$\boxed{ct_B = \frac{ct_A - (V/c)x_A}{\sqrt{1-(V/c)^2}} \qquad x_B = \frac{x_A - (V/c)ct_A}{\sqrt{1-(V/c)^2}} \qquad y_B = y_A \qquad z_B = z_A} \tag{1.6}$$

To get the transformation of the coordinates of B into those of A, just swap A and B and replace V by $-V$. Indeed, if observer B is moving in the positive x-direction with speed V compared to observer A, then observer A is moving in the negative x-direction with speed V compared to observer B, as in figure 1.2. In the limit that the speed of light c becomes infinite, the Lorentz transformation becomes the nonrelativistic "Galilean transformation" in which t_B is simply t_A and $x_B = x_A - Vt$, i.e. x_B equals x_A except for a shift of magnitude Vt.

The made assumptions are that A and B are at the origin of their coordinate system. And that their spatial coordinate systems are aligned. And that their relative motion is along the x axes. And that they take the zero of time to be the instant that they meet. These simplifying assumptions may look very restrictive. But they are not. A different observer A' at rest relative to A can still use any coordinate system he wants, with any arbitrary orientation, origin, and zero of time. Since A' is at rest relative to A, the two fundamentally agree about space and time. So whatever coordinates and times A' uses for events are easily converted to those that A uses in the classical way, {A.3}. Similarly an observer B' at rest compared to B can still use any arbitrary coordinate system that she wants. The coordinates and times of events observed by the arbitrary observers A' and B' can then be related to each other in stages. First relate the coordinates of A' to those of A in the classical way. Next use the Lorentz transformation as given above to relate those to the coordinates of B.

Then relate those in the classical way to the coordinates of B'. In this way, the coordinates and times of any two observers in relative motion to each other, using arbitrary coordinate systems and zeros of time, can be related. The simple Lorentz transformation above describes the *nontrivial* part of how the observations of different observers relate.

Time dilation is one special case of the Lorentz transformation. Assume that two events 1 and 2 happen at the same location $x_{\rm A}, y_{\rm A}, z_{\rm A}$ in system A. Then the first Lorentz transformation formula (1.6) gives

$$t_{2,\rm B} - t_{1,\rm B} = \frac{t_{2,\rm A} - t_{1,\rm A}}{\sqrt{1-(V/c)^2}}$$

So observer B finds that the time difference between the events is larger. The same is of course true vice-versa, just use the inverse formulae.

Lorentz-Fitzgerald contraction is another special case of the Lorentz transformation. Assume that two stationary locations in system B are apart by a distance $x_{2,B} - x_{1,B}$ in the direction of relative motion. The second Lorentz transformation formula (1.6) then says how far these points are apart in system A at any given time $t_{\rm A}$:

$$x_{2,\rm B} - x_{1,\rm B} = \frac{x_{2,\rm A} - x_{1,\rm A}}{\sqrt{1-(V/c)^2}}$$

Taking the square root to the other side gives the contraction.

As a result of the Lorentz transformation, measured velocities are related as

$$v_{x,B} = \frac{v_{x,A} - V}{1-(V/c^2)v_{x,A}} \quad v_{y,B} = \frac{v_{y,A}\sqrt{1-(V/c)^2}}{1-(V/c^2)v_{x,A}} \quad v_{z,B} = \frac{v_{z,A}\sqrt{1-(V/c)^2}}{1-(V/c^2)v_{x,A}} \tag{1.7}$$

Note that v_x, v_y, v_z refer here to the perceived velocity components of some moving object; they are not components of the velocity difference V between the coordinate systems.

1.2.2 Proper time and distance

In classical Newtonian mechanics, time is absolute. All observers agree about the difference in time Δt between any two events:

nonrelativistic: Δt is independent of the observer

The time difference is an "invariant;" it is the same for all observers.

All observers, regardless of how their spatial coordinate systems are oriented, also agree over the distance $|\Delta\vec{r}|$ between two events that occur at the same time:

nonrelativistic: $|\Delta\vec{r}|$ is independent of the observer if $\Delta t = 0$

Here the distance between any two points 1 and 2 is found as

$$|\Delta\vec{r}| \equiv \sqrt{(\Delta\vec{r})\cdot(\Delta\vec{r})} = \sqrt{(\Delta x)^2+(\Delta y)^2+(\Delta z)^2} \qquad \Delta\vec{r} \equiv \vec{r}_2 - \vec{r}_1$$

The fact that the distance may be expressed as a square root of the sum of the square components is known as the "Pythagorean theorem."

Relativity messes all these things up big time. As time dilation shows, the time between events now depends on who is doing the observing. And as Lorentz-Fitzgerald contraction shows, distances now depend on who is doing the observing. For example, consider a moving ticking clock. Not only do different observers disagree over the distance $|\Delta\vec{r}|$ traveled between ticks, (as they would do nonrelativistically), but they also disagree about the time difference Δt between ticks, (which they would not do nonrelativistically).

However, there is one thing that all observers can agree on. They do agree on how much time between ticks an observer moving along with the clock would measure. That time difference is called the "proper time" difference. (The word proper is a wrongly translated French "propre," which here means "own." So proper time really means the clock's own time.) The time difference Δt that an observer actually perceives is longer than the proper time difference Δt_0 due to the time dilation:

$$\Delta t = \frac{\Delta t_0}{\sqrt{1-(v/c)^2}}$$

Here v is the velocity of the clock as perceived by the observer.

To clean this up, take the square root to the other side and write v as the distance $|\Delta\vec{r}|$ traveled by the clock divided by Δt. That gives the proper time difference Δt_0 between two events, like the ticks of a clock here, as

$$\boxed{\Delta t_0 = \Delta t\sqrt{1-\frac{(\Delta x)^2+(\Delta y)^2+(\Delta z)^2}{(c\Delta t)^2}}} \tag{1.8}$$

The numerator in the ratio is the square of the distance between the events.

Note however that the proper time difference is imaginary if the quantity under the square root is negative. For example, if an observer perceives two events as happening simultaneously at two different locations, then the proper time difference between those two events is imaginary. To avoid dealing with complex numbers, it is then more convenient to define the "proper distance" Δs between the two events as

$$\boxed{\Delta s = \sqrt{(\Delta x)^2+(\Delta y)^2+(\Delta z)^2-(c\Delta t)^2}} \tag{1.9}$$

Note that this is the ordinary distance between the two events if they are at the same time, i.e. $\Delta t = 0$. The proper distance is different from the proper time difference by a factor $\sqrt{-c^2}$. Because of the minus sign under the square

root, this factor is imaginary. As a result, Δs is imaginary if Δt_0 is real and vice-versa.

All observers agree about the values of the proper time difference Δt_0 and the proper distance Δs for any two given events.

Physicists define the *square* of the proper distance to be the "space-time interval" I. The term is obviously confusing, as a dictionary defines an interval as a difference in time or space, not as the square of such a difference. To add more confusion, some physicists change sign in the definition, and others divide by the square speed of light. And some rightly define the interval to be Δs without the square, unfortunately causing still more confusion.

If the interval, defined as $(\Delta s)^2$, is positive, then the proper distance Δs between the two events is real. Such an interval is called "space-like." On the other hand, if the interval is negative, then the proper distance is imaginary. In that case it is the proper time difference between the events that is real. Such an interval is called "time-like."

If the proper time difference is real, the earlier event can affect, or even cause, the later event. If the proper time difference is imaginary however, then the effects of either event cannot reach the other event even if traveling at the speed of light. It follows that the sign of the interval is directly related to "causality," to what can cause what. Since all observers agree about the value of the proper time difference, they all agree about what can cause what.

For small differences in time and location, all differences Δ above become differentials d.

1.2.3 Subluminal and superluminal effects

Suppose you stepped off the curb at the wrong moment and are now in the hospital. The pain is agonizing, so you contact one of the telecommunications microchips buzzing in the sky overhead. These chips are capable of sending out a "superluminal" beam; a beam that propagates with a speed greater than the speed of light. The factor with which the speed of the beam exceeds the speed of light is called the "warp factor" w. A beam with a high warp factor is great for rapid communication with space ships at distant locations in the solar system and beyond. A beam with a warp factor of 10 allows ten times quicker communication than those old-fashioned radio waves that propagate at the speed of light. And these chips have other very helpful uses, like for your predicament.

You select a microchip that is moving at high speed away from the location where the accident occurred. The microchip sends out its superluminal beam. In its coordinate system, the beam reaches the location of the accident at a time t_m, at which time the beam has traveled a distance x_m equal to wct_m. According to the Lorentz transformation (1.6), in the coordinate system fixed to the earth, the beam reaches the location of the accident at a position and

time equal to

$$t = \frac{1-(wV/c)}{\sqrt{1-(V/c)^2}}t_m \qquad x = \frac{wc-V}{\sqrt{1-(V/c)^2}}t_m$$

Because of the high speed V of the microchip and the additional warp factor, the time that the beam reaches the location of the accident is negative; the beam has entered into the past. Not far enough in the past however, so another microchip picks up the message and beams it back, achieving another reduction in time. After a few more bounces, the message is beamed to your cell phone. It reaches you just when you are about to step off the curb. The message will warn you of the approaching car, but it is not really needed. The mere distraction of your buzzing cell phone causes you to pause for just a second, and the car rushes past safely. So the accident never happens; you are no longer in agony in the hospital, but on your Bermuda vacation as planned. And these microchips are great for investing in the stock market too.

Sounds good, does it not? Unfortunately, there is a hitch. Physicists refuse to work on the underlying physics to enable this technology. They claim it will not be workable, since it will force them to think up answers to tough questions like: "if you did not end up in the hospital after all, then why did you still send the message?" Until they change their mind, our reality will be that observable matter or radiation cannot propagate faster than the speed of light.

Therefore, manipulating the past is not possible. An event can only affect later events. Even more specifically, an event can only affect a later event if the location of that later event is sufficiently close that it can be reached with a speed of no more than the speed of light. A look at the definition of the proper time interval then shows that this means that the proper time interval between the events must be real, or "time-like." And while different observers may disagree about the location and time of the events, they all agree about the proper time interval. So all observers, regardless of their velocity, agree on whether an event can affect another event. And they also all agree on which event is the earlier one, because before the time interval Δt could change sign for some observer speeds, it would have to pass through zero. It cannot, because it must be the same for all observers. Relativity maintains a single reality, even though observers may disagree about precise times and locations.

A more visual interpretation of those concepts can also be given. Imagine a hypothetical spherical wave front spreading out from the earlier event with the speed of light. Then a later event can be affected by the earlier event only if that later event is within or on that spherical wave front. If you restrict attention to events in the x, y plane, you can use the z-coordinate to plot the values of time. In such a plot, the expanding circular wave front becomes a cone, called the "light-cone." Only events within this light cone can be affected. Similarly in three dimensions and time, an event can only be affected if it is within the light

cone in four-dimensional space-time. But of course, a cone in four dimensions is hard to visualize geometrically.

1.2.4 Four-vectors

The Lorentz transformation mixes up the space and time coordinates badly. In relativity, it is therefore best to think of the spatial coordinates and time as coordinates in a four-dimensional "space-time."

Since you would surely like all components in a vector to have the same units, you probably want to multiply time by the speed of light, because ct has units of length. So the four-dimensional "position vector" can logically be defined to be (ct, x, y, z); ct is the "zeroth" component of the vector where x, y, and z are components number 1, 2, and 3 as usual. This four-dimensional position vector will be indicated by

$$\vec{\imath r} \equiv \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ r_3 \end{pmatrix} \tag{1.10}$$

The hook on the arrow indicates that time has been hooked into it.

How about the important dot product between vectors? In three dimensional space this produces such important quantities as the length of vectors and the angle between vectors. Moreover, the dot product between two vectors is the same regardless of the orientation of the coordinate system in which it is viewed.

It turns out that the proper way to define the dot product for four-vectors reverses the sign of the contribution of the time components:

$$\vec{\imath r}_1 \cdot \vec{\imath r}_2 \equiv -c^2 t_1 t_2 + x_1 x_2 + y_1 y_2 + z_1 z_2 \tag{1.11}$$

It can be checked by simple substitution that the Lorentz transformation (1.6) preserves this dot product. In more expensive words, this "inner product" is "invariant under the Lorentz transformation." Different observers may disagree about the individual components of four-vectors, but not about their dot products.

The difference between the four-vector positions of two events has a "proper length" equal to the proper distance between the events

$$\Delta s = \sqrt{(\Delta \vec{\imath r}) \cdot (\Delta \vec{\imath r})} \tag{1.12}$$

So, the fact that all observers agree about proper distance can be seen as a consequence of the fact that they all agree about dot products.

It should be pointed out that many physicist reverse the sign of the *spatial* components instead of the time in their inner product. Obviously, this is completely inconsistent with the nonrelativistic analysis, which is often still a valid

approximation. And this inconsistent sign convention seems to be becoming the dominant one too. Count on physicists to argue for more than a century about a sign convention and end up getting it all wrong in the end. One very notable exception is [49]; you can see why he would end up with a Nobel Prize in physics.

Some physicists also like to point out that if time is replaced by it, then the above dot product becomes the normal one. The Lorentz transformation can then be considered as a mere rotation of the coordinate system in this four-dimensional space-time. Gee, thanks physicists! This will be very helpful when examining what happens in universes in which time is imaginary, unlike our own universe, in which it is real. The good thing you can say about these physicists is that they define the dot product the right way: the i^2 takes care of the minus sign on the zeroth component.

Returning to our own universe, the proper length of a four-vector can be imaginary, and a zero proper length does not imply that the four-vector is zero as it does in normal three-dimensional space. In fact, a zero proper length merely indicates that it requires motion at the speed of light to go from the start point of the four-vector to its end point.

1.2.5 Index notation

The notations used in the previous subsection are not standard. In literature, you will almost invariably find the four-vectors and the Lorentz transform written out in index notation. Fortunately, it does not require courses in linear algebra and tensor algebra to make some basic sense out of it.

First of all, in the nonrelativistic case position vectors are normally indicated by $\vec{r}$. The three components of this vector are commonly indicated by x, y, and z, or using indices, by r_1, r_2, and r_3. To handle space-time, physicists do not simply add a zeroth component r_0 equal to ct. That would make the meaning too easy to guess. Instead physicists like to indicate the components of four-vectors by x^0, x^1, x^2, x^3. It is harder to guess correctly what that means, especially since the letter x is already greatly over-used as it is. A generic component may be denoted as x^μ. An entire four-vector can then be indicated by $\{x^\mu\}$ where the brackets indicate the set of all four components. Needless to say, most physicists forget about the brackets, because using a component where a vector is required can have hilarious consequences.

In short,

$$\begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \equiv \vec{r} \equiv \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ r_3 \end{pmatrix} \qquad \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \equiv \{x^\mu\} \equiv \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix}$$

shows this book's common sense notation to the left and the index notation commonly used in physics to the right.

Recall now the Lorentz transformation (1.6). It described the relationship between the positions and times of events as observed by two different observers A and B. These observers were in motion compared to each other with a relative speed V. Physicists like to put the coefficients of such a Lorentz transformation into a table, as follows:

$$\Lambda \equiv \begin{pmatrix} \lambda^0{}_0 & \lambda^0{}_1 & \lambda^0{}_2 & \lambda^0{}_3 \\ \lambda^1{}_0 & \lambda^1{}_1 & \lambda^1{}_2 & \lambda^1{}_3 \\ \lambda^2{}_0 & \lambda^2{}_1 & \lambda^2{}_2 & \lambda^2{}_3 \\ \lambda^3{}_0 & \lambda^3{}_1 & \lambda^3{}_2 & \lambda^3{}_3 \end{pmatrix} \equiv \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.13}$$

where

$$\gamma \equiv \frac{1}{\sqrt{1-(V/c)^2}} \qquad \beta \equiv \frac{V}{c} \qquad \gamma^2 - \beta^2\gamma^2 = 1$$

A table like Λ is called a "matrix" or "second-order tensor." The individual entries in the matrix are indicated by $\lambda^\mu{}_\nu$ where μ is the number of the row and ν the number of the column. Note also the convention of showing the first index as a superscript. That is a tensor algebra convention. In linear algebra, you normally make all indices subscripts.

(Different sources use different letters for the Lorentz matrix and its entries. Some common examples are $\Lambda^\mu{}_\nu$ and $a^\mu{}_\nu$. The name "Lorentz" starts with L and the Greek letter for L is Λ. And Lorentz was Dutch, which makes him a European just like the Greek. Therefore Λ is a good choice for the name of the Lorentz matrix, and Λ or lower case λ for the entries of the matrix. An L for the matrix and l for its entries would be just too easy to guess. Also, λ is the standard name for the eigenvalues of matrices and Λ for the matrix of those eigenvalues. So there is some potential for hilarious confusion there. An "a" for the Lorentz matrix is good too: the name "Lorentz" consists of roman letters and a is the first letter of the roman alphabet.)

The values of the entries $\lambda^\mu{}_\nu$ may vary. The ones shown in the final matrix in (1.13) above apply only in the simplest nontrivial case. In particular, they require that the relative motion of the observers is aligned with the x axes as in figure 1.2. If that is not the case, the values become a lot more messy.

In terms of the above notations, the Lorentz transformation (1.6) can be written as

$$x^\mu_{\rm B} = \sum_{\nu=0}^{3} \lambda^\mu{}_\nu x^\nu_{\rm A} \qquad \text{for all four values } \mu = 0, 1, 2, 3$$

That is obviously a lot more concise than (1.6). Some further shorthand notation is now used. In particular, the "Einstein summation convention" is to leave away

the summation symbol $\sum$. So, you will likely find the Lorentz transformation written more concisely as

$$x^\mu_{\rm B} = \lambda^\mu{}_\nu x^\nu_{\rm A}$$

Whenever an index like ν appears twice in an expression, summation over that index is to be understood. In other words, you yourself are supposed to mentally add back the missing summation over all four values of ν to the expression above. Also, if an index appears only once, like μ above, it is to be understood that the equation is valid for all four values of that index.

It should be noted that mathematicians call the matrix Λ the transformation matrix *from* B *to* A, even though it produces the coordinates *of* B *from* those of A. However, after you have read some more in this book, insane notation will no longer surprise you. Just that in this case it comes from mathematicians.

In understanding tensor algebra, it is essential to recognize one thing. It is that a quantity like a position differential transforms different from a quantity like a gradient:

$$\mathrm{d}x^\mu_{\rm B} = \frac{\partial x^\mu_{\rm B}}{\partial x^\nu_{\rm A}} \mathrm{d}x^\nu_{\rm A} \qquad \frac{\partial f}{\partial x^\mu_{\rm B}} = \frac{\partial f}{\partial x^\nu_{\rm A}} \frac{\partial x^\nu_{\rm A}}{\partial x^\mu_{\rm B}}$$

In the first expression, the partial derivatives are by definition the entries of the Lorentz matrix Λ,

$$\frac{\partial x^\mu_{\rm B}}{\partial x^\nu_{\rm A}} \equiv \lambda^\mu{}_\nu$$

In the second expression, the corresponding partial derivatives will be indicated by

$$\frac{\partial x^\mu_{\rm A}}{\partial x^\nu_{\rm B}} \equiv (\lambda^{-1})^\mu{}_\nu$$

The entries $(\lambda^{-1})^\mu{}_\nu$ form the so-called "inverse" Lorentz matrix Λ^{-1}. If the Lorentz transformation describes a transformation from an observer A to an observer B, then the inverse transformation describes the transformation from B to A.

Assuming that the Lorentz transformation matrix is the simple one to the right in (1.13), the inverse matrix Λ^{-1} looks exactly the same as Λ except that $-\beta$ gets replaced by $+\beta$. The reason is fairly simple. The quantity β is the velocity between the observers scaled with the speed of light. And the relative velocity of B seen by A is the opposite of the one of A seen by B, if their coordinate systems are aligned.

Consider now the reason why tensor analysis raises some indices. Physicists use a superscript index on a vector if it transforms using the normal Lorentz transformation matrix Λ. Such a vector is called a "contravariant" vector for reasons not worth describing. As an example, a position differential is a contravariant vector. So the components of a position differential are indicated by $\mathrm{d}x^\mu$ with a superscript index.

If a vector transforms using the inverse matrix Λ^{-1}, it is called a "covariant" vector. In that case subscript indices are used. For example, the gradient of a function f is a covariant vector. So a component $\partial f/\partial x^\mu$ is commonly indicated by $\partial_\mu f$.

Now suppose that you flip over the sign of the zeroth, time, component of a four-vector like a position or a position differential. It turns out that the resulting four-vector then transforms using the inverse Lorentz transformation matrix. That means that it has become a covariant vector. (You can easily verify this in case of the simple Lorentz transform above.) Therefore lower indices are used for the flipped-over vector:

$$\{x_\mu\} \equiv (-ct, x, y, z) \equiv (x_0, x_1, x_2, x_3)$$

The convention of showing covariant vectors as rows instead of columns comes from linear algebra. Tensor notation by itself does not have such a graphical interpretation.

Keep one important thing in mind though. If you flip the sign of a component of a vector, you get a fundamentally different vector. The vector $\{x_\mu\}$ is *not* just a somewhat different way to write the position four-vector $\{x^\mu\}$ of the space-time point that you are interested in. Now normally if you define some new vector that is different from a vector that you are already using, you change the name. For example, you might change the name from x to y or to $x^{\rm L}$ say. Tensor algebra does not do that. Therefore the golden rule is:

> *The names of tensors are only correct if the indices are at the right height.*

If you remember that, tensor algebra becomes a lot less confusing. The expression $\{x^\mu\}$ is only your space-time location named x if the index is a superscript as shown. The four-vector $\{x_\mu\}$ is simply a different animal. How do you know what is the right height? You just have to remember, you know.

Now consider two different contravariant four-vectors, call them $\{x^\mu\}$ and $\{y^\mu\}$. The dot product between these two four-vectors can be written as

$$x_\mu y^\mu$$

To see why, recall that since the index μ appears twice, summation over that index is understood. Also, the lowered index of x_μ indicates that the sign of the zeroth component is flipped over. That produces the required minus sign on the product of the time components in the dot product.

Note also from the above examples that summation indices appear once as a subscript and once as a superscript. That is characteristic of tensor algebra.

Addendum {A.4} gives a more extensive description of the most important tensor algebra formulae for those with a good knowledge of linear algebra.

1.2.6 Group property

The derivation of the Lorentz transformation as given earlier examined two observers A and B. But now assume that a third observer C is in motion compared to observer B. The coordinates of an event as perceived by observer C may then be computed from those of B using the corresponding Lorentz transformation, and the coordinates of B may in turn be computed from those of A using that Lorentz transformation. Schematically,

$$\vec{r}_{\rm C} = \Lambda_{C\leftarrow B}\vec{r}_{\rm B} = \Lambda_{C\leftarrow B}\Lambda_{B\leftarrow A}\vec{r}_{\rm A}$$

But if everything is OK, that means that the Lorentz transformations from A to B followed by the Lorentz transformation from B to C must be the same as the Lorentz transformation from A directly to C. In other words, the combination of two Lorentz transformations must be another Lorentz transformation.

Mathematicians say that Lorentz transformations must form a "group." It is much like rotations of a coordinate system in three spatial dimensions: a rotation followed by another one is equivalent to a single rotation over some combined angle. In fact, such spatial rotations *are* Lorentz transformations; just between coordinate systems that do not move compared to each other.

Using a lot of linear algebra, it may be verified that indeed the Lorentz transformations form a group, {D.5}.

1.3 Relativistic Mechanics

1.3.1 Intro to relativistic mechanics

Nonrelativistic mechanics is often based on the use of a potential energy to describe the forces. For example, in a typical molecular dynamics computation, the forces between the molecules are derived from a potential that depends on the differences in position between the atoms. Unfortunately, this sort of description fails badly in the truly relativistic case.

The basic problem is not difficult to understand. If a potential depends only on the spatial configuration of the atoms involved, then the motion of an atom instantaneously affects all the other ones. Relativity simply cannot handle instantaneous effects; they must be limited by the speed of light or major problems appear. It makes relativistic mechanics more difficult.

The simplest way to deal with the problem is to look at collisions between particles. Direct collisions inherently avoid erroneous action at a distance. They allow simple dynamics to be done without the use of a potential between particles that is relativistically suspect.

As a first example, consider two particles of equal mass and opposite speeds that collide as shown in the center of figure 1.3. You might think of the particles as two helium atoms. It will be assumed that while the speed of the atoms may

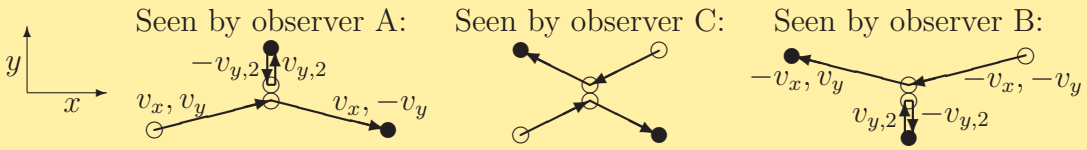


Figure 1.3: Example elastic collision seen by different observers.

be quite high, the collision is at a shallow enough angle that it does not excite the atoms. In other words, it is assumed that the collision is elastic.

As seen by observer C, the collision is perfectly symmetric. Regardless of the mechanics of the actual collision, observer C sees nothing wrong with it. The energy of the helium atoms is the same after the collision as before. Also, the net linear momentum was zero before the collision and still zero afterwards. And whatever little angular momentum there is, it too is still the same after the collision.

But now consider an observer A that moves horizontally along with the top helium atom. For this observer, the top helium atom comes down vertically and bounces back vertically. Observer B moves along with the bottom helium atom in the horizontal direction and sees that atom moving vertically. Now consider the Lorentz transformation (1.7) of the vertical velocity $v_{y,2}$ of the top atom as seen by observer A into the vertical velocity v_y of that atom as seen by observer B:

$$v_y = \sqrt{1-(v_x/c)^2}\,v_{y,2}$$

They are different! In particular, v_y is smaller than $v_{y,2}$. Therefore, if the masses of the helium atoms that the observers perceive would be their rest mass, linear momentum would not be conserved. For example, observer A would perceive a net downwards linear momentum before the collision and a net upwards linear momentum after it.

Clearly, linear momentum conservation is too fundamental a concept to be summarily thrown out. Instead, observer A perceives the mass of the rapidly moving lower atom to be the moving mass m_v, which is larger than the rest mass m by the Lorentz factor:

$$m_v = \frac{m}{\sqrt{1-(v/c)^2}}$$

and that exactly compensates for the lower vertical velocity in the expression for the momentum. (Remember that it was assumed that the collision is under a shallow angle, so the vertical velocity components are too small to have an effect on the masses.)

It is not difficult to understand why things are like this. The nonrelativistic definition of momentum allows two plausible generalizations to the relativistic

case:

$$\vec{p} = m\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} \qquad \Longrightarrow \qquad \begin{cases} \overrightarrow{p} = m\dfrac{\mathrm{d}\overrightarrow{r}}{\mathrm{d}t} \ ? \\ \overrightarrow{p} = m\dfrac{\mathrm{d}\overrightarrow{r}}{\mathrm{d}t_0} \ ? \end{cases}$$

Indeed, nonrelativistically, all observers agree about time intervals. However, relativistically the question arises whether the right time differential in momentum is $\mathrm{d}t$ as perceived by the observer, or the proper time difference $\mathrm{d}t_0$ as perceived by a hypothetical second observer moving along with the particle.

A little thought shows that the right time differential has to be $\mathrm{d}t_0$. For, after collisions the sum of the momenta should be the same as before them. However, the Lorentz velocity transformation (1.7) shows that perceived velocities transform nonlinearly from one observer to the next. For a nonlinear transformation, there is no reason to assume that if the momenta after a collision are the same as before for one observer, they are also so for another observer. On the other hand, since all observers agree about the proper time intervals, momentum based on the proper time interval $\mathrm{d}t_0$ transforms like $\mathrm{d}\overrightarrow{r}$, like position, and that is linear. A linear transformation does assure that if an observer A perceives that the sum of the momenta of a collection of particles $j = 1,2,\ldots$ is the same before and after,

$$\sum_j \overrightarrow{p}_{jA,\text{after}} = \sum_j \overrightarrow{p}_{jA,\text{before}}$$

then so does any other observer B:

$$\sum_j \Lambda_{B\leftarrow A}\overrightarrow{p}_{jA,\text{after}} = \sum_j \Lambda_{B\leftarrow A}\overrightarrow{p}_{jA,\text{before}} \quad \Rightarrow \quad \sum_j \overrightarrow{p}_{jB,\text{after}} = \sum_j \overrightarrow{p}_{jB,\text{before}}$$

Using the chain rule of differentiation, the components of the momentum four-vector $\overrightarrow{p}$ can be written out as

$$p_0 = mc\frac{\mathrm{d}t}{\mathrm{d}t_0} \quad p_1 = m\frac{\mathrm{d}t}{\mathrm{d}t_0}\frac{\mathrm{d}x}{\mathrm{d}t} \quad p_2 = m\frac{\mathrm{d}t}{\mathrm{d}t_0}\frac{\mathrm{d}y}{\mathrm{d}t} \quad p_3 = m\frac{\mathrm{d}t}{\mathrm{d}t_0}\frac{\mathrm{d}z}{\mathrm{d}t} \tag{1.14}$$

The components p_1, p_2, p_3 can be written in the same form as in the nonrelativistic case by defining a moving mass

$$m_v = m\frac{\mathrm{d}t}{\mathrm{d}t_0} = \frac{m}{\sqrt{1-(v/c)^2}} \tag{1.15}$$

How about the zeroth component? Since it too is part of the conservation law, reasonably speaking it can only be the relativistic equivalent of the nonrelativistic kinetic energy. Indeed, it equals $m_v c^2$ except for a trivial scaling factor $1/c$ to give it units of momentum.

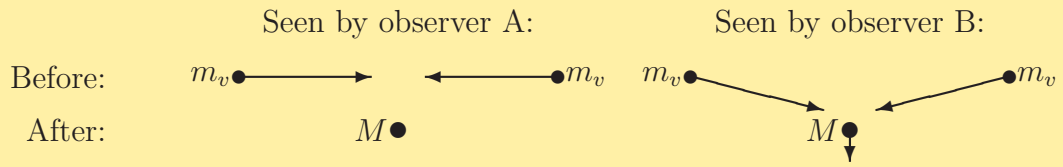


Figure 1.4: A completely inelastic collision.

Note that so far, this only indicates that the difference between $m_v c^2$ and mc^2 gives the kinetic energy. It does not imply that mc^2 by itself also corresponds to a meaningful energy. However, there is a beautifully simple argument to show that indeed kinetic energy can be converted into rest mass, [21]. Consider two identical rest masses m that are accelerated to high speed and then made to crash into each other head-on, as in the left part of figure 1.4. In this case, think of the masses as macroscopic objects, so that thermal energy is a meaningful concept for them. Assume that the collision has so much energy that the masses melt and merge without any rebound. By symmetry, the combined mass M has zero velocity. Momentum is conserved: the net momentum was zero before the collision because the masses had opposite velocity, and it is still zero after the collision. All very straightforward.

But now consider the same collision from the point of view of a second observer who is moving upwards slowly compared to the first observer with a small speed $v_{\rm B}$. No relativity involved here at all; going up so slowly, the second observer sees almost the same thing as the first one, with one difference. According to the second observer, the entire collision process seems to have a small downward velocity $v_{\rm B}$. The two masses have a slight downward velocity $v_{\rm B}$ before the collision and so has the mass M after the collision. But then vertical momentum conservation inevitably implies

$$2m_v v_{\rm B} = M v_{\rm B}$$

So M must be twice the moving mass m_v. The combined rest mass M is not the sum of the *rest* masses m, but of the *moving* masses m_v. All the kinetic energy given to the two masses has ended up as additional rest mass in M.

1.3.2 Lagrangian mechanics

Lagrangian mechanics can simplify many complicated dynamics problems. As an example, in this section it is used to derive the relativistic motion of a particle in an electromagnetic field.

Consider first the nonrelativistic motion of a particle in an electrostatic field. That is an important case for this book, because it is a good approximation for the electron in the hydrogen atom. To describe such purely nonrelativistic

motion, physicists like to define a Lagrangian as

$$\mathcal{L} = \tfrac{1}{2}m|\vec{v}|^2 - q\varphi \tag{1.16}$$

where m is the mass of the particle, $\vec{v}$ its velocity, and q its charge, while $q\varphi$ is the potential energy due to the electrostatic field, which depends on the position of the particle. (It is important to remember that the Lagrangian should mathematically be treated as a function of velocity and position of the particle. While for a given motion, the position and velocity are in turn functions of time, time derivatives must be implemented through the chain rule, i.e. by means of total derivatives of the Lagrangian.)

Physicists next define canonical, or generalized, momentum as the partial derivative of the Lagrangian with respect to velocity. An arbitrary component p_i^{c} of the canonical momentum is found as

$$p_i^{\mathrm{c}} = \frac{\partial \mathcal{L}}{\partial v_i} \tag{1.17}$$

This works out to be simply component $p_i = mv_i$ of the normal momentum. The equations of motion are taken to be

$$\frac{\mathrm{d}p_i^{\mathrm{c}}}{\mathrm{d}t} = \frac{\partial \mathcal{L}}{\partial r_i} \tag{1.18}$$

which is found to be

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = -q\frac{\partial \varphi}{\partial r_i}$$

That is simply Newton's second law; the left hand side is just mass times acceleration while in the right hand side minus the spatial derivative of the potential energy gives the force. It can also be seen that the sum of kinetic and potential energy of the particle remains constant, by multiplying Newton's equation by v_i and summing over i.

Since the Lagrangian is a just a scalar, it is relatively simple to guess its form in the relativistic case. To get the momentum right, simply replace the kinetic energy by an reciprocal Lorentz factor,

$$-mc^2\sqrt{1 - (|\vec{v}|/c)^2}$$

For velocities small compared to the speed of light, a two term Taylor series shows this is equivalent to mc^2 plus the kinetic energy. The constant mc^2 is of no importance since only derivatives of the Lagrangian are used. For any velocity, big or small, the canonical momentum as defined above produces the relativistic momentum based on the moving mass as it should.

The potential energy part of the Lagrangian is a bit trickier. The previous section showed that momentum is a four-vector including energy. Therefore,

going from one observer to another mixes up energy and momentum nontrivially, just like it mixes up space and time. That has consequences for energy conservation. In the classical solution, kinetic energy of the particle can temporarily be stored away as electrostatic potential energy and recovered later intact. But relativistically, the kinetic energy seen by one observer becomes momentum seen by another one. If that momentum is to be recovered intact later, there should be something like potential momentum. Since momentum is a vector, obviously so should potential momentum be: there must be something like a vector potential $\vec{A}$.

Based on those arguments, you might guess that the Lagrangian should be something like

$$\boxed{\mathcal{L} = -mc^2\sqrt{1-(|\vec{v}|/c)^2} + q\vec{A}\cdot\frac{\mathrm{d}\vec{r}}{\mathrm{d}t} \qquad \vec{A} = \left(\frac{1}{c}\varphi, A_x, A_y, A_z\right)} \tag{1.19}$$

And that is in fact right. Component zero of the potential four-vector is the classical electrostatic potential. The spatial vector $\vec{A} = (A_x, A_y, A_z)$ is called the "magnetic vector potential."

The canonical momentum is now

$$p_i^{\mathrm{c}} = \frac{\partial\mathcal{L}}{\partial v_i} = m_v v_i + qA_i \tag{1.20}$$

and that is no longer just the normal momentum, $p_i = m_v v_i$, but includes the magnetic vector potential.

The Lagrangian equations of motion become, the same way as before, but after clean up and in vector notation, {D.6}:

$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = q\vec{\mathcal{E}} + q\vec{v}\times\vec{\mathcal{B}} \tag{1.21}$$

The right-hand side in this equation of motion is called the Lorentz force. In it, $\vec{\mathcal{E}}$ is called the electric field and $\vec{\mathcal{B}}$ the magnetic field. These fields are related to the four-vector potential as

$$\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial\vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla\times\vec{A}$$

where by definition

$$\nabla = \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

is the vector operator called nabla or del.

Of course, if the Lagrangian above is right, it should apply to all observers, regardless of their relative motion. In particular, all observers should agree that the so-called "action" integral $\int\mathcal{L}\mathrm{d}t$ is stationary for the way that the

particle moves, {A.1.3}, {D.3.1} That requires that $\overrightarrow{A}$ transforms according to the Lorentz transformation.

(To see why, recall that dot products are the same for all observers, and that the square root in the Lagrangian (1.19) equals $\mathrm{d}t_0/\mathrm{d}t$ where the proper time interval $\mathrm{d}t_0$ is the same for all observers. So the action is the same for all observers.)

From the Lorentz transformation of $\overrightarrow{A}$, that of the electric and magnetic fields may be found; that is not a Lorentz transformation. Note that this suggests that $\overrightarrow{A}$ might be more fundamental physically than the more intuitive electric and magnetic fields. And that is in fact exactly what more advanced quantum mechanics shows, chapter 13.1.

It may be noted that the field strengths are unchanged in a "gauge transformation" that modifies φ and $\vec{A}$ into

$$\varphi' = \varphi - \frac{\partial \chi}{\partial t} \qquad \vec{A}' = \vec{A} + \nabla \chi \tag{1.22}$$

where χ is any arbitrary function of position and time. This might at first seem no more than a neat mathematical trick. But actually, in advanced quantum mechanics it is of decisive importance, chapter 7.3, {A.19.5}.

The energy can be found following addendum {A.1} as

$$E = \vec{v} \cdot \vec{p}^{\,\mathrm{c}} - \mathcal{L} = m_v c^2 + q\varphi$$

The Hamiltonian is the energy expressed in terms of the canonical momentum $\vec{p}^{\,\mathrm{c}}$ instead of $\vec{v}$; that works out to

$$H = \sqrt{(mc^2)^2 + (\vec{p}^{\,\mathrm{c}} - q\vec{A})^2 c^2} + q\varphi$$

using the formula given in the overview subsection. The Hamiltonian is of great importance in quantum mechanics.

Part II

Basic Quantum Mechanics

Chapter 2

Mathematical Prerequisites

Abstract

Quantum mechanics is based on a number of advanced mathematical ideas that are described in this chapter.

First the normal real numbers will be generalized to complex numbers. A number such as $\mathrm{i} = \sqrt{-1}$ is an invalid real number, but it is considered to be a valid complex one. The mathematics of quantum mechanics is most easily described in terms of complex numbers.

Classical physics tends to deal with numbers such as the position, velocity, and acceleration of particles. However, quantum mechanics deals primarily with functions rather than with numbers. To facilitate manipulating functions, they will be modeled as vectors in infinitely many dimensions. Dot products, lengths, and orthogonality can then all be used to manipulate functions. Dot products will however be renamed to be "inner products" and lengths to be "norms."

"Operators" will be defined that turn functions into other functions. Particularly important for quantum mechanics are "eigenvalue" cases, in which an operator turns a function into a simple multiple of itself.

A special class of operators, "Hermitian" operators will be defined. These will eventually turn out to be very important, because quantum mechanics associates physical quantities like position, momentum, and energy with corresponding Hermitian operators and their eigenvalues.

2.1 Complex Numbers

Quantum mechanics is full of complex numbers, numbers involving

$$\mathrm{i} = \sqrt{-1}.$$

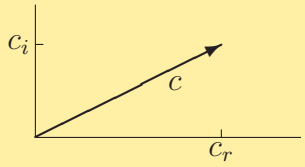
Note that $\sqrt{-1}$ is not an ordinary, "real", number, since there is no real number whose square is -1; the square of a real number is always positive. This section summarizes the most important properties of complex numbers.

First, any complex number, call it c, can by definition always be written in the form

$$c = c_r + \mathrm{i}c_i \tag{2.1}$$

where both c_r and c_i are ordinary real numbers, not involving $\sqrt{-1}$. The number c_r is called the real part of c and c_i the imaginary part.

You can think of the real and imaginary parts of a complex number as the components of a two-dimensional vector:



The length of that vector is called the "magnitude," or "absolute value" $|c|$ of the complex number. It equals

$$|c| = \sqrt{c_r^2 + c_i^2}.$$

Complex numbers can be manipulated pretty much in the same way as ordinary numbers can. A relation to remember is:

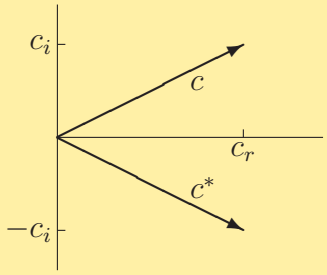
$$\frac{1}{\mathrm{i}} = -\mathrm{i} \tag{2.2}$$

which can be verified by multiplying the top and bottom of the fraction by i and noting that by definition $\mathrm{i}^2 = -1$ in the bottom.

The complex conjugate of a complex number c, denoted by c^*, is found by replacing i everywhere by $-\mathrm{i}$. In particular, if $c = c_r + \mathrm{i}c_i$, where c_r and c_i are real numbers, the complex conjugate is

$$c^* = c_r - \mathrm{i}c_i \tag{2.3}$$

The following picture shows that graphically, you get the complex conjugate of a complex number by flipping it over around the horizontal axis:



You can get the magnitude of a complex number c by multiplying c with its complex conjugate c^* and taking a square root:

$$|c| = \sqrt{c^* c} \tag{2.4}$$

If $c = c_r + \mathrm{i}c_i$, where c_r and c_i are real numbers, multiplying out c^*c shows the magnitude of c to be

$$|c| = \sqrt{c_r^2 + c_i^2}$$

which is indeed the same as before.

From the above graph of the vector representing a complex number c, the real part is $c_r = |c| \cos\alpha$ where α is the angle that the vector makes with the horizontal axis, and the imaginary part is $c_i = |c| \sin\alpha$. So you can write any complex number in the form

$$c = |c| \left(\cos\alpha + \mathrm{i}\sin\alpha\right)$$

The critically important Euler formula says that:

$$\cos\alpha + \mathrm{i}\sin\alpha = e^{\mathrm{i}\alpha} \tag{2.5}$$

So, any complex number can be written in "polar form" as

$$c = |c| e^{\mathrm{i}\alpha} \tag{2.6}$$

where both the magnitude $|c|$ and the phase angle (or argument) α are real numbers.

Any complex number of magnitude one can therefore be written as $e^{\mathrm{i}\alpha}$. Note that the only two real numbers of magnitude one, 1 and -1, are included for $\alpha = 0$, respectively $\alpha = \pi$. The number i is obtained for $\alpha = \pi/2$ and $-\mathrm{i}$ for $\alpha = -\pi/2$.

(See derivation {D.7} if you want to know where the Euler formula comes from.)

Key Points

- Complex numbers include the square root of minus one, i, as a valid number.
- All complex numbers can be written as a real part plus i times an imaginary part, where both parts are normal real numbers.
- The complex conjugate of a complex number is obtained by replacing i everywhere by $-$i.
- The magnitude of a complex number is obtained by multiplying the number by its complex conjugate and then taking a square root.

0— The Euler formula relates exponentials to sines and cosines.

2.1 Review Questions

1. Multiply out $(2+3\mathrm{i})^2$ and then find its real and imaginary part.
   Solution mathcplx-a
2. Show more directly that $1/\mathrm{i} = -\mathrm{i}$.
   Solution mathcplx-b
3. Multiply out $(2+3\mathrm{i})(2-3\mathrm{i})$ and then find its real and imaginary part.
   Solution mathcplx-c
4. Find the magnitude or absolute value of $2+3\mathrm{i}$.
   Solution mathcplx-d
5. Verify that $(2-3\mathrm{i})^2$ is still the complex conjugate of $(2+3\mathrm{i})^2$ if both are multiplied out.
   Solution mathcplx-e
6. Verify that $e^{-2\mathrm{i}}$ is still the complex conjugate of $e^{2\mathrm{i}}$ after both are rewritten using the Euler formula.
   Solution mathcplx-f
7. Verify that $\left(e^{\mathrm{i}\alpha} + e^{-\mathrm{i}\alpha}\right)/2 = \cos\alpha$.
   Solution mathcplx-g
8. Verify that $\left(e^{\mathrm{i}\alpha} - e^{-\mathrm{i}\alpha}\right)/2\mathrm{i} = \sin\alpha$.
   Solution mathcplx-h

2.2 Functions as Vectors

The second mathematical idea that is crucial for quantum mechanics is that functions can be treated in a way that is fundamentally not that much different from vectors.

A vector $\vec{f}$ (which might be velocity $\vec{v}$, linear momentum $\vec{p} = m\vec{v}$, force $\vec{F}$, or whatever) is usually shown in physics in the form of an arrow:

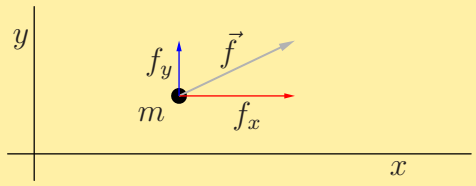


Figure 2.1: The classical picture of a vector.

However, the same vector may instead be represented as a spike diagram, by plotting the value of the components versus the component index:

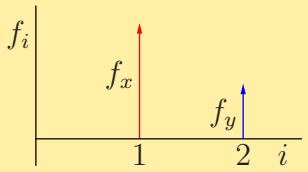


Figure 2.2: Spike diagram of a vector.

(The symbol i for the component index is not to be confused with $\mathrm{i} = \sqrt{-1}$.)

In the same way as in two dimensions, a vector in three dimensions, or, for that matter, in thirty dimensions, can be represented by a spike diagram:



Figure 2.3: More dimensions.

And just like vectors can be interpreted as spike diagrams, spike diagrams can be interpreted as vectors. So a spike diagram with very many spikes can be considered to be a single vector in a space with a very high number of dimensions.

In the limit of infinitely many spikes, the large values of i can be rescaled into a continuous coordinate, call it x. For example, x might be defined as i divided by the number of dimensions. In any case, the spike diagram becomes a function f of a continuous coordinate x:

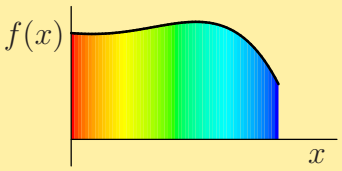


Figure 2.4: Infinite dimensions.

For functions, the spikes are usually not shown:

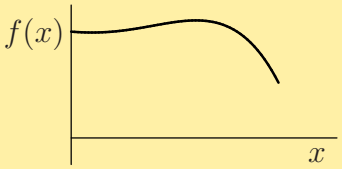


Figure 2.5: The classical picture of a function.

In this way, a function is just a single vector in an infinite dimensional space.

Note that the (x) in $f(x)$ does not mean "multiply by x." Here the (x) is only a way of reminding you that f is not a simple number but a function. Just like the arrow in $\vec{f}$ is only a way of reminding you that that f is not a simple number but a vector.

(It should be noted that to make the transition to infinite dimensions mathematically meaningful, you need to impose some smoothness constraints on the function. Typically, it is required that the function is continuous, or at least integrable in some sense. These details are not important for this book.)

Key Points

- A function can be thought of as a vector with infinitely many components.
- This allows quantum mechanics do the same things with functions as you can do with vectors.

2.2 Review Questions

1. Graphically compare the spike diagram of the 10-dimensional vector $\vec{v}$ with components (0.5,1,1.5,2,2.5,3,3.5,4,4.5,5) with the plot of the function $f(x) = 0.5\ x$.
   Solution funcvec-a
2. Graphically compare the spike diagram of the 10-dimensional unit vector $\hat{\imath}_3$, with components (0,0,1,0,0,0,0,0,0,0), with the plot of the function $f(x) = 1$. (No, they do not look alike.)
   Solution funcvec-b

2.3 The Dot, oops, INNER Product

The dot product of vectors is an important tool. It makes it possible to find the length of a vector, by multiplying the vector by itself and taking the square root. It is also used to check if two vectors are orthogonal: if their dot product is zero, they are. In this subsection, the dot product is defined for complex vectors and functions.

The usual dot product of two vectors $\vec{f}$ and $\vec{g}$ can be found by multiplying components with the same index i together and summing that:

$$\vec{f} \cdot \vec{g} \equiv f_1 g_1 + f_2 g_2 + f_3 g_3$$

(The emphatic equal, $\equiv$, is commonly used to indicate "is by definition equal" or "is always equal.") Figure 2.6 shows multiplied components using equal colors.



Figure 2.6: Forming the dot product of two vectors.

Note the use of numeric subscripts, f_1, f_2, and f_3 rather than f_x, f_y, and f_z; it means the same thing. Numeric subscripts allow the three term sum above to be written more compactly as:

$$\vec{f} \cdot \vec{g} \equiv \sum_{\text{all } i} f_i g_i$$

The Σ is called the "summation symbol."

The length of a vector $\vec{f}$, indicated by $|\vec{f}|$ or simply by f, is normally computed as

$$|\vec{f}| = \sqrt{\vec{f} \cdot \vec{f}} = \sqrt{\sum_{\text{all } i} f_i^2}$$

However, this does not work correctly for complex vectors. The difficulty is that terms of the form f_i^2 are no longer necessarily positive numbers. For example, $\mathrm{i}^2 = -1$.

Therefore, it is necessary to use a generalized "inner product" for complex vectors, which puts a complex conjugate on the first vector:

$$\boxed{\left\langle \vec{f} \middle| \vec{g} \right\rangle \equiv \sum_{\text{all } i} f_i^* g_i} \tag{2.7}$$

If the vector $\vec{f}$ is real, the complex conjugate does nothing, and the inner product $\left\langle \vec{f} \middle| \vec{g} \right\rangle$ is the same as the dot product $\vec{f} \cdot \vec{g}$. Otherwise, in the inner product $\vec{f}$ and $\vec{g}$ are no longer interchangeable; the conjugates are only on the *first* factor, $\vec{f}$. Interchanging $\vec{f}$ and $\vec{g}$ changes the inner product's value into its complex conjugate.

The length of a nonzero vector is now always a positive number:

$$\boxed{|\vec{f}| = \sqrt{\left\langle \vec{f} \middle| \vec{f} \right\rangle} = \sqrt{\sum_{\text{all } i} |f_i|^2}} \tag{2.8}$$

Physicists take the inner product "bracket" verbally apart as

$$\underset{\text{bra}}{\left\langle \vec{f} \right|} \quad \not\in \quad \underset{\text{ket}}{\left| \vec{g} \right\rangle}$$

and refer to vectors as bras and kets.

The inner product of functions is defined in exactly the same way as for vectors, by multiplying values at the same x-position together and summing. But since there are infinitely many x values, the sum becomes an integral:

$$\boxed{\langle f|g\rangle = \int_{\text{all } x} f^*(x)g(x)\,\mathrm{d}x} \tag{2.9}$$

Figure 2.7 shows multiplied function values using equal colors:



Figure 2.7: Forming the inner product of two functions.

The equivalent of the length of a vector is in the case of a function called its "norm:"

$$\boxed{||f|| \equiv \sqrt{\langle f|f\rangle} = \sqrt{\int_{\text{all } x} |f(x)|^2\,\mathrm{d}x}} \tag{2.10}$$

The double bars are used to avoid confusion with the absolute value of the function.

A vector or function is called "normalized" if its length or norm is one:

$$\boxed{\langle f|f\rangle = 1 \text{ iff } f \text{ is normalized.}} \tag{2.11}$$

("iff" should really be read as "if and only if.")

Two vectors, or two functions, f and g, are by definition orthogonal if their inner product is zero:

$$\boxed{\langle f|g\rangle = 0 \text{ iff } f \text{ and } g \text{ are orthogonal.}} \tag{2.12}$$

Sets of vectors or functions that are all

- mutually orthogonal, and
- normalized

occur a lot in quantum mechanics. Such sets should be called "orthonormal", though the less precise term "orthogonal" is often used instead. This document will refer to them correctly as being orthonormal.

So, a set of functions or vectors $f_1, f_2, f_3, \ldots$ is orthonormal if

$$0 = \langle f_1|f_2\rangle = \langle f_2|f_1\rangle = \langle f_1|f_3\rangle = \langle f_3|f_1\rangle = \langle f_2|f_3\rangle = \langle f_3|f_2\rangle = \ldots$$

and

$$1 = \langle f_1|f_1\rangle = \langle f_2|f_2\rangle = \langle f_3|f_3\rangle = \ldots$$

Key Points

- For complex vectors and functions, the normal dot product becomes the inner product.
- To take an inner product of vectors,
  - take complex conjugates of the components of the first vector;
  - multiply corresponding components of the two vectors together;
  - sum these products.
- To take an inner product of functions,
  - take the complex conjugate of the first function;
  - multiply the two functions;
  - integrate the product function.
- To find the length of a vector, take the inner product of the vector with itself, and then a square root.
- To find the norm of a function, take the inner product of the function with itself, and then a square root.
- A pair of vectors, or a pair of functions, is orthogonal if their inner product is zero.
- A set of vectors forms an orthonormal set if every one is orthogonal to all the rest, and every one is of unit length.
- A set of functions forms an orthonormal set if every one is orthogonal to all the rest, and every one is of unit norm.

2.3 Review Questions

1. Find the following inner product of the two vectors:

$$\left\langle \begin{pmatrix} 1+\mathrm{i} \\ 2-\mathrm{i} \end{pmatrix} \middle| \begin{pmatrix} 2\mathrm{i} \\ 3 \end{pmatrix} \right\rangle$$

   Solution dot-a

2. Find the length of the vector

$$\begin{pmatrix} 1+\mathrm{i} \\ 3 \end{pmatrix}$$

   Solution dot-b

3. Find the inner product of the functions $\sin(x)$ and $\cos(x)$ on the interval $0 \leqslant x \leqslant 1$.
   Solution dot-c

4. Show that the functions $\sin(x)$ and $\cos(x)$ are orthogonal on the interval $0 \leqslant x \leqslant 2\pi$.
   Solution dot-d

5. Verify that $\sin(x)$ is not a normalized function on the interval $0 \leqslant x \leqslant 2\pi$, and normalize it by dividing by its norm.
   Solution dot-e
6. Verify that the most general multiple of $\sin(x)$ that is normalized on the interval $0 \leqslant x \leqslant 2\pi$ is $e^{\mathrm{i}\alpha}\sin(x)/\sqrt{\pi}$ where α is any arbitrary real number. So, using the Euler formula, the following multiples of $\sin(x)$ are all normalized: $\sin(x)/\sqrt{\pi}$, (for $\alpha = 0$), $-\sin(x)/\sqrt{\pi}$, (for $\alpha = \pi$), and $\mathrm{i}\sin(x)/\sqrt{\pi}$, (for $\alpha = \pi/2$).
   Solution dot-f
7. Show that the functions $e^{4\mathrm{i}\pi x}$ and $e^{6\mathrm{i}\pi x}$ are an orthonormal set on the interval $0 \leqslant x \leqslant 1$.
   Solution dot-g

2.4 Operators

This section defines operators, which are a generalization of matrices. Operators are the principal components of quantum mechanics.

In a finite number of dimensions, a matrix A can transform any arbitrary vector v into a different vector $A\vec{v}$:

$$\vec{v} \xrightarrow{\text{matrix } A} \vec{w} = A\vec{v}$$

Similarly, an operator transforms a function into another function:

$$f(x) \xrightarrow{\text{operator } A} g(x) = Af(x)$$

Some simple examples of operators:

$$f(x) \xrightarrow{\widehat{x}} g(x) = xf(x)$$

$$f(x) \xrightarrow{\frac{\mathrm{d}}{\mathrm{d}x}} g(x) = f'(x)$$

Note that a hat is often used to indicate operators; for example, $\widehat{x}$ is the symbol for the operator that corresponds to multiplying by x. If it is clear that something is an operator, such as $\mathrm{d}/\mathrm{d}x$, no hat will be used.

It should really be noted that the operators that you are interested in in quantum mechanics are "linear" operators. If you increase a function f by a factor, Af increases by that same factor. Also, for any two functions f and g, $A(f+g)$ will be $(Af) + (Ag)$. For example, differentiation is a linear operator:

$$\frac{\mathrm{d}\Big(c_1 f(x) + c_2 g(x)\Big)}{\mathrm{d}x} = c_1 \frac{\mathrm{d}f(x)}{\mathrm{d}x} + c_2 \frac{\mathrm{d}g(x)}{\mathrm{d}x}$$

Squaring is *not* a linear operator:

$$\Big(c_1 f(x) + c_2 g(x)\Big)^2 = c_1^2 f^2(x) + 2c_1 c_2 f(x) g(x) + c_2^2 g^2(x) \neq c_1 f^2(x) + c_2 g^2(x)$$

However, it is not something to really worry about. You will not find a single nonlinear operator in the rest of this entire book.

Key Points

- Matrices turn vectors into other vectors.
- Operators turn functions into other functions.

2.4 Review Questions

1. So what is the result if the operator $\mathrm{d}/\mathrm{d}x$ is applied to the function $\sin(x)$?
   Solution mathops-a
2. If, say, $\widehat{x^2}\sin(x)$ is simply the function $x^2\sin(x)$, then what *is* the difference between $\widehat{x^2}$ and x^2?
   Solution mathops-b
3. A less self-evident operator than the above examples is a translation operator like $\mathcal{T}_{\pi/2}$ that translates the graph of a function towards the left by an amount $\pi/2$: $\mathcal{T}_{\pi/2} f(x) = f\left(x + \frac{1}{2}\pi\right)$. (Curiously enough, translation operators turn out to be responsible for the law of conservation of momentum.) Show that $\mathcal{T}_{\pi/2}$ turns $\sin(x)$ into $\cos(x)$.
   Solution mathops-c
4. The inversion, or parity, operator Π turns $f(x)$ into $f(-x)$. (It plays a part in the question to what extent physics looks the same when seen in the mirror.) Show that Π leaves $\cos(x)$ unchanged, but turns $\sin(x)$ into $-\sin(x)$.
   Solution mathops-d

2.5 Eigenvalue Problems

To analyze quantum mechanical systems, it is normally necessary to find so-called eigenvalues and eigenvectors or eigenfunctions. This section defines what they are.

A nonzero vector $\vec{v}$ is called an eigenvector of a matrix A if $A\vec{v}$ is a multiple of the same vector:

$$A\vec{v} = a\vec{v} \text{ iff } \vec{v} \text{ is an eigenvector of } A \tag{2.13}$$

The multiple a is called the eigenvalue. It is just a number.

A nonzero function f is called an eigenfunction of an operator A if Af is a multiple of the same function:

$$Af = af \text{ iff } f \text{ is an eigenfunction of } A. \tag{2.14}$$

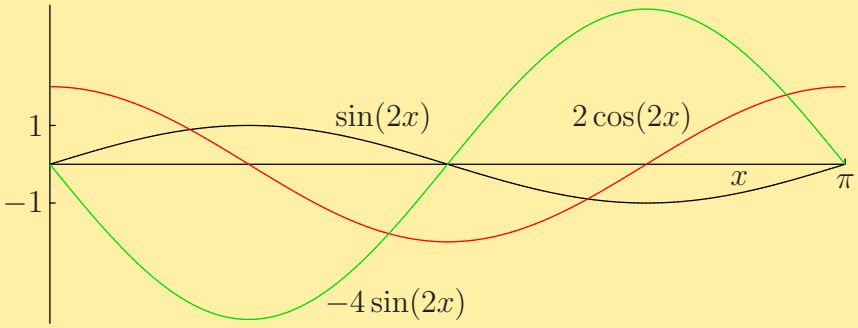


Figure 2.8: Illustration of the eigenfunction concept. Function $\sin(2x)$ is shown in black. Its first derivative $2\cos(2x)$, shown in red, is not just a multiple of $\sin(2x)$. Therefore $\sin(2x)$ is *not* an eigenfunction of the first derivative operator. However, the second derivative of $\sin(2x)$ is $-4\sin(2x)$, which is shown in green, and that is indeed a multiple of $\sin(2x)$. So $\sin(2x)$ is an eigenfunction of the second derivative operator, and with eigenvalue -4.

For example, e^x is an eigenfunction of the operator $\mathrm{d}/\mathrm{d}x$ with eigenvalue 1, since $\mathrm{d}e^x/\mathrm{d}x = 1\ e^x$. Another simple example is illustrated in figure 2.8; the function $\sin(2x)$ is *not* an eigenfunction of the first derivative operator $\mathrm{d}/\mathrm{d}x$. However it *is* an eigenfunction of the second derivative operator $\mathrm{d}^2/\mathrm{d}x^2$, and with eigenvalue -4.

Eigenfunctions like e^x are not very common in quantum mechanics since they become very large at large x, and that typically does not describe physical situations. The eigenfunctions of the first derivative operator $\mathrm{d}/\mathrm{d}x$ that do appear a lot are of the form $e^{\mathrm{i}kx}$, where $\mathrm{i} = \sqrt{-1}$ and k is an arbitrary real number. The eigenvalue is $\mathrm{i}k$:

$$\frac{\mathrm{d}}{\mathrm{d}x} e^{\mathrm{i}kx} = \mathrm{i}k e^{\mathrm{i}kx}$$

Function $e^{\mathrm{i}kx}$ does not blow up at large x; in particular, the Euler formula (2.5) says:

$$e^{\mathrm{i}kx} = \cos(kx) + \mathrm{i}\sin(kx)$$

The constant k is called the "wave number."

Key Points

- If a matrix turns a nonzero vector into a multiple of that vector, then that vector is an eigenvector of the matrix, and the multiple is the eigenvalue.
- If an operator turns a nonzero function into a multiple of that function, then that function is an eigenfunction of the operator, and the multiple is the eigenvalue.

2.5 Review Questions

1. Show that $e^{\mathrm{i}kx}$, above, is also an eigenfunction of $\mathrm{d}^2/\mathrm{d}x^2$, but with eigenvalue $-k^2$. In fact, it is easy to see that the square of any operator has the same eigenfunctions, but with the square eigenvalues.
   Solution eigvals-a
2. Show that any function of the form $\sin(kx)$ and any function of the form $\cos(kx)$, where k is a constant called the wave number, is an eigenfunction of the operator $\mathrm{d}^2/\mathrm{d}x^2$, though they are not eigenfunctions of $\mathrm{d}/\mathrm{d}x$.
   Solution eigvals-b
3. Show that $\sin(kx)$ and $\cos(kx)$, with k a constant, are eigenfunctions of the inversion operator Π, which turns any function $f(x)$ into $f(-x)$, and find the eigenvalues.
   Solution eigvals-c

2.6 Hermitian Operators

Most operators in quantum mechanics are of a special kind called "Hermitian". This section lists their most important properties.

An operator is called Hermitian when it can always be flipped over to the other side if it appears in a inner product:

$$\langle f|Ag\rangle = \langle Af|g\rangle \text{ always iff } A \text{ is Hermitian.} \tag{2.15}$$

That is the definition, but Hermitian operators have the following additional special properties:

- They always have real eigenvalues, not involving $\mathrm{i} = \sqrt{-1}$. (But the eigenfunctions, or eigenvectors if the operator is a matrix, might be complex.) Physical values such as position, momentum, and energy are ordinary real numbers since they are eigenvalues of Hermitian operators {N.3}.
- Their eigenfunctions can always be chosen so that they are normalized and mutually orthogonal, in other words, an orthonormal set. This tends to simplify the various mathematics a lot.
- Their eigenfunctions form a "complete" set. This means that *any* function can be written as some linear combination of the eigenfunctions. (There is a proof in derivation {D.8} for an important example. But see also {N.4}.) In practical terms, it means that you only need to look at the eigenfunctions to completely understand what the operator does.

In the linear algebra of real matrices, Hermitian operators are simply symmetric matrices. A basic example is the inertia matrix of a solid body in Newtonian dynamics. The orthonormal eigenvectors of the inertia matrix give the directions of the principal axes of inertia of the body.

An orthonormal complete set of eigenvectors or eigenfunctions is an example of a so-called "basis." In general, a basis is a minimal set of vectors or functions that you can write all other vectors or functions in terms of. For example, the unit vectors $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ are a basis for normal three-dimensional space. Every three-dimensional vector can be written as a linear combination of the three.

The following properties of inner products involving Hermitian operators are often needed, so they are listed here:

$$\text{If } A \text{ is Hermitian:} \quad \langle g|Af\rangle = \langle f|Ag\rangle^*, \quad \langle f|Af\rangle \text{ is real.} \tag{2.16}$$

The first says that you can swap f and g if you take the complex conjugate. (It is simply a reflection of the fact that if you change the sides in an inner product, you turn it into its complex conjugate. Normally, that puts the operator at the other side, but for a Hermitian operator, it does not make a difference.) The second is important because ordinary real numbers typically occupy a special place in the grand scheme of things. (The fact that the inner product is real merely reflects the fact that if a number is equal to its complex conjugate, it must be real; if there was an i in it, the number would change by a complex conjugate.)

Key Points

- Hermitian operators can be flipped over to the other side in inner products.
- Hermitian operators have only real eigenvalues.
- Hermitian operators have a complete set of orthonormal eigenfunctions (or eigenvectors).

2.6 Review Questions

1. A matrix A is defined to convert any vector $\vec{r} = x\hat{\imath}+y\hat{\jmath}$ into $\vec{r}_2 = 2x\hat{\imath}+4y\hat{\jmath}$. Verify that $\hat{\imath}$ and $\hat{\jmath}$ are orthonormal eigenvectors of this matrix, with eigenvalues 2, respectively 4.
   Solution herm-a
2. A matrix A is defined to convert any vector $\vec{r} = (x, y)$ into the vector $\vec{r}_2 = (x + y, x + y)$. Verify that $(\cos 45°, \sin 45°)$ and $(\cos 45°, -\sin 45°)$ are orthonormal eigenvectors of this matrix, with eigenvalues 2 respectively 0. Note: $\cos 45° = \sin 45° = \frac{1}{2}\sqrt{2}$.
   Solution herm-b
3. Show that the operator $\widehat{2}$ is a Hermitian operator, but $\widehat{\imath}$ is not.
   Solution herm-c
4. Generalize the previous question, by showing that any complex constant c comes out of the right hand side of an inner product unchanged, but out of the left hand side as its complex conjugate;

$$\langle f|cg\rangle = c\langle f|g\rangle \qquad \langle cf|g\rangle = c^*\langle f|g\rangle.$$

As a result, a number c is only a Hermitian operator if it is real: if c is complex, the two expressions above are not the same.

Solution herm-d

5. Show that an operator such as $\widehat{x}^2$, corresponding to multiplying by a real function, is an Hermitian operator.

   Solution herm-e

6. Show that the operator $\mathrm{d}/\mathrm{d}x$ is *not* a Hermitian operator, but $\mathrm{i}\mathrm{d}/\mathrm{d}x$ is, assuming that the functions on which they act vanish at the ends of the interval $a \leqslant x \leqslant b$ on which they are defined. (Less restrictively, it is only required that the functions are "periodic"; they must return to the same value at $x = b$ that they had at $x = a$.)

   Solution herm-f

7. Show that if A is a Hermitian operator, then so is A^2. As a result, under the conditions of the previous question, $-\mathrm{d}^2/\mathrm{d}x^2$ is a Hermitian operator too. (And so is just $\mathrm{d}^2/\mathrm{d}x^2$, of course, but $-\mathrm{d}^2/\mathrm{d}x^2$ is the one with the positive eigenvalues, the squares of the eigenvalues of $\mathrm{i}\mathrm{d}/\mathrm{d}x$.)

   Solution herm-g

8. A complete set of orthonormal eigenfunctions of $-\mathrm{d}^2/\mathrm{d}x^2$ on the interval $0 \leqslant x \leqslant \pi$ that are zero at the end points is the infinite set of functions

$$\frac{\sin(x)}{\sqrt{\pi/2}}, \frac{\sin(2x)}{\sqrt{\pi/2}}, \frac{\sin(3x)}{\sqrt{\pi/2}}, \frac{\sin(4x)}{\sqrt{\pi/2}}, \ldots$$

   Check that these functions are indeed zero at $x = 0$ and $x = \pi$, that they are indeed orthonormal, and that they are eigenfunctions of $-\mathrm{d}^2/\mathrm{d}x^2$ with the positive real eigenvalues

$$1, 4, 9, 16, \ldots$$

   Completeness is a much more difficult thing to prove, but they are. The completeness proof in the notes covers this case.

   Solution herm-h

9. A complete set of orthonormal eigenfunctions of the operator $\mathrm{i}\mathrm{d}/\mathrm{d}x$ that are periodic on the interval $0 \leqslant x \leqslant 2\pi$ are the infinite set of functions

$$\ldots, \frac{e^{-3\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{-2\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{-\mathrm{i}x}}{\sqrt{2\pi}}, \frac{1}{\sqrt{2\pi}}, \frac{e^{\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{2\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{3\mathrm{i}x}}{\sqrt{2\pi}}, \ldots$$

   Check that these functions are indeed periodic, orthonormal, and that they are eigenfunctions of $\mathrm{i}\mathrm{d}/\mathrm{d}x$ with the real eigenvalues

$$\ldots, 3, 2, 1, 0, -1, -2, -3, \ldots$$

   Completeness is a much more difficult thing to prove, but they are. The completeness proof in the notes covers this case.

   Solution herm-i

2.7 Additional Points

This subsection describes a few further issues of importance for this book.

2.7.1 Dirac notation

Physicists like to write inner products such as $\langle f|Ag\rangle$ in "Dirac notation":

$$\langle f|A|g\rangle \equiv \langle f|Ag\rangle$$

since this conforms more closely to how you would think of it in linear algebra:

$$\begin{array}{ccc} \langle f| & A & |g\rangle \\ \text{bra} & \text{operator} & \text{ket} \end{array}$$

The various advanced ideas of linear algebra can be extended to operators in this way, but they will not be needed in this book.

One thing will be needed in some more advanced addenda, however. That is the case that operator A is *not* Hermitian. In that case, if you want to take A to the other side of the inner product, you need to change it into a different operator. That operator is called the "Hermitian conjugate" of A. In physics, it is almost always indicated as $A^\dagger$. So, simply by definition,

$$\langle f|Ag\rangle \equiv \int_{\text{all } x} f^*(x)\,(Ag(x))\,\mathrm{d}x \equiv \int_{\text{all } x} (A^\dagger f(x))^* g(x)\,\mathrm{d}x \equiv \left\langle A^\dagger f \middle| g \right\rangle$$

Then there are some more things that this book will not use. However, you will almost surely encounter these when you read other books on quantum mechanics.

First, the dagger is used much like a generalization of "complex conjugate,"

$$f^\dagger \equiv f^* \qquad |f\rangle^\dagger \equiv \langle f|$$

etcetera. Applying a dagger a second time gives the original back. Also, if you work out the dagger on a product, you need to reverse the order of the factors. For example

$$\Big(A^\dagger|f\rangle\Big)^\dagger|g\rangle = \langle f|A|g\rangle$$

In words, putting $A^\dagger|f\rangle$ into the left side of an inner product gives $\langle f|A$.

The second point will be illustrated for the case of vectors in three dimensions. Such a vector can be written as

$$\vec{v} = \hat{\imath}v_x + \hat{\jmath}v_y + \hat{k}v_z$$

Here $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ are the three unit vectors in the axial directions. The components v_x, v_y and v_z can be found using dot products:

$$\vec{v} = \hat{\imath}(\hat{\imath}\cdot\vec{v}) + \hat{\jmath}(\hat{\jmath}\cdot\vec{v}) + \hat{k}(\hat{k}\cdot\vec{v})$$

Symbolically, you can write this as

$$\vec{v} = (\hat{\imath}\hat{\imath}\cdot + \hat{\jmath}\hat{\jmath}\cdot + \hat{k}\hat{k}\cdot)\vec{v}$$

In fact, the operator in parentheses can be *defined* by saying that for any vector $\vec{v}$, it gives the exact same vector back. Such an operator is called an "identity operator."

The relation

$$(\hat{\imath}\hat{\imath}\cdot + \hat{\jmath}\hat{\jmath}\cdot + \hat{k}\hat{k}\cdot) = 1$$

is called the "completeness relation." To see why, suppose you leave off the third part of the operator. Then

$$(\hat{\imath}\hat{\imath}\cdot + \hat{\jmath}\hat{\jmath}\cdot)\vec{v} = \hat{\imath}v_x + \hat{\jmath}v_y$$

The z-component is gone! Now the vector $\vec{v}$ gets projected onto the x, y-plane. The operator has become a "projection operator" instead of an identity operator by not suming over the complete set of unit vectors.

You will almost always find these things in terms of bras and kets. To see how that looks, define

$$\hat{\imath} \equiv |1\rangle \qquad \hat{\jmath} \equiv |2\rangle \qquad \hat{k} \equiv |3\rangle \qquad \vec{v} \equiv |v\rangle$$

Then

$$|v\rangle = |1\rangle\langle 1||v\rangle + |2\rangle\langle 2||v\rangle + |3\rangle\langle 3||v\rangle = \sum_{\text{all } i} |i\rangle\langle i||v\rangle$$

so the completeness relation looks like

$$\sum_{\text{all } i} |i\rangle\langle i| = 1$$

If you do not sum over the complete set of kets, you get a projection operator instead of an identity one.

2.7.2 Additional independent variables

In many cases, the functions involved in an inner product may depend on more than a single variable x. For example, they might depend on the position (x, y, z) in three-dimensional space.

The rule to deal with that is to ensure that the inner product integrations are over *all* independent variables. For example, in three spatial dimensions:

$$\langle f|g\rangle = \int_{\text{all } x}\int_{\text{all } y}\int_{\text{all } z} f^*(x,y,z)g(x,y,z)\,\mathrm{d}x\mathrm{d}y\mathrm{d}z$$

Note that the time t is a somewhat different variable from the rest, and time is *not* included in the inner product integrations.

Chapter 3

Basic Ideas of Quantum Mechanics

Abstract

In this chapter the basic ideas of quantum mechanics are described and then a basic but very important example is worked out.

Before embarking on this chapter, do take note of the very sage advice given by Richard Feynman, Nobel-prize winning pioneer of relativistic quantum mechanics:

> "Do not keep saying to yourself, if you can possibly avoid it, 'But how can it be like that?' because you will get 'down the drain,' into a blind alley from which nobody has yet escaped. Nobody knows how it can be like that." [Richard P. Feynman (1965) *The Character of Physical Law* 129. BBC/Penguin]
>
> "So do not take the lecture too seriously, . . . , but just relax and enjoy it." [*ibid.*]

And it may be uncertain whether Niels Bohr, Nobel-prize winning pioneer of early quantum mechanics actually said it to Albert Einstein, and if so, exactly what he said, but it may be the sanest statement about quantum mechanics of all:

> "Stop telling God what to do." [Neils Bohr, reputed].

First of all, this chapter will throw out the classical picture of particles with positions and velocities. Completely.

Quantum mechanics substitutes instead a function called the "wave function" that associates a numerical value with *every possible state of the universe*. If the "universe" that you are considering is just a single particle, the wave function of interest associates a numerical value with every possible position of that particle, at every time.

The physical meaning of the value of the wave function, also called the "quantum amplitude," itself is somewhat hazy. It is just a complex number. However, the square magnitude of the number has a clear meaning, first stated by Born: The square magnitude of the wave function at a point is a measure of the probability of finding the particle at that point, *if* you conduct such a search.

But if you do, watch out. Heisenberg has shown that if you eliminate the uncertainty in the position of a particle, its linear momentum explodes. If the position is precise, the linear momentum has infinite uncertainty. The same thing also applies in reverse. Neither position nor linear momentum can have an precise value for a particle. And usually other quantities like energy do not either.

Which brings up the question what meaning to attach to such physical quantities. Quantum mechanics answers that by associating a separate Hermitian operator with every physical quantity. The most important ones will be described. These operators act on the wave function. If, and only if, the wave function is an eigenfunction of such a Hermitian operator, only then does the corresponding physical quantity have a definite value: the eigenvalue. In all other cases the physical quantity is uncertain.

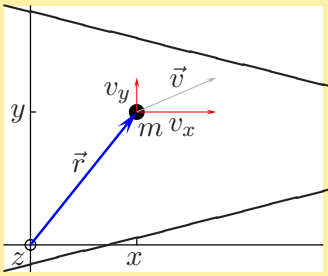
The most important Hermitian operator is called the "Hamiltonian." It is associated with the total energy of the particle. The eigenvalues of the Hamiltonian describe the only possible values that the total energy of the particle can have.

The chapter will conclude by analyzing a simple quantum system in detail. It is a particle stuck in a pipe of square cross section. While relatively simple, this case describes some of the quantum effects encountered in nanotechnology. In later chapters, it will be found that this case also provides a basic model for such systems as valence electrons in metals, molecules in ideal gases, nucleons in nuclei, and much more.

3.1 The Revised Picture of Nature

This section describes the view quantum mechanics has of nature, which is in terms of a mysterious function called the "wave function".

According to quantum mechanics, the way that the old Newtonian physics describes nature is wrong if examined closely enough. Not just a bit wrong. Totally wrong. For example, the Newtonian picture for a particle of mass m looks like figure 3.1:



velocity: $v_x = \dfrac{\mathrm{d}x}{\mathrm{d}t}$, etc.

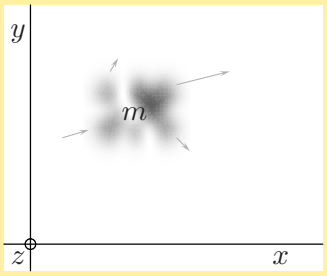
linear momentum: $p_x = mv_x$, etc.

force: $F_x = m\dfrac{\mathrm{d}v_x}{\mathrm{d}t}$, etc. (Newton's second law)

Figure 3.1: The old incorrect Newtonian physics.

The problems? A numerical position for the particle simply *does not exist.* A numerical velocity or linear momentum for the particle *does not exist.*

What does exist according to quantum mechanics is the so-called wave function $\Psi(x, y, z; t)$. Its square magnitude, $|\Psi|^2$, can be shown as grey tones (darker where the magnitude is larger), as in figure 3.2:



A precise position does not exist. Neither does a precise velocity or linear momentum.

Quantum mechanics does not use forces, but the potential energy V. It is implicitly defined by:

$$F_x = -\frac{\partial V}{\partial x} \qquad F_y = -\frac{\partial V}{\partial y} \qquad F_z = -\frac{\partial V}{\partial z}$$

Figure 3.2: The correct quantum physics.

The physical meaning of the wave function is known as "Born's statistical interpretation": darker regions are regions where the particle is more likely to be found if the location is narrowed down. More precisely, if $\vec{r} = (x, y, z)$ is a given location, then

$$|\Psi(\vec{r}; t)|^2 \, \mathrm{d}^3\vec{r} \tag{3.1}$$

is the probability of finding the particle within a small volume, of size $\mathrm{d}^3\vec{r} = \mathrm{d}x\mathrm{d}y\mathrm{d}z$, around that given location, *if* such a measurement is attempted.

(And if such a position measurement is actually done, it affects the wave function: after the measurement, the new wave function will be restricted to the volume to which the position was narrowed down. But it will spread out again in time if allowed to do so afterwards.)

The particle must be found somewhere if you look everywhere. In quantum mechanics, that is expressed by the fact that the total probability to find the particle, integrated over all possible locations, must be 100% (certainty):

$$\int_{\text{all } \vec{r}} |\Psi(\vec{r}; t)|^2 \, \mathrm{d}^3\vec{r} = 1 \tag{3.2}$$

In other words, proper wave functions are normalized, $\langle\Psi|\Psi\rangle = 1$.

The position of macroscopic particles is typically very much narrowed down by incident light, surrounding objects, earlier history, etcetera. For such particles, the "blob size" of the wave function is extremely small. As a result, claiming that a macroscopic particle, is, say, at the center point of the wave function blob may be just fine in practical applications. But when you are interested in what happens on very small scales, the nonzero blob size can make a big difference.

In addition, even on macroscopic scales, position can be ill defined. Consider what happens if you take the wave function blob apart and send half to Mars and half to Venus. Quantum mechanics allows it; this is what happens in a "scattering" experiment. You would presumably need to be extremely careful to do it on such a large scale, but there is no *fundamental* theoretical objection in quantum mechanics. So, where is the particle now? Hiding on Mars? Hiding on Venus?

Orthodox quantum mechanics says: *neither*. It will pop up on one of the two planets if measurements force it to reveal its presence. But until that moment, it is just as ready to pop up on Mars as on Venus, at an instant's notice. If it was hiding on Mars, it could not possibly pop up on Venus on an instant's notice; the fastest it would be allowed to move is at the speed of light. Worse, when the electron does pop up on Mars, it must communicate that fact instantaneously to Venus to prevent itself from also popping up there. That requires that quantum mechanics internally communicates at speeds faster than the speed of light. That is called the Einstein-Podolski-Rosen paradox. A famous theorem by John Bell in 1964 implies that nature really does communicate instantaneously; it is not just some unknown deficiency in the theory of quantum mechanics, chapter 8.2.

Of course, quantum mechanics is largely a matter of inference. The wave function cannot be directly observed. But that is not as strong an argument against quantum mechanics as it may seem. The more you learn about quantum mechanics, the more its weirdness will probably become inescapable. After almost a century, quantum mechanics is still standing, with no real "more reasonable" competitors, ones that stay closer to the common sense picture. And the best minds in physics have tried.

From a more practical point of view, you might object that the Born interpretation cheats: it only explains what the absolute value of the wave function is, not what the wave function itself is. And you would have a very good point there. Ahem. The wave function $\Psi(\vec{r}, t)$ itself gives the "quantum amplitude" that the particle can be found at position $\vec{r}$. No, indeed that does not help at all. That is just a definition.

However, for unknown reasons nature always "computes" with a wave function, never with probabilities. The classical example is where you shoot electrons at random at a tiny pinhole in a wall. Open up a second hole, and you would

expect that every point behind the wall would receive more electrons, with another hole open. The probability of getting the electron from the second hole should add to the probability of getting it from the first one. But that is not what happens. Some points may now get no electrons at all. The wave function trace passing through the second hole may arrive with the opposite sign of the wave function trace passing through the first hole. If that happens, the net wave function becomes zero, and so its square magnitude, the probability of finding an electron, does too. Electrons are prevented from reaching the location by giving them an additional way to get there. It is weird. Then again, there is little profit in worrying about it.

Key Points

- According to quantum mechanics, particles do not have precise values of position or velocity when examined closely enough.
- What they do have is a "wave function" that depends on position.
- Larger values of the magnitude of the wave function, (indicated in this book by darker regions,) correspond to regions where the particle is more likely to be found if a location measurement is done.
- Such a measurement changes the wave function; the measurement itself creates the reduced uncertainty in position that exists immediately after the measurement.
- In other words, the wave function is all there is; you cannot identify a hidden position in a *given* wave function, just create a *new* wave function that more precisely locates the particle.
- The creation of such a more localized wave function during a position measurement is governed by laws of chance: the more localized wave function is more likely to end up in regions where the initial wave function had a larger magnitude.
- Proper wave functions are normalized.

3.2 The Heisenberg Uncertainty Principle

The Heisenberg uncertainty principle is a way of expressing the qualitative properties of quantum mechanics in an easy to visualize way.

Figure 3.3 is a combination plot of the position x of a particle and the corresponding linear momentum $p_x = mv_x$, (with m the mass and v_x the velocity in the x-direction). To the left in the figure, both the position and the linear momentum have some uncertainty.

The right of the figure shows what happens if you squeeze down on the particle to try to restrict it to one position x: it stretches out in the momentum direction.

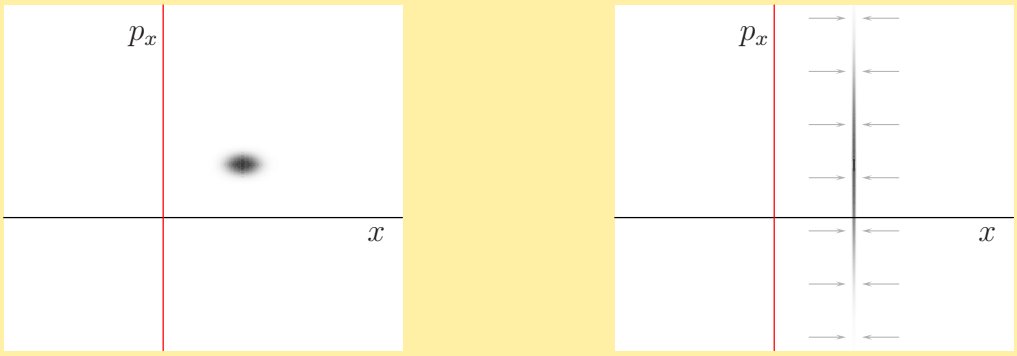


Figure 3.3: Illustration of the Heisenberg uncertainty principle. A combination plot of position and linear momentum components in a single direction is shown. Left: Fairly localized state with fairly low linear momentum. Right: narrowing down the position makes the linear momentum explode.

Heisenberg showed that according to quantum mechanics, the area of the "blob" cannot be contracted to a point. When you try to narrow down the position of a particle, you get into trouble with momentum. Conversely, if you try to pin down a precise momentum, you lose all hold on the position.

The area of the blob has a minimum value below which you cannot go. This minimum area is comparable in size to the so-called "Planck constant," roughly 10^{-34} kg m^2/s. That is an extremely small area for macroscopic systems, relatively speaking. But it is big enough to dominate the motion of microscopic systems, like say electrons in atoms.

Key Points

- The Heisenberg uncertainty principle says that there is always a minimum combined uncertainty in position and linear momentum.
- It implies that a particle cannot have a mathematically precise position, because that would require an infinite uncertainty in linear momentum.
- It also implies that a particle cannot have a mathematically precise linear momentum (velocity), since that would imply an infinite uncertainty in position.

3.3 The Operators of Quantum Mechanics

The numerical quantities that the old Newtonian physics uses, (position, momentum, energy, ...), are just "shadows" of what really describes nature: operators. The operators described in this section are the key to quantum mechanics.

As the first example, while a mathematically precise value of the position x of a particle never exists, instead there is an x-position *operator* $\widehat{x}$. It turns the wave function Ψ into $x\Psi$:

$$\Psi(x,y,z,t) \quad \xrightarrow{\quad \widehat{x} \quad} \quad x\Psi(x,y,z,t) \tag{3.3}$$

The operators $\widehat{y}$ and $\widehat{z}$ are defined similarly as $\widehat{x}$.

Instead of a linear momentum $p_x = mu$, there is an x-momentum *operator*

$$\boxed{\widehat{p}_x = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}} \tag{3.4}$$

that turns Ψ into its x-derivative:

$$\Psi(x,y,z,t) \quad \xrightarrow{\quad \widehat{p}_x = \dfrac{\hbar}{\mathrm{i}}\dfrac{\partial}{\partial x} \quad} \quad \frac{\hbar}{\mathrm{i}}\Psi_x(x,y,z,t) \tag{3.5}$$

The constant $\hbar$ is called "Planck's constant." (Or rather, it is Planck's original constant h divided by 2π.) If it would have been zero, all these troubles with quantum mechanics would not occur. The blobs would become points. Unfortunately, $\hbar$ is very small, but nonzero. It is about 10^{-34} kg m^2/s.

The factor i in $\widehat{p}_x$ makes it a Hermitian operator (a proof of that is in derivation {D.9}). All operators reflecting macroscopic physical quantities are Hermitian.

The operators $\widehat{p}_y$ and $\widehat{p}_z$ are defined similarly as $\widehat{p}_x$:

$$\boxed{\widehat{p}_y = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial y} \qquad \widehat{p}_z = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial z}} \tag{3.6}$$

The *kinetic energy operator* $\widehat{T}$ is:

$$\widehat{T} = \frac{\widehat{p}_x^2 + \widehat{p}_y^2 + \widehat{p}_z^2}{2m} \tag{3.7}$$

Its shadow is the Newtonian notion that the kinetic energy equals:

$$T = \frac{1}{2}m\left(u^2+v^2+w^2\right) = \frac{(mu)^2+(mv)^2+(mw)^2}{2m}$$

This is an example of the "Newtonian analogy": the relationships between the different operators in quantum mechanics are in general the same as those between the corresponding numerical values in Newtonian physics. But since the momentum *operators* are gradients, the actual kinetic energy operator is, from the momentum operators above:

$$\widehat{T} = -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right). \tag{3.8}$$

Mathematicians call the set of second order derivative operators in the kinetic energy operator the "Laplacian", and indicate it by ∇^2:

$$\boxed{\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}} \tag{3.9}$$

In those terms, the kinetic energy operator can be written more concisely as:

$$\boxed{\widehat{T} = -\frac{\hbar^2}{2m}\nabla^2} \tag{3.10}$$

Following the Newtonian analogy once more, the total energy operator, indicated by H, is the the sum of the kinetic energy operator above and the potential energy operator $V(x, y, z, t)$:

$$\boxed{H = -\frac{\hbar^2}{2m}\nabla^2 + V} \tag{3.11}$$

This total energy operator H is called the *Hamiltonian* and it is very important. Its eigenvalues are indicated by E (for energy), for example E_1, E_2, E_3, ... with:

$$H\psi_n = E_n\psi_n \quad \text{for } n = 1, 2, 3, ... \tag{3.12}$$

where ψ_n is eigenfunction number n of the Hamiltonian.

It is seen later that in many cases a more elaborate numbering of the eigenvalues and eigenvectors of the Hamiltonian is desirable instead of using a single counter n. For example, for the electron of the hydrogen atom, there is more than one eigenfunction for each different eigenvalue E_n, and additional counters l and m are used to distinguish them. It is usually best to solve the eigenvalue problem first and decide on how to number the solutions afterwards.

(It is also important to remember that in the literature, the Hamiltonian eigenvalue problem is commonly referred to as the "time-independent Schrödinger equation." However, this book prefers to reserve the term Schrödinger equation for the unsteady evolution of the wave function.)

Key Points

- Physical quantities correspond to operators in quantum mechanics.
- Expressions for various important operators were given.
- Kinetic energy is in terms of the so-called Laplacian operator.
- The important total energy operator, (kinetic plus potential energy,) is called the Hamiltonian.

3.4 The Orthodox Statistical Interpretation

In addition to the operators defined in the previous section, quantum mechanics requires rules on how to use them. This section gives those rules, along with a critical discussion what they really mean.

3.4.1 Only eigenvalues

According to quantum mechanics, the only "measurable values" of position, momentum, energy, etcetera, are the *eigenvalues* of the corresponding operator. For example, if the total energy of a particle is "measured", the only numbers that can come out are the eigenvalues of the total energy Hamiltonian.

There is really no controversy that only the eigenvalues come out; this has been verified overwhelmingly in experiments, often to astonishingly many digits accuracy. It is the reason for the line spectra that allow the elements to be recognized, either on earth or halfway across the observable universe, for lasers, for the blackbody radiation spectrum, for the value of the speed of sound, for the accuracy of atomic clocks, for the properties of chemical bonds, for the fact that a Stern-Gerlach apparatus does not fan out a beam of atoms but splits it into discrete rays, and countless other basic properties of nature.

But the question *why and how* only the eigenvalues come out is much more tricky. In general the wave function that describes physics is a *combination* of eigenfunctions, not a single eigenfunction. (Even if the wave function was an eigenfunction of one operator, it would not be one of another operator.) If the wave function is a combination of eigenfunctions, then why is the measured value not a combination, (maybe some average), of eigenvalues, but a *single* eigenvalue? And what happens to the eigenvalues in the combination that do not come out? It is a question that has plagued quantum mechanics since the beginning.

The most generally given answer in the physics community is the "orthodox interpretation." It is commonly referred to as the "Copenhagen Interpretation", though that interpretation, as promoted by Niels Bohr, was actually much more circumspect than what is usually presented.

> *According to the orthodox interpretation, "measurement" causes the wave function Ψ to "collapse" into one of the eigenfunctions of the quantity being measured.*

Staying with energy measurements as the example, any total energy "measurement" will cause the wave function to collapse into one of the eigenfunctions ψ_n of the total energy Hamiltonian. The energy that is measured is the corresponding eigenvalue:

$$\left.\begin{array}{l}\Psi = c_1\psi_1 + c_2\psi_2 + \ldots \\ \text{Energy is uncertain}\end{array}\right\} \xrightarrow{\text{energy measurement}} \left\{\begin{array}{l}\Psi = c_n\psi_n \\ \text{Energy} = E_n\end{array}\right. \quad \text{for } \textit{some } n$$

This story, of course, is nonsense. It makes a distinction between "nature" (the particle, say) and the "measurement device" supposedly producing an exact value. But the measurement device is a part of nature too, and therefore also uncertain. What measures the measurement device?

Worse, there is no definition at all of what "measurement" is or is not, so anything physicists, and philosophers, want to put there goes. Needless to say, theories have proliferated, many totally devoid of common sense. The more reasonable "interpretations of the interpretation" tend to identify measurements as interactions with macroscopic systems. Still, there is no indication how and when a system would be sufficiently macroscopic, and how that would produce a collapse or at least something approximating it.

If that is not bad enough, quantum mechanics *already has* a law, called the Schrödinger equation (chapter 7.1), that says how the wave function evolves. This equation contradicts the collapse, (chapter 8.5.)

The collapse in the orthodox interpretation is what the classical theater world would have called "Deus ex Machina". It is a god that appears out of thin air to make things right. A god that has the power to distort the normal laws of nature at will. Mere humans may not question the god. In fact, physicists tend to actually get upset if you do.

However, it is a fact that after a real-life measurement has been made, further follow-up measurements have statistics that are consistent with a collapsed wave function, (which can be computed.) The orthodox interpretation does describe what happens practically in actual laboratory settings well. It just offers no practical help in circumstances that are not so clear cut, being phrased in terms that are essentially meaningless.

Key Points

- Even if a system is initially in a combination of the eigenfunctions of a physical quantity, a measurement of that quantity pushes the measured system into a single eigenfunction.
- The measured value is the corresponding eigenvalue.

3.4.2 Statistical selection

There is another hot potato besides the collapse itself; it is the selection of the eigenfunction to collapse to. If the wave function before a "measurement" is a combination of many different eigenfunctions, then *what* eigenfunction will the measurement produce? Will it be ψ_1? ψ_2? ψ_{10}?

The answer of the orthodox interpretation is that nature contains a mysterious random number generator. If the wave function Ψ *before* the "measurement"

equals, in terms of the eigenfunctions,

$$\Psi = c_1\psi_1 + c_2\psi_2 + c_3\psi_3 + \ldots$$

then this random number generator will, in Einstein's words, "throw the dice" and select one of the eigenfunctions based on the result. It will collapse the wave function to eigenfunction ψ_1 in on average a fraction $|c_1|^2$ of the cases, it will collapse the wave function to ψ_2 in a fraction $|c_2|^2$ of the cases, etc.

> *The orthodox interpretation says that the square magnitudes of the coefficients of the eigenfunctions give the probabilities of the corresponding eigenvalues.*

This too describes very well what happens practically in laboratory experiments, but offers again no insight into why and when. And the notion that nature would somehow come with, maybe not a physical random number generator, but certainly an endless sequence of *truly* random numbers seemed very hard to believe even for an early pioneer of quantum mechanics like Einstein. Many have proposed that the eigenfunction selections are not truly random, but reflect unobserved "hidden variables" that merely seem random to us humans. Yet, after almost a century, none of these theories have found convincing evidence or general acceptance. Physicists still tend to insist quite forcefully on a *literal* random number generator. Somehow, when belief is based on faith, rather than solid facts, tolerance of alternative views is much less, even among scientists.

While the usual philosophy about the orthodox interpretation can be taken with a big grain of salt, the bottom line to remember is:

> *Random collapse of the wave function, with chances governed by the square magnitudes of the coefficients, is indeed the correct way for us humans to describe what happens in our observations.*

As explained in chapter 8.6, this is despite the fact that the wave function *does not* collapse: the collapse is an artifact produced by limitations in our capability to see the entire picture. We humans have no choice but to work within our limitations, and within these, the rules of the orthodox interpretation do apply.

Schrödinger gave a famous, rather cruel, example of a cat in a box to show how weird the predictions of quantum mechanics really are. It is discussed in chapter 8.1.

Key Points

- If a system is initially in a combination of the eigenfunctions of a physical quantity, a measurement of that quantity picks one of the eigenvalues at random.

- The chances of a given eigenvalue being picked are proportional to the square magnitude of the coefficient of the corresponding eigenfunction in the combination.

3.5 A Particle Confined Inside a Pipe

This section demonstrates the general procedure for analyzing quantum systems using a very elementary example. The system to be studied is that of a particle, say an electron, confined to the inside of a narrow pipe with sealed ends. This example will be studied in some detail, since if you understand it thoroughly, it becomes much easier not to get lost in the more advanced examples of quantum mechanics discussed later. And as the final subsection 3.5.9 shows, as well as much of chapter 6, the particle in a pipe is really quite interesting despite its simplicity.

3.5.1 The physical system

The system to be analyzed is shown in figure 3.4 as it would appear in classical nonquantum physics. A particle is bouncing around between the two ends of a



Figure 3.4: Classical picture of a particle in a closed pipe.

pipe. It is assumed that there is no friction, so the particle will keep bouncing back and forward forever. (Friction is a macroscopic effect that has no place in the sort of quantum-scale systems analyzed here.) Typically, classical physics draws the particles that it describes as little spheres, so that is what figure 3.4 shows.

The actual quantum system to be analyzed is shown in figure 3.5. A particle



Figure 3.5: Quantum mechanics picture of a particle in a closed pipe.

like an electron has no (known) specific shape or size, but it does have a wave function "blob." So in quantum mechanics the equivalent of a particle bouncing around is a wave function blob bouncing around between the ends of the pipe.

Please do not ask what this impenetrable pipe is made off. It is obviously a crude idealization. You could imagine that the electron is a valence electron in

a very tiny bar of copper. In that case the pipe walls would correspond to the surface of the copper bar, and it is assumed that the electron cannot get off the bar.

But of course, a copper bar would have nuclei, and other electrons, and the analysis here does not consider those. So maybe it is better to think of the particle as being a lone helium atom stuck inside a carbon nanotube.

Key Points

- An idealized problem of a particle bouncing about in a pipe will be considered.

3.5.2 Mathematical notations

The first step in the solution process is to describe the problem mathematically. To do so, an x-coordinate that measures longitudinal position inside the pipe will be used, as shown in figure 3.6. Also,the length of the pipe will be called ℓ_x.

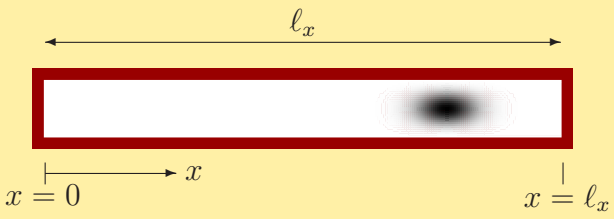


Figure 3.6: Definitions for one-dimensional motion in a pipe.

To make the problem as easy to solve as possible, it will be assumed that the only position coordinate that exists is the longitudinal position x along the pipe. For now, the existence of any coordinates y and z that measure the location in cross section will be completely ignored.

Key Points

- The only position coordinate to be considered for now is x.
- The notations have been defined.

3.5.3 The Hamiltonian

To analyze a quantum system you must find the Hamiltonian. The Hamiltonian is the total energy operator, equal to the sum of kinetic plus potential energy.

The potential energy V is the easiest to find: since it is assumed that the particle does not experience forces inside the pipe, (until it hits the ends of the pipe, that is), the potential energy must be constant inside the pipe:

$$V = \text{constant}$$

(The force is the derivative of the potential energy, so a constant potential energy produces zero force.) Further, since the value of the constant does not make any difference physically, you may as well assume that it is zero and save some writing:

$$V = 0$$

Next, the kinetic energy operator $\widehat{T}$ is needed. You can just look up its precise form in section 3.3 and find it is:

$$\widehat{T} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$$

Note that only the x term is used here; the existence of the other two coordinates y and z is completely ignored. The constant m is the mass of the particle, and $\hbar$ is Planck's constant.

Since the potential energy is zero, the Hamiltonian H is just this kinetic energy:

$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} \tag{3.13}$$

Key Points

- The one-dimensional Hamiltonian (3.13) has been written down.

3.5.4 The Hamiltonian eigenvalue problem

With the Hamiltonian H found, the next step is to formulate the Hamiltonian eigenvalue problem, (or "time-independent Schrödinger equation."). This problem is always of the form

$$H\psi = E\psi$$

Any nonzero solution ψ of this equation is called an energy eigenfunction and the corresponding constant E is called the energy eigenvalue.

Substituting the Hamiltonian for the pipe as found in the previous subsection, the eigenvalue problem is:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = E\psi \tag{3.14}$$

The problem is not complete yet. These problems also need so called "boundary conditions", conditions that say what happens at the *ends* of the x range. In this case, the ends of the x range are the ends of the pipe. Now recall that the square magnitude of the wave function gives the probability of finding the particle. So the wave function must be zero wherever there is no possibility of finding the particle. That is outside the pipe: it is assumed that the particle is confined to the pipe. So the wave function is zero outside the pipe. And since the outside of the pipe starts at the ends of the pipe, that means that the wave function must be zero at the ends {N.5}:

$$\psi = 0 \text{ at } x = 0 \qquad \text{and} \qquad \psi = 0 \text{ at } x = \ell_x \tag{3.15}$$

Key Points

- The Hamiltonian eigenvalue problem (3.14)has been found.
- It also includes the boundary conditions (3.15).

3.5.5 All solutions of the eigenvalue problem

The previous section found the Hamiltonian eigenvalue problem to be:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = E\psi$$

Now you need to solve this equation. Mathematicians call an equation of this type an ordinary differential equation; "differential" because it has a derivative in it, and "ordinary" since there are no derivatives with respect to variables other than x.

If you do not know how to solve ordinary differential equations, it is no big deal. The best way is usually to look them up anyway. The equation above can be found in most mathematical table books, e.g. [41, item 19.7]. According to what it says there, (with changes in notation), if you assume that the energy E is negative, the solution is

$$\psi = C_1 e^{\kappa x} + C_2 e^{-\kappa x} \qquad \kappa = \frac{\sqrt{-2mE}}{\hbar}$$

This solution may easily by checked by simply substituting it into the ordinary differential equation.

As far as the ordinary differential equation is concerned, the constants C_1 and C_2 could be any two numbers. But you also need to satisfy the two boundary conditions given in the previous subsection. The boundary condition that $\psi = 0$ when $x = 0$ produces, if ψ is as above,

$$C_1 e^0 + C_2 e^0 = 0$$

and since $e^0 = 1$, this can be used to find an expression for C_2:

$$C_2 = -C_1$$

The second boundary condition, that $\psi = 0$ at $x = \ell_x$, produces

$$C_1 e^{\kappa \ell_x} + C_2 e^{-\kappa \ell_x} = 0$$

or, since you just found out that $C_2 = -C_1$,

$$C_1 \left(e^{\kappa \ell_x} - e^{-\kappa \ell_x} \right) = 0$$

This equation spells trouble because the term between parentheses cannot be zero; the exponentials are not equal. Instead C_1 will have to be zero; that is bad news since it implies that $C_2 = -C_1$ is zero too, and then so is the wave function ψ:

$$\psi = C_1 e^{\kappa x} + C_2 e^{-\kappa x} = 0$$

A zero wave function is not acceptable, since there would be no possibility to find the particle anywhere!

Everything was done right. So the problem must be the initial assumption that the energy is negative. Apparently, the energy cannot be negative. This can be understood from the fact that for this particle, the energy is all kinetic energy. Classical physics would say that the kinetic energy cannot be negative because it is proportional to the square of the velocity. You now see that quantum mechanics agrees that the kinetic energy cannot be negative, but says it is because of the boundary conditions on the wave function.

Try again, but now assume that the energy E is zero instead of negative. In that case the solution of the ordinary differential equation is according to [41, item 19.7]

$$\psi = C_1 + C_2 x$$

The boundary condition that $\psi = 0$ at $x = 0$ now produces:

$$C_1 + C_2 0 = C_1 = 0$$

so C_1 must be zero. The boundary condition that $\psi = 0$ at $x = \ell_x$ gives:

$$0 + C_2 \ell_x = 0$$

so C_2 must be zero too. Once again there is no nonzero solution, so the assumption that the energy E can be zero must be wrong too.

Note that classically, it is perfectly OK for the energy to be zero: it would simply mean that the particle is sitting in the pipe at rest. But in quantum mechanics, zero kinetic energy is not acceptable, and it all has to do with Heisenberg's uncertainty principle. Since the particle is restricted to the inside

of the pipe, its position is constrained, and so the uncertainty principle requires that the linear momentum must be uncertain. Uncertain momentum cannot be zero momentum; measurements will show a range of values for the momentum of the particle, implying that it is in motion and therefore has kinetic energy.

Try, try again. The only possibility left is that the energy E is positive. In that case, the solution of the ordinary differential equation is according to [41, item 19.7]:

$$\psi = C_1 \cos(kx) + C_2 \sin(kx) \qquad k = \frac{\sqrt{2mE}}{\hbar}$$

Here the constant k is called the "wave number."

The boundary condition that $\psi = 0$ at $x = 0$ is:

$$C_1 1 + C_2 0 = C_1 = 0$$

so C_1 must be zero. The boundary condition $\psi = 0$ at $x = \ell_x$ is then:

$$0 + C_2 \sin(k\ell_x) = 0$$

There finally is possibility to get a nonzero coefficient C_2: this equation can be satisfied if $\sin(k\ell_x) = 0$ instead of C_2. In fact, there is not just one possibility for this to happen: a sine is zero when its argument equals π, 2π, 3π, ... So there is a nonzero solution for each of the following values of the positive constant k:

$$k = \frac{\pi}{\ell_x},\ k = \frac{2\pi}{\ell_x},\ k = \frac{3\pi}{\ell_x},\ \ldots$$

Each of these possibilities gives one solution ψ. Different solutions ψ will be distinguished by giving them a numeric subscript:

$$\psi_1 = C_2 \sin\left(\frac{\pi}{\ell_x} x\right),\ \psi_2 = C_2 \sin\left(\frac{2\pi}{\ell_x} x\right),\ \psi_3 = C_2 \sin\left(\frac{3\pi}{\ell_x} x\right),\ \ldots$$

The generic solution can be written more concisely using a counter n as:

$$\psi_n = C_2 \sin\left(\frac{n\pi}{\ell_x} x\right) \quad \text{for } n = 1, 2, 3, \ldots$$

Let's check the solutions. Clearly each is zero when $x = 0$ and when $x = \ell_x$. Also, substitution of each of the solutions into the ordinary differential equation

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} = E\psi$$

shows that they all satisfy it, provided that their energy values are, respectively:

$$E_1 = \frac{\hbar^2 \pi^2}{2m\ell_x^2},\ E_2 = \frac{2^2 \hbar^2 \pi^2}{2m\ell_x^2},\ E_3 = \frac{3^2 \hbar^2 \pi^2}{2m\ell_x^2},\ \ldots$$

or generically:

$$E_n = \frac{n^2\hbar^2\pi^2}{2m\ell_x^2} \quad \text{for } n = 1, 2, 3, \ldots$$

There is one more condition that must be satisfied: each solution must be normalized so that the total probability of finding the particle integrated over all possible positions is 1 (certainty). That requires:

$$1 = \langle\psi_n|\psi_n\rangle = \int_{x=0}^{\ell_x} |C_2|^2 \sin^2\left(\frac{n\pi}{\ell_x}x\right) \mathrm{d}x$$

which after integration fixes C_2 (assuming you choose it to be a positive real number):

$$C_2 = \sqrt{\frac{2}{\ell_x}}$$

Summarizing the results of this subsection, there is not just one energy eigenfunction and corresponding eigenvalue, but an infinite set of them:

$$\begin{aligned}
\psi_1 &= \sqrt{\frac{2}{\ell_x}}\sin\left(\frac{\pi}{\ell_x}x\right) \quad E_1 = \frac{\hbar^2\pi^2}{2m\ell_x^2} \\
\psi_2 &= \sqrt{\frac{2}{\ell_x}}\sin\left(\frac{2\pi}{\ell_x}x\right) \quad E_2 = \frac{2^2\hbar^2\pi^2}{2m\ell_x^2} \\
\psi_3 &= \sqrt{\frac{2}{\ell_x}}\sin\left(\frac{3\pi}{\ell_x}x\right) \quad E_3 = \frac{3^2\hbar^2\pi^2}{2m\ell_x^2} \\
&\vdots \qquad\qquad\qquad\qquad \vdots
\end{aligned} \tag{3.16}$$

or in generic form:

$$\psi_n = \sqrt{\frac{2}{\ell_x}}\sin\left(\frac{n\pi}{\ell_x}x\right) \quad E_n = \frac{n^2\hbar^2\pi^2}{2m\ell_x^2} \quad \text{for } n = 1, 2, 3, 4, 5, \ldots \tag{3.17}$$

The next thing will be to take a better look at these results.

Key Points

- After a lot of grinding mathematics armed with table books, the energy eigenfunctions and eigenvalues have finally been found
- There are infinitely many of them.
- They are as listed in (3.17) above. The first few are also written out explicitly in (3.16).

3.5.5 Review Questions

1. Write down eigenfunction number 6.
   Solution piped-a
2. Write down eigenvalue number 6.
   Solution piped-b

3.5.6 Discussion of the energy values

This subsection discusses the energy that the particle in the pipe can have. It was already discovered in the previous subsection that the particle cannot have negative energy, nor zero energy. In fact, according to the orthodox interpretation, the only values that the total energy of the particle can take are the energy eigenvalues

$$E_1 = \frac{\hbar^2\pi^2}{2m\ell_x^2},\ E_2 = \frac{2^2\hbar^2\pi^2}{2m\ell_x^2},\ E_3 = \frac{3^2\hbar^2\pi^2}{2m\ell_x^2},\ \ldots$$

derived in the previous subsection.

Energy values are typically shown graphically in the form of an "energy spectrum", as in figure 3.7. Energy is plotted upwards, so the vertical height

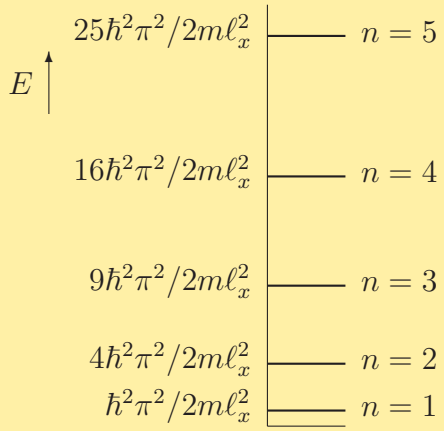


Figure 3.7: One-dimensional energy spectrum for a particle in a pipe.

of each energy level indicates the amount of energy it has. To the right of each energy level, the solution counter, or "quantum number", n is listed.

Classically, the total energy of the particle can have any nonnegative value. But according to quantum mechanics, that is not true: the total energy must be one of the levels shown in the energy spectrum figure 3.7. It should be noted that for a macroscopic particle, you would not know the difference; the spacing between the energy levels is macroscopically very fine, since Planck's constant $\hbar$ is so small. However, for a quantum-scale system, the discreteness of the energy values can make a major difference.

Another point: at absolute zero temperature, the particle will be stuck in the lowest possible energy level, $E_1 = \hbar^2\pi^2/2m\ell_x^2$, in the spectrum figure 3.7.

This lowest possible energy level is called the "ground state." Classically you would expect that at absolute zero the particle has no kinetic energy, so zero total energy. But quantum mechanics does not allow it. Heisenberg's principle requires some momentum, hence kinetic energy to remain for a confined particle even at zero temperature.

Key Points

- Energy values can be shown as an energy spectrum.
- The possible energy levels are discrete.
- But for a macroscopic particle, they are extremely close together.
- The ground state of lowest energy has nonzero kinetic energy.

3.5.6 Review Questions

1. Plug the mass of an electron, $m = 9.109\,38\ 10^{-31}$ kg, and the rough size of an hydrogen atom, call it $\ell_x = 2\ 10^{-10}$ m, into the expression for the ground state kinetic energy and see how big it is. Note that $\hbar = 1.054\,57\ 10^{-34}$ J s. Express in units of eV, where one eV equals $1.602\,18\ 10^{-19}$ J.
   Solution pipee-a
2. Just for fun, plug macroscopic values, $m = 1$ kg and $\ell_x = 1$ m, into the expression for the ground state energy and see how big it is. Note that $\hbar = 1.054\,57\ 10^{-34}$ J s.
   Solution pipee-b
3. What is the eigenfunction number, or quantum number, n that produces a macroscopic amount of energy, 1 J, for macroscopic values $m = 1$ kg and $\ell_x = 1$ m? With that many energy levels involved, would you see the difference between successive ones?
   Solution pipee-c

3.5.7 Discussion of the eigenfunctions

This subsection discusses the one-dimensional energy eigenfunctions of the particle in the pipe. The solution of subsection 3.5.5 found them to be:

$$\psi_1 = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{\pi}{\ell_x}x\right),\ \psi_2 = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{2\pi}{\ell_x}x\right),\ \psi_3 = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{3\pi}{\ell_x}x\right),\ \ldots$$

The first one to look at is the ground state eigenfunction

$$\psi_1 = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{\pi}{\ell_x}x\right).$$

It is plotted at the top of figure 3.8. As noted in section 3.1, it is the *square* magnitude of a wave function that gives the probability of finding the particle.

So, the second graph in figure 3.8 shows the square of the ground state wave function, and the higher values of this function then give the locations where the particle is more likely to be found. This book shows regions where the particle is more likely to be found as darker regions, and in those terms the probability of finding the particle is as shown in the bottom graphic of figure 3.8. It is seen

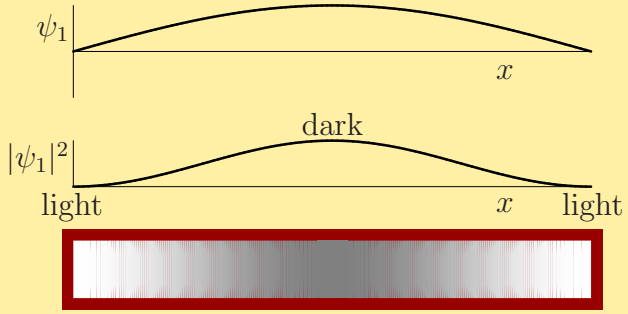


Figure 3.8: One-dimensional ground state of a particle in a pipe.

that in the ground state, the particle is much more likely to be found somewhere in the middle of the pipe than close to the ends.

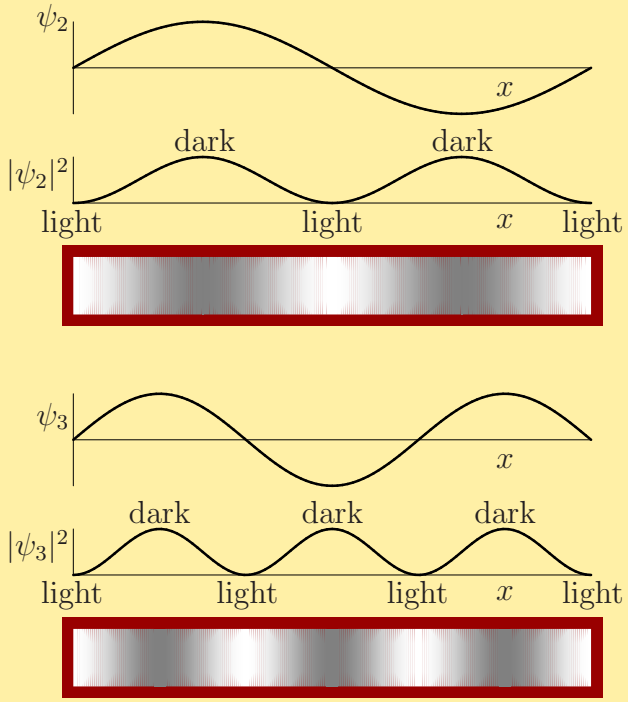


Figure 3.9: Second and third lowest one-dimensional energy states.

Figure 3.9 shows the two next lowest energy states

$$\psi_2 = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{2\pi}{\ell_x} x\right) \text{ and } \psi_3 = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{3\pi}{\ell_x} x\right)$$

as grey tones. Regions where the particle is relatively likely to be found alternate with ones where it is less likely to be found. And the higher the energy, the more such regions there are. Also note that in sharp contrast to the ground state, for eigenfunction ψ_2 there is almost no likelihood of finding the particle close to the center.

Needless to say, none of those energy states looks at all like the wave function blob bouncing around in figure 3.5. Moreover, it turns out that energy eigenstates are stationary states: the probabilities shown in figures 3.8 and 3.9 do not change with time.

In order to describe a localized wave function blob bouncing around, states of different energy must be combined. It will take until chapter 7.11.4 before the analytical tools to do so have been described. For now, the discussion must remain restricted to just finding the energy levels. And these are important enough by themselves anyway, sufficient for many practical applications of quantum mechanics.

Key Points

- In the energy eigenfunctions, the particle is not localized to within any particular small region of the pipe.
- In general there are regions where the particle may be found separated by regions in which there is little chance to find the particle.
- The higher the energy level, the more such regions there are.

3.5.7 Review Questions

1. So how does, say, the one-dimensional eigenstate ψ_6 look?
   Solution pipef-a
2. Generalizing the results above, for eigenfunction ψ_n, any n, how many distinct regions are there where the particle may be found?
   Solution pipef-b
3. If you are up to a trick question, consider the following. There are no forces inside the pipe, so the particle has to keep moving until it hits an end of the pipe, then reflect backward until it hits the other side and so on. So, it has to cross the center of the pipe regularly. But in the energy eigenstate ψ_2, the particle has *zero* chance of ever being found at the center of the pipe. What gives?
   Solution pipef-c

3.5.8 Three-dimensional solution

The solution for the particle stuck in a pipe that was obtained in the previous subsections cheated. It pretended that there was only one spatial coordinate x.

Real life is three-dimensional. And yes, as a result, the solution as obtained is simply wrong.

Fortunately, it turns out that you can fix up the problem pretty easily if you assume that the pipe has a square cross section. There is a way of combining one-dimensional solutions for all three coordinates into full three-dimensional solutions. This is called the "separation of variables" idea: Solve each of the three variables x, y, and z separately, then combine the results.

The full coordinate system for the problem is shown in figure 3.10: in addition to the x-coordinate along the length of the pipe, there is also a y-coordinate giving the vertical position in cross section, and similarly a z-coordinate giving the position in cross section towards you.

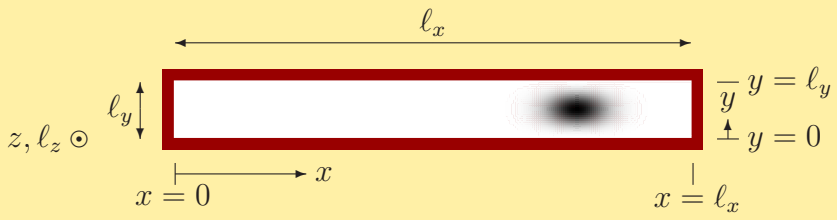


Figure 3.10: Definition of all variables for motion in a pipe.

Now recall the one-dimensional solutions that were obtained assuming there is just an x-coordinate, but add subscripts "x" to keep them apart from any solutions for y and z:

$$\begin{aligned}
\psi_{x1} &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{\pi}{\ell_x} x\right) \quad & E_{x1} &= \frac{\hbar^2 \pi^2}{2m\ell_x^2} \\
\psi_{x2} &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{2\pi}{\ell_x} x\right) \quad & E_{x2} &= \frac{2^2 \hbar^2 \pi^2}{2m\ell_x^2} \\
\psi_{x3} &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{3\pi}{\ell_x} x\right) \quad & E_{x3} &= \frac{3^2 \hbar^2 \pi^2}{2m\ell_x^2} \\
&\vdots & &\vdots
\end{aligned} \tag{3.18}$$

or in generic form:

$$\psi_{xn_x} = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{n_x \pi}{\ell_x} x\right) \quad E_{xn_x} = \frac{n_x^2 \hbar^2 \pi^2}{2m\ell_x^2} \quad \text{for } n_x = 1, 2, 3, \ldots \tag{3.19}$$

Since it is assumed that the cross section of the pipe is square or rectangular of dimensions $\ell_y \times \ell_z$, the y and z directions have *one-dimensional* solutions

completely equivalent to the x direction:

$$\psi_{yn_y} = \sqrt{\frac{2}{\ell_y}} \sin\left(\frac{n_y \pi}{\ell_y} y\right) \quad E_{yn_y} = \frac{n_y^2 \hbar^2 \pi^2}{2m\ell_y^2} \quad \text{for } n_y = 1, 2, 3, \ldots \tag{3.20}$$

and

$$\psi_{zn_z} = \sqrt{\frac{2}{\ell_z}} \sin\left(\frac{n_z \pi}{\ell_z} z\right) \quad E_{zn_z} = \frac{n_z^2 \hbar^2 \pi^2}{2m\ell_z^2} \quad \text{for } n_z = 1, 2, 3, \ldots \tag{3.21}$$

After all, there is no fundamental difference between the three coordinate directions; each is along an edge of a rectangular box.

Now it turns out, {D.11}, that the full three-dimensional problem has eigenfunctions $\psi_{n_x n_y n_z}$ that are simply *products* of the one-dimensional ones:

$$\boxed{\psi_{n_x n_y n_z} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{n_x \pi}{\ell_x} x\right) \sin\left(\frac{n_y \pi}{\ell_y} y\right) \sin\left(\frac{n_z \pi}{\ell_z} z\right)} \tag{3.22}$$

There is one such three-dimensional eigenfunction for each *set* of three numbers (n_x, n_y, n_z). These numbers are the three "quantum numbers" of the eigenfunction.

Further, the energy eigenvalues $E_{n_x n_y n_z}$ of the three-dimensional problem are the *sum* of those of the one-dimensional problems:

$$\boxed{E_{n_x n_y n_z} = \frac{n_x^2 \hbar^2 \pi^2}{2m\ell_x^2} + \frac{n_y^2 \hbar^2 \pi^2}{2m\ell_y^2} + \frac{n_z^2 \hbar^2 \pi^2}{2m\ell_z^2}} \tag{3.23}$$

For example, the ground state of lowest energy occurs when all three quantum numbers are lowest, $n_x = n_y = n_z = 1$. The three-dimensional ground state wave function is therefore:

$$\psi_{111} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{\pi}{\ell_x} x\right) \sin\left(\frac{\pi}{\ell_y} y\right) \sin\left(\frac{\pi}{\ell_z} z\right) \tag{3.24}$$

This ground state is shown in figure 3.11. The y and z factors ensure that the wave function is now zero at all the surfaces of the pipe.

The ground state energy is:

$$E_{111} = \frac{\hbar^2 \pi^2}{2m\ell_x^2} + \frac{\hbar^2 \pi^2}{2m\ell_y^2} + \frac{\hbar^2 \pi^2}{2m\ell_z^2} \tag{3.25}$$

Since the cross section dimensions ℓ_y and ℓ_z are small compared to the length of the pipe, the last two terms are large compared to the first one. They make

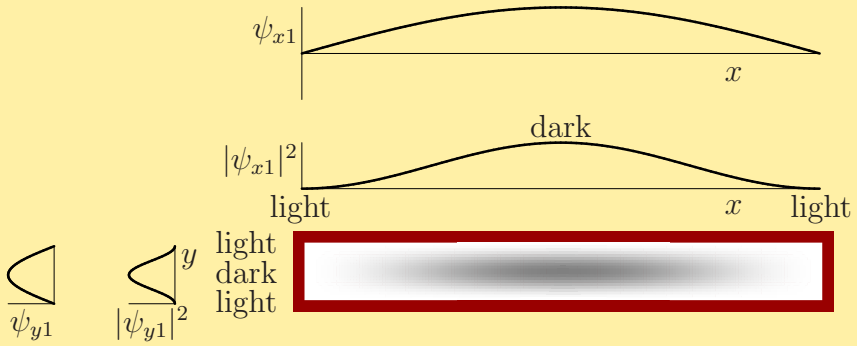


Figure 3.11: True ground state of a particle in a pipe.



Figure 3.12: True second and third lowest energy states.

the true ground state energy much larger than the one-dimensional value, which was just the first term.

The next two lowest energy levels occur for $n_x = 2$, $n_y = n_z = 1$ respectively $n_x = 3$, $n_y = n_z = 1$. (The latter assumes that the cross section dimensions are small enough that the alternative possibilities $n_y = 2$, $n_x = n_z$ =1 and n_z =2, $n_x = n_y$ =1 have more energy.) The energy eigenfunctions

$$\psi_{211} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{2\pi}{\ell_x}x\right) \sin\left(\frac{\pi}{\ell_y}y\right) \sin\left(\frac{\pi}{\ell_z}z\right) \tag{3.26}$$

$$\psi_{311} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{3\pi}{\ell_x}x\right) \sin\left(\frac{\pi}{\ell_y}y\right) \sin\left(\frac{\pi}{\ell_z}z\right) \tag{3.27}$$

are shown in figure 3.12. They have energy levels:

$$E_{211} = \frac{4\hbar^2\pi^2}{2m\ell_x^2} + \frac{\hbar^2\pi^2}{2m\ell_y^2} + \frac{\hbar^2\pi^2}{2m\ell_z^2} \qquad E_{311} = \frac{9\hbar^2\pi^2}{2m\ell_x^2} + \frac{\hbar^2\pi^2}{2m\ell_y^2} + \frac{\hbar^2\pi^2}{2m\ell_z^2} \tag{3.28}$$

Key Points

- Three-dimensional energy eigenfunctions can be found as products of one-dimensional ones.
- Three-dimensional energies can be found as sums of one-dimensional ones.

Example three-dimensional eigenstates have been shown.

3.5.8 Review Questions

1. If the cross section dimensions ℓ_y and ℓ_z are one tenth the size of the pipe length, how much bigger are the energies E_{y1} and E_{z1} compared to E_{x1}? So, by what percentage is the one-dimensional ground state energy E_{x1} as an approximation to the three-dimensional one, E_{111}, then in error?
   Solution pipeg-a
2. At what ratio of ℓ_y/ℓ_x does the energy E_{121} become higher than the energy E_{311}?
   Solution pipeg-b
3. Shade the regions where the particle is likely to be found in the ψ_{322} energy eigenstate.
   Solution pipeg-c

3.5.9 Quantum confinement

Normally, motion in physics occurs in three dimensions. Even in a narrow pipe, in classical physics a point particle of zero size would be able to move in all three directions. But in quantum mechanics, if the pipe gets very narrow, the motion becomes truly one-dimensional.

To understand why, the first problem that must be addressed is what "motion" means in the first place, because normally motion is defined as change in position, and in quantum mechanics particles *do not have* a well-defined position.

Consider the particle in the ground state of lowest energy, shown in figure 3.11. This is one boring state; the picture never changes. You might be surprised by that; after all, it was found that the ground state has energy, and it is all kinetic energy. If the particle has kinetic energy, should not the positions where the particle is likely to be found change with time?

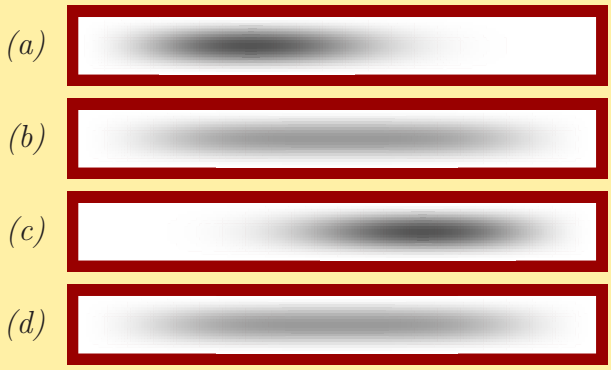
The answer is no; kinetic energy is *not* directly related to changes in likely positions of a particle; that is only an *approximation* valid for macroscopic systems. It is not necessarily true for quantum-scale systems, certainly not if they are in the ground state. Like it or not, in quantum mechanics kinetic energy is second-order derivatives of the wave function, and nothing else.

Next, as already pointed out, all the other energy eigenstates, like those in figure 3.12, have the same boring property of not changing with time.

Things only become somewhat interesting when you combine states of different energy. As the simplest possible example, consider the possibility that the particle has the wave function:

$$\Psi = \sqrt{\tfrac{4}{5}}\psi_{111} + \sqrt{\tfrac{1}{5}}\psi_{211}$$

at some starting time, which will be taken as $t = 0$. According to the orthodox interpretation, in an energy measurement this particle would have a $\frac{4}{5} = 80\%$ chance of being found at the ground state energy E_{111} and a 20% chance of being found at the elevated energy level E_{211}. So there is now uncertainty in energy; that is critical.



Animation: http://www.dommelen.net/quansup/pipemv.gif

Figure 3.13: A combination of ψ_{111} and ψ_{211} seen at some typical times.

In chapter 7.1 it will be found that for nonzero times, the wave function of this particle is given by

$$\Psi = \sqrt{\tfrac{4}{5}} e^{-\mathrm{i}E_{111}t/\hbar}\psi_{111} + \sqrt{\tfrac{1}{5}} e^{-\mathrm{i}E_{211}t/\hbar}\psi_{211}.$$

Using this expression, the probability of finding the particle, $|\Psi|^2$, can be plotted for various times. That is done in figure 3.13 for four typical times. It shows that with uncertainty in energy, the wave function blob does move. It performs a periodic oscillation: after figure 3.13*(d)*, the wave function returns to state 3.13*(a)*, and the cycle repeats.

You would not yet want to call the particle localized, but at least the locations where the particle can be found are now bouncing back and forwards between the ends of the pipe. And if you add additional wave functions ψ_{311}, ψ_{411}, ..., you can get closer and closer to a localized wave function blob bouncing around.

But if you look closer at figure 3.13, you will note that the wave function blob does not move at all in the y-direction; it remains at all times centered around the horizontal pipe centerline. It may seem that this is no big deal; just add one or more wave functions with an n_y value greater than one, like ψ_{121}, and bingo, there will be interesting motion in the y-direction too.

But there is a catch, and it has to do with the required energy. According to the previous section, the kinetic energy in the y-direction takes the values

$$E_{y1} = \frac{\hbar^2\pi^2}{2m\ell_y^2},\ E_{y2} = \frac{4\hbar^2\pi^2}{2m\ell_y^2},\ E_{y3} = \frac{9\hbar^2\pi^2}{2m\ell_y^2},\ \ldots$$

That will be very large energies for a narrow pipe in which ℓ_y is small. The particle will certainly have the large energy E_{y1} in the y-direction; if it is in the pipe at all it has at least that amount of energy. But if the pipe is really narrow, it will simply not have enough *additional*, say thermal, energy to get anywhere close to the next level E_{y2}. The kinetic energy in the y-direction will therefore be stuck at the lowest possible level E_{y1}.

The result is that absolutely nothing interesting goes on in the y-direction. As far as a particle in a narrow pipe is concerned, the y-direction might just as well not exist. It is ironic that while the kinetic energy in the y-direction, E_{y1}, is very large, nothing actually happens in that direction.

If the pipe is also narrow in the z-direction, the only interesting motion is in the x-direction, making the nontrivial physics truly one-dimensional. It becomes a "quantum wire". However, if the pipe size in the z-direction is relatively wide, the particle will have lots of different energy states in the z-direction available too and the motion will be two-dimensional, a "quantum well". Conversely, if the pipe is narrow in all three directions, you get a zero-dimensional "quantum dot" in which the particle does nothing unless it gets a sizable chunk of energy.

An isolated atom can be regarded as an example of a quantum dot; the electrons are confined to a small region around the nucleus and will be at a single energy level unless they are given a considerable amount of energy. But note that when people talk about quantum confinement, they are normally talking about semi-conductors, for which similar effects occur at significantly larger scales, maybe tens of times as large, making them much easier to manufacture. An actual quantum dot is often referred to as an "artificial atom", and has similar properties as a real atom.

It may give you a rough idea of all the interesting things you can do in nanotechnology when you restrict the motion of particles, in particular of electrons, in various directions. You truly change the dimensionality of the normal three-dimensional world into a lower dimensional one. Only quantum mechanics can explain why, by making the energy levels discrete instead of continuously varying. And the lower dimensional worlds can have your choice of topology (a ring, a letter 8, a sphere, a cylinder, a Möbius strip?, ...,) to make things really exciting.

Key Points

- Quantum mechanics allows you to create lower-dimensional worlds for particles.

Chapter 4

Single-Particle Systems

Abstract

In this chapter, the machinery to deal with single particles is worked out, culminating in the vital solutions for the hydrogen atom and hydrogen molecular ion.

The first section covers the harmonic oscillator. This vibrating system is a simple model for such systems as an atom in a trap, crystal vibrations, and electromagnetic waves.

Next, before the hydrogen atom can be discussed, first the quantum mechanics of angular momentum needs to be covered. Just like you need angular momentum to solve the motion of a planet around the sun in classical physics, so do you need angular momentum for the motion of an electron around a nucleus in quantum mechanics. The eigenvalues of angular momentum and their quantum numbers are critically important for many other reasons besides the hydrogen atom.

After angular momentum, the hydrogen atom can be discussed. The solution is messy, but fundamentally not much different from that of the particle in the pipe or the harmonic oscillator of the previous chapter.

The hydrogen atom is the major step towards explaining heavier atoms and then chemical bonds. One rather unusual chemical bond can already be discussed in this chapter: that of a ionized hydrogen molecule. A hydrogen molecular ion has only one electron.

But the hydrogen molecular ion cannot readily be solved exactly, even if the motion of the nuclei is ignored. So an approximate method will be used. Before this can be done, however, a problem must be addressed. The hydrogen molecular ion ground state is defined to be the state of lowest energy. But an approximate ground state is not an exact energy eigenfunction and has uncertain energy. So how should the term "lowest energy" be defined for the approximation?

To answer that, before tackling the molecular ion, first systems with uncertainty in a variable of interest are discussed. The "expectation value" of a variable will be defined to be the average of the eigenvalues, weighted by their probability. The "standard deviation" will be defined as a measure of how much uncertainty there is to that expectation value.

With a precise mathematical definition of uncertainty, the obvious next question is whether two different variables can be certain at the same time. The "commutator" of the two operators will be introduced to answer it. That then allows the Heisenberg uncertainty relationship to be formulated. Not only can position and linear momentum not be certain at the same time; a specific equation can be written down for how big the uncertainty must be, at the very least.

With the mathematical machinery of uncertainty defined, the hydrogen molecular ion is solved last.

4.1 The Harmonic Oscillator

This section provides an in-depth discussion of a basic quantum system. The case to be analyzed is a particle that is constrained by some kind of forces to remain at approximately the same position. This can describe systems such as an atom in a solid or in a molecule. If the forces pushing the particle back to its nominal position are proportional to the distance that the particle moves away from it, you have what is called an harmonic oscillator. Even if the forces vary nonlinearly with position, they can often still be approximated to vary linearly as long as the distances from the nominal position remain small.

The particle's displacement from the nominal position will be indicated by (x, y, z). The forces keeping the particle constrained can be modeled as springs, as sketched in figure 4.1. The stiffness of the springs is characterized by the

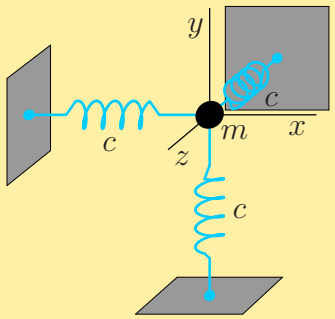


Figure 4.1: Classical picture of an harmonic oscillator.

so called "spring constant" c, giving the ratio between force and displacement. Note that it will be assumed that the three spring stiffnesses are equal.

For a quantum picture of a harmonic oscillator, imagine a light atom like a carbon atom surrounded by much heavier atoms. When the carbon atom tries to move away from its nominal position, the heavy atoms push it back. The harmonic oscillator is also the basic relativistic model for the quantum electromagnetic field.

According to classical Newtonian physics, the particle vibrates back and forth around its nominal position with a frequency

$$\omega = \sqrt{\frac{c}{m}} \tag{4.1}$$

in radians per second. In quantum mechanics, a particle does not have a precise position. But the natural frequency above remains a convenient computational quantity in the quantum solution.

Key Points

- The system to be described is that of a particle held in place by forces that increase proportional to the distance that the particle moves away from its equilibrium position.
- The relation between distance and force is assumed to be the same in all three coordinate directions.
- Number c is a measure of the strength of the forces and ω is the frequency of vibration according to classical physics.

4.1.1 The Hamiltonian

In order to find the energy levels that the oscillating particle can have, you must first write down the total energy Hamiltonian.

As far as the potential energy is concerned, the spring in the x-direction holds an amount of potential energy equal to $\frac{1}{2}cx^2$, and similarly the ones in the y and z directions.

To this total potential energy, you need to add the kinetic energy operator $\widehat{T}$ from section 3.3 to get the Hamiltonian:

$$H = -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + \tfrac{1}{2}c\left(x^2 + y^2 + z^2\right) \tag{4.2}$$

Key Points

- The Hamiltonian (4.2) has been found.

4.1.2 Solution using separation of variables

This section finds the energy eigenfunctions and eigenvalues of the harmonic oscillator using the Hamiltonian as found in the previous subsection. Every energy eigenfunction ψ and its eigenvalue E must satisfy the Hamiltonian eigenvalue problem, (or "time-independent Schrödinger equation"):

$$\left[-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+\tfrac{1}{2}c\left(x^2+y^2+z^2\right)\right]\psi = E\psi \tag{4.3}$$

The boundary condition is that ψ becomes zero at large distance from the nominal position. After all, the magnitude of ψ tells you the relative probability of finding the particle at that position, and because of the rapidly increasing potential energy, the chances of finding the particle very far from the nominal position should be vanishingly small.

Like for the particle in the pipe of the previous section, it will be assumed that each eigenfunction is a product of *one-dimensional* eigenfunctions, one in each direction:

$$\psi = \psi_x(x)\psi_y(y)\psi_z(z) \tag{4.4}$$

Finding the eigenfunctions and eigenvalues by making such an assumption is known in mathematics as the "method of separation of variables".

Substituting the assumption in the eigenvalue problem above, and dividing everything by $\psi_x(x)\psi_y(y)\psi_z(z)$ reveals that E consists of three parts that will be called E_x, E_y, and E_z:

$$\begin{aligned}
E &= E_x + E_y + E_z \\
E_x &= -\frac{\hbar^2}{2m}\frac{\psi_x''(x)}{\psi_x(x)} + \tfrac{1}{2}cx^2 \\
E_y &= -\frac{\hbar^2}{2m}\frac{\psi_y''(y)}{\psi_y(y)} + \tfrac{1}{2}cy^2 \\
E_z &= -\frac{\hbar^2}{2m}\frac{\psi_z''(z)}{\psi_z(z)} + \tfrac{1}{2}cz^2
\end{aligned} \tag{4.5}$$

where the primes indicate derivatives. The three parts represent the x, y, and z dependent terms.

By the definition above, the quantity E_x can only depend on x; variables y and z do not appear in its definition. But actually, E_x cannot depend on x either, since $E_x = E - E_y - E_z$, and none of those quantities depends on x. The inescapable conclusion is that E_x must be a constant, independent of all three variables (x, y, z). The same way E_y and E_z must be constants.

If now in the definition of E_x above, both sides are multiplied by $\psi_x(x)$, a one-dimensional eigenvalue problem results:

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \tfrac{1}{2}cx^2\right]\psi_x = E_x\psi_x \tag{4.6}$$

The operator within the square brackets here, call it H_x, involves only the x-related terms in the full Hamiltonian. Similar problems can be written down for E_y and E_z. Separate problems in each of the three variables x, y, and z have been obtained, explaining why this mathematical method is called separation of variables.

Solving the one-dimensional problem for ψ_x can be done by fairly elementary but elaborate means. If you are interested, you can find how it is done in derivation {D.12}, but that is mathematics and it will not teach you much about quantum mechanics. It turns out that, like for the particle in the pipe of the previous section, there is again an infinite number of different solutions for E_x and ψ_x:

$$\begin{aligned} E_{x0} &= \tfrac{1}{2}\hbar\omega \quad \psi_{x0}(x) = h_0(x) \\ E_{x1} &= \tfrac{3}{2}\hbar\omega \quad \psi_{x1}(x) = h_1(x) \\ E_{x2} &= \tfrac{5}{2}\hbar\omega \quad \psi_{x2}(x) = h_2(x) \\ &\vdots \qquad\qquad \vdots \end{aligned} \tag{4.7}$$

Unlike for the particle in the pipe, here by convention the solutions are numbered starting from 0, rather than from 1. So the first eigenvalue is E_{x0} and the first eigenfunction ψ_{x0}. That is just how people choose to do it.

Also, the eigenfunctions are not sines like for the particle in the pipe; instead, as table 4.1 shows, they take the form of some polynomial times an exponential. But you will probably really not care much about what kind of functions they are anyway unless you end up writing a textbook on quantum mechanics and have to plot them. In that case, you can find a general expression, (D.4), in derivation {D.12}.

But the eigenvalues are what you want to remember from this solution. According to the orthodox interpretation, these are the measurable values of the total energy in the x-direction (potential energy in the x-direction spring plus kinetic energy of the motion in the x-direction.) Instead of writing them all out as was done above, they can be described using the generic expression:

$$E_{xn_x} = \frac{2n_x + 1}{2}\hbar\omega \quad \text{for } n_x = 0, 1, 2, 3, \ldots \tag{4.8}$$

The eigenvalue problem has now been solved, because the equations for Y and Z are mathematically the same and must therefore have corresponding solutions:

$$E_{yn_y} = \frac{2n_y + 1}{2}\hbar\omega \quad \text{for } n_y = 0, 1, 2, 3, \ldots \tag{4.9}$$

$h_0(x) = \dfrac{1}{(\pi\ell^2)^{1/4}}\, e^{-\xi^2/2}$	
$h_1(x) = \dfrac{2\xi}{(4\pi\ell^2)^{1/4}}\, e^{-\xi^2/2}$	$\omega = \sqrt{\dfrac{c}{m}}$
$h_2(x) = \dfrac{2\xi^2 - 1}{(4\pi\ell^2)^{1/4}}\, e^{-\xi^2/2}$	$\ell = \sqrt{\dfrac{\hbar}{m\omega}}$
$h_3(x) = \dfrac{2\xi^3 - 3\xi}{(9\pi\ell^2)^{1/4}}\, e^{-\xi^2/2}$	$\xi = \dfrac{x}{\ell}$
$h_4(x) = \dfrac{4\xi^4 - 12\xi^2 + 3}{(576\pi\ell^2)^{1/4}}\, e^{-\xi^2/2}$	

Table 4.1: First few one-dimensional eigenfunctions of the harmonic oscillator.

$$E_{zn_z} = \frac{2n_z + 1}{2}\hbar\omega \quad \text{for } n_z = 0, 1, 2, 3, \ldots \tag{4.10}$$

The total energy E of the complete system is the sum of E_x, E_y, and E_z. Any nonnegative choice for number n_x, combined with any nonnegative choice for number n_y, and for n_z, produces *one* combined total energy value $E_{xn_x} + E_{yn_y} + E_{zn_z}$, which will be indicated by $E_{n_x n_y n_z}$. Putting in the expressions for the three partial energies above, these total energy eigenvalues become:

$$E_{n_x n_y n_z} = \frac{2n_x + 2n_y + 2n_z + 3}{2}\,\hbar\omega \tag{4.11}$$

where the "quantum numbers" n_x, n_y, and n_z may each have any value in the range 0, 1, 2, 3, ...

The corresponding eigenfunction of the complete system is:

$$\psi_{n_x n_y n_z} = h_{n_x}(x) h_{n_y}(y) h_{n_z}(z) \tag{4.12}$$

where the functions h_0, h_1, ... are in table 4.1 or in (D.4) if you need them.

Note that the n_x, n_y, n_z numbering system for the solutions arose naturally from the solution process; it was not imposed a priori.

Key Points

- The eigenvalues and eigenfunctions have been found, skipping a lot of tedious math that you can check when the weather is bad during spring break.

o⊸ Generic expressions for the eigenvalues are above in (4.11) and for the eigenfunctions in (4.12).

4.1.2 Review Questions

1. Write out the ground state energy.
   Solution harmb-a
2. Write out the ground state wave function fully.
   Solution harmb-b
3. Write out the energy E_{100}.
   Solution harmb-c
4. Write out the eigenstate ψ_{100} fully.
   Solution harmb-d

4.1.3 Discussion of the eigenvalues

As the previous subsection showed, for every set of three nonnegative whole numbers n_x, n_y, n_z, there is one unique energy eigenfunction, or eigenstate, (4.12) and a corresponding energy eigenvalue (4.11). The "quantum numbers" n_x, n_y, and n_z correspond to the numbering system of the one-dimensional solutions that make up the full solution.

This section will examine the energy eigenvalues. These are of great physical importance, because according to the orthodox interpretation, they are the only measurable values of the total energy, the only energy levels that the oscillator can ever be found at.

The energy levels can be plotted in the form of a so-called "energy spectrum", as in figure 4.2. The energy values are listed along the vertical axis, and the sets of quantum numbers n_x, n_y, n_z for which they occur are shown to the right of the plot.

The first point of interest illustrated by the energy spectrum is that the energy of the oscillating particle cannot take on any arbitrary value, but only certain discrete values. Of course, that is just like for the particle in the pipe of the previous section, but for the harmonic oscillator, the energy levels are evenly spaced. In particular the energy value is always an odd multiple of $\frac{1}{2}\hbar\omega$. It contradicts the Newtonian notion that a harmonic oscillator can have any energy level. But since $\hbar$ is so small, about 10^{-34} kg m^2/s, macroscopically the different energy levels are extremely close together. Though the old Newtonian theory is strictly speaking incorrect, it remains an excellent approximation for macroscopic oscillators.

Also note that the energy levels have no largest value; however high the energy of the particle in a true harmonic oscillator may be, it will never escape. The further it tries to go, the larger the forces that pull it back. It can't win.

Another striking feature of the energy spectrum is that the lowest possible energy is again nonzero. The lowest energy occurs for $n_x = n_y = n_z = 0$ and

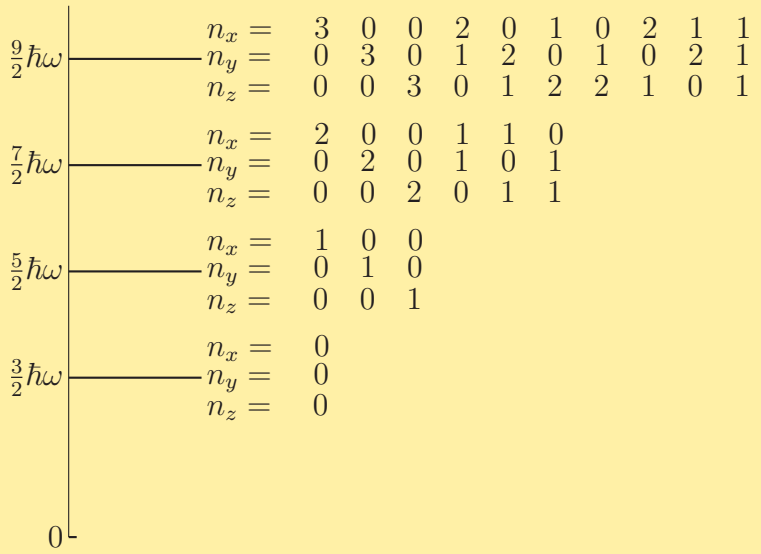


Figure 4.2: The energy spectrum of the harmonic oscillator.

has a value:

$$E_{000} = \tfrac{3}{2}\hbar\omega \tag{4.13}$$

So, even at absolute zero temperature, the particle is not completely at rest at its nominal position; it still has $\frac{3}{2}\hbar\omega$ worth of kinetic and potential energy left that it can never get rid of. This lowest energy state is the ground state.

The reason that the energy cannot be zero can be understood from the uncertainty principle. To get the potential energy to be zero, the particle would have to be at its nominal position for certain. But the uncertainty principle does not allow a precise position. Also, to get the kinetic energy to be zero, the linear momentum would have to be zero for certain, and the uncertainty principle does not allow that either.

The actual ground state is a compromise between uncertainties in momentum and position that make the total energy as small as Heisenberg's relationship allows. There is enough uncertainty in momentum to keep the particle near the nominal position, minimizing potential energy, but there is still enough uncertainty in position to keep the momentum low, minimizing kinetic energy. In fact, the compromise results in potential and kinetic energies that are exactly equal, {D.13}.

For energy levels above the ground state, figure 4.2 shows that there is a rapidly increasing number of different sets of quantum numbers n_x, n_y, and n_z that all produce that energy. Since each set represents one eigenstate, it means that multiple states produce the same energy.

Key Points

- Energy values can be graphically represented as an energy spectrum.

- The energy values of the harmonic oscillator are equally spaced, with a constant energy difference of $\hbar\omega$ between successive levels.
- The ground state of lowest energy has nonzero kinetic and potential energy.
- For any energy level above the ground state, there is more than one eigenstate that produces that energy.

4.1.3 Review Questions

1. Verify that the sets of quantum numbers shown in the spectrum figure 4.2 do indeed produce the indicated energy levels.
   Solution harmc-a
2. Verify that there are no sets of quantum numbers missing in the spectrum figure 4.2; the listed ones are the only ones that produce those energy levels.
   Solution harmc-b

4.1.4 Discussion of the eigenfunctions

This section takes a look at the energy eigenfunctions of the harmonic oscillator to see what can be said about the position of the particle at various energy levels.

At absolute zero temperature, the particle will be in the ground state of lowest energy. The eigenfunction describing this state has the lowest possible numbering $n_x = n_y = n_z = 0$, and is according to (4.12) of subsection 4.1.2 equal to

$$\psi_{000} = h_0(x)h_0(y)h_0(z) \tag{4.14}$$

where function h_0 is in table 4.1. The wave function in the ground state must be equal to the eigenfunction to within a constant:

$$\Psi_{\rm gs} = c_{000}h_0(x)h_0(y)h_0(z) \tag{4.15}$$

where the magnitude of the constant c_{000} must be one. Using the expression for function h_0 from table 4.1, the properties of the ground state can be explored.

As noted earlier in section 3.1, it is useful to plot the square magnitude of Ψ as grey tones, because the darker regions will be the ones where the particle is more likely to be found. Such a plot for the ground state is shown in figure 4.3. It shows that in the ground state, the particle is most likely to be found near the nominal position, and that the probability of finding the particle falls off quickly to zero beyond a certain distance from the nominal position.

The region in which the particle is likely to be found extends, roughly speaking, about a distance $\ell = \sqrt{\hbar/m\omega}$ from the nominal position. For a macroscopic oscillator, this will be a very small distance because of the smallness of $\hbar$. That

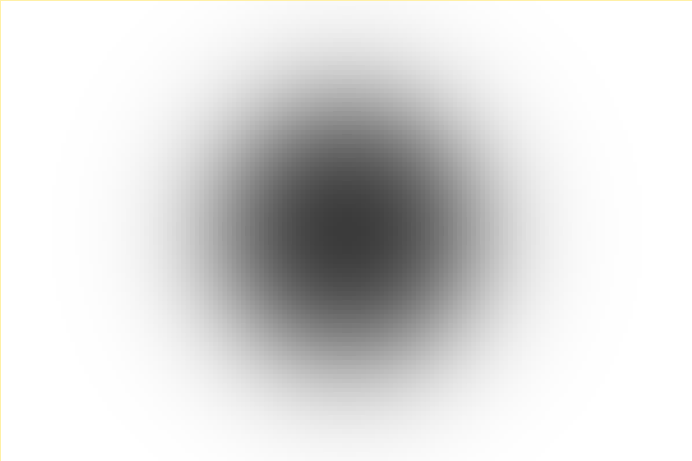


Figure 4.3: Ground state of the harmonic oscillator

is somewhat comforting, because macroscopically, you would expect an oscillator to be able to be at rest at the nominal position. While quantum mechanics does not allow it, at least the distance ℓ from the nominal position, and the energy $\frac{3}{2}\hbar\omega$ are extremely small.

But obviously, the bad news is that the ground state probability density of figure 4.3 does not at all resemble the classical Newtonian picture of a localized particle oscillating back and forwards. In fact, the probability density does not even depend on time: the chances of finding the particle in any given location are the same for all times. The probability density is also spherically symmetric; it only depends on the distance from the nominal position, and is the same at all angular orientations. To get something that can start to resemble a Newtonian spring-mass oscillator, one requirement is that the energy is well above the ground level.

Turning now to the second lowest energy level, this energy level is achieved by three different energy eigenfunctions, ψ_{100}, ψ_{010}, and ψ_{001}. The probability distribution of each of the three takes the form of two separate "blobs"; figure 4.4 shows ψ_{100} and ψ_{010} when seen along the z-direction. In case of ψ_{001}, one blob hides the other, so this eigenfunction was not shown.

Obviously, these states too do not resemble a Newtonian oscillator at all. The probability distributions once again stay the same at all times. (This is a consequence of energy conservation, as discussed later in chapter 7.1.) Also, while in each case there are two blobs occupied by a single particle, the particle will never be be caught on the symmetry plane in between the blobs, which naively could be taken as a sign of the particle moving from one blob to the other.

The eigenfunctions for still higher energy levels show similar lack of resemblance to the classical motion. As an arbitrary example, figure 4.5 shows eigen-

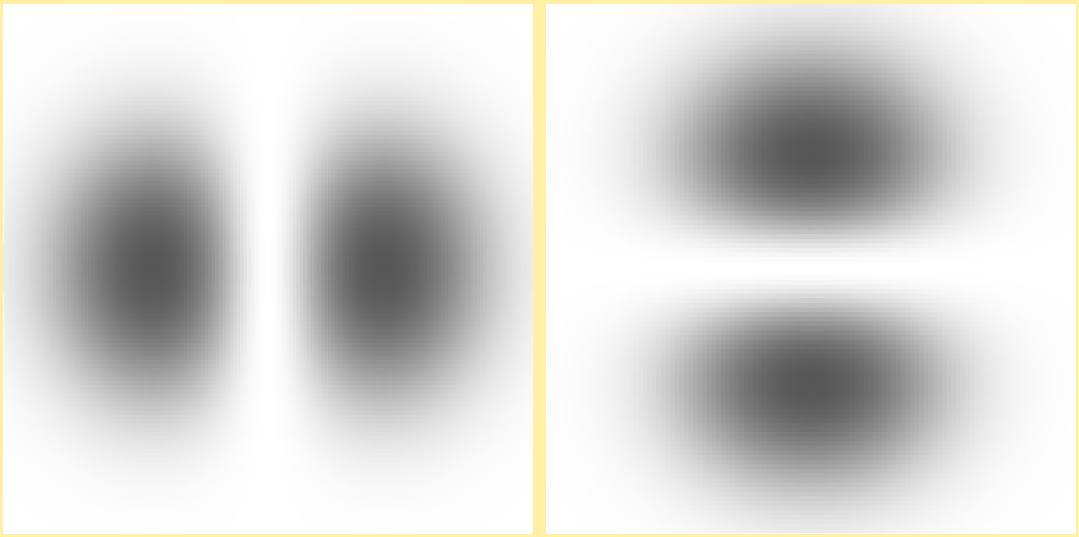


Figure 4.4: Wave functions ψ_{100} and ψ_{010}.

function ψ_{213} when looking along the z-axis. To resemble a classical oscillator, the particle would need to be restricted to, maybe not an exact moving point, but at most a very small moving region. Instead, all energy eigenfunctions have steady probability distributions and the locations where the particle may be found extend over large regions. It turns out that there is an uncertainty principle involved here: in order to get some localization of the position of the particle, you need to allow some uncertainty in its energy. This will have to wait until much later, in chapter 7.11.4.

The basic reason that quantum mechanics is so slow is simple. To analyze, say the x motion, classical physics says: "the *value* of the total energy E_x is

$$E_x = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{2}cx^2,$$

now go analyze the motion!". Quantum mechanics says: "the total energy *operator* H_x is

$$H_x = \tfrac{1}{2}m\left(\frac{\hbar}{\mathrm{i}m}\frac{\partial}{\partial x}\right)^2 + \tfrac{1}{2}c\widehat{x}^2,$$

now first figure out the possible energy *values* $E_{x0}, E_{x1}, \ldots$ before you can even start thinking about analyzing the motion."

Key Points

- The ground state wave function is spherically symmetric: it looks the same seen from any angle.
- In energy eigenstates the particle position is uncertain.

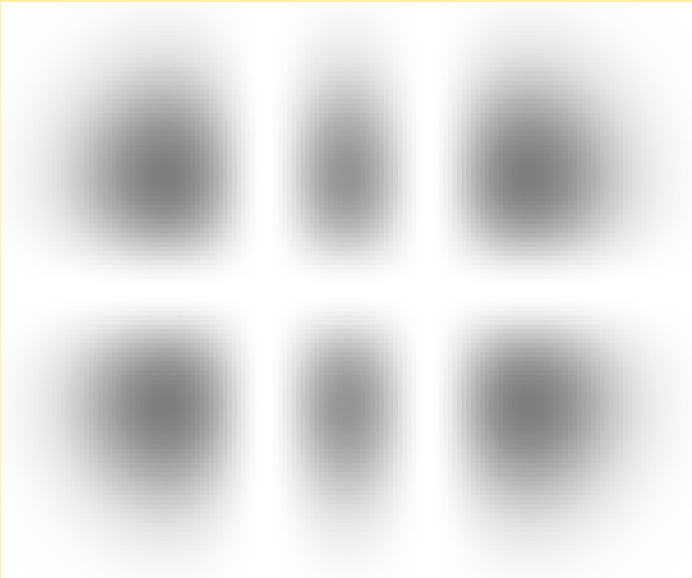


Figure 4.5: Energy eigenfunction ψ_{213}.

4.1.4 Review Questions

1. Write out the ground state wave function and show that it is indeed spherically symmetric.

   Solution harmd-a

2. Show that the ground state wave function is maximal at the origin and, like all the other energy eigenfunctions, becomes zero at large distances from the origin.

   Solution harmd-b

3. Write down the explicit expression for the eigenstate ψ_{213} using table 4.1, then verify that it looks like figure 4.5 when looking along the z-axis, with the x-axis horizontal and the y-axis vertical.

   Solution harmd-c

4.1.5 Degeneracy

As the energy spectrum figure 4.2 illustrated, the only energy level for which there is only a single energy eigenfunction is the ground state. All higher energy levels are what is called "degenerate"; there is more than one eigenfunction that produces that energy. (In other words, more than one set of three quantum numbers n_x, n_y, and n_z.)

It turns out that degeneracy always results in nonuniqueness of the eigenfunctions. That is important for a variety of reasons. For example, in the quantum mechanics of molecules, chemical bonds often select among nonunique theoretical solutions those that best fit the given conditions. Also, to find spe-

cific mathematical or numerical solutions for the eigenfunctions of a quantum system, the nonuniquenesses will somehow have to be resolved.

Nonuniqueness also poses problems for advanced analysis. For example, suppose you try to analyze the effect of various small perturbations that a harmonic oscillator might experience in real life. Analyzing the effect of small perturbations is typically a relatively easy mathematical problem: the perturbation will slightly change an eigenfunction, but it can still be approximated by the unperturbed one. So, if you know the unperturbed eigenfunction you are in business; unfortunately, if the unperturbed eigenfunction is not unique, you may not know which is the right one to use in the analysis.

The nonuniqueness arises from the fact that:

> *Linear combinations of eigenfunctions at the same energy level produce alternative eigenfunctions that still have that same energy level.*

For example, the eigenfunctions ψ_{100}, and ψ_{010} of the harmonic oscillator have the same energy $E_{100} = E_{010} = \frac{5}{2}\hbar\omega$ (as does ψ_{001}, but this example will be restricted to two eigenfunctions.) Any linear combination of the two has that energy too, so you could replace eigenfunctions ψ_{100} and ψ_{010} by two alternative ones such as:

$$\frac{\psi_{100} + \psi_{010}}{\sqrt{2}} \quad \text{and} \quad \frac{\psi_{010} - \psi_{100}}{\sqrt{2}}$$

It is readily verified these linear combinations are indeed still eigenfunctions with eigenvalue $E_{100} = E_{010}$: applying the Hamiltonian H to either one will multiply each term by $E_{100} = E_{010}$, hence the entire combination by that amount. How do these alternative eigenfunctions look? Exactly like ψ_{100} and ψ_{010} in figure 4.4, except that they are rotated over 45 degrees. Clearly then, they are just as good as the originals, just seen under a different angle.

Which raises the question, how come the analysis ended up with the ones that it did in the first place? The answer is in the method of separation of variables that was used in subsection 4.1.2. It produced eigenfunctions of the form $h_{n_x}(x)h_{n_y}(y)h_{n_z}(z)$ that were not just eigenfunctions of the full Hamiltonian H, but also of the partial Hamiltonians H_x, H_y, and H_z, being the x, y, and z parts of it.

For example, $\psi_{100} = h_1(x)h_0(y)h_0(z)$ is an eigenfunction of H_x with eigenvalue $E_{x1} = \frac{3}{2}\hbar\omega$, of H_y with eigenvalue $E_{y0} = \frac{1}{2}\hbar\omega$, and of H_z with eigenvalue $E_{z0} = \frac{1}{2}\hbar\omega$, as well as of H with eigenvalue $E_{100} = \frac{5}{2}\hbar\omega$.

The alternative eigenfunctions are still eigenfunctions of H, but no longer of the partial Hamiltonians. For example,

$$\frac{\psi_{100} + \psi_{010}}{\sqrt{2}} = \frac{h_1(x)h_0(y)h_0(z) + h_0(x)h_1(y)h_0(z)}{\sqrt{2}}$$

is not an eigenfunction of H_x: taking H_x times this eigenfunction would multiply the first term by E_{x1} but the second term by E_{x0}.

So, the obtained eigenfunctions were really made determinate by ensuring that they are simultaneously eigenfunctions of H, H_x, H_y, and H_z. The nice thing about them is that they can answer questions not just about the total energy of the oscillator, but also about how much of that energy is in each of the three directions.

Key Points

- Degeneracy occurs when different eigenfunctions produce the same energy.
- It causes nonuniqueness: alternative eigenfunctions will exist.
- That can make various analysis a lot more complex.

4.1.5 Review Questions

1. Just to check that this book is not lying, (you cannot be too careful), write down the analytical expression for ψ_{100} and ψ_{010} using table 4.1. Next write down $(\psi_{100} + \psi_{010})/\sqrt{2}$ and $(\psi_{010} - \psi_{100})/\sqrt{2}$. Verify that the latter two are the functions ψ_{100} and ψ_{010} in a coordinate system $(\bar{x}, \bar{y}, z)$ that is rotated 45 degrees counter-clockwise around the z-axis compared to the original (x, y, z) coordinate system.

   Solution harme-a

4.1.6 Noneigenstates

It should not be thought that the harmonic oscillator only exists in energy eigenstates. The opposite is more like it. Anything that somewhat localizes the particle will produce an uncertainty in energy. This section explores the procedures to deal with states that are not energy eigenstates.

First, even if the wave function is not an energy eigenfunction, it can still always be written as a combination of the eigenfunctions:

$$\Psi(x, y, z, t) = \sum_{n_x=0}^{\infty} \sum_{n_y=0}^{\infty} \sum_{n_z=0}^{\infty} c_{n_x n_y n_z} \psi_{n_x n_y n_z} \tag{4.16}$$

That this is always possible is a consequence of the completeness of the eigenfunctions of Hermitian operators such as the Hamiltonian. An arbitrary example of such a combination state is shown in figure 4.6.

The coefficients $c_{n_x n_y n_z}$ in the combination are important: according to the orthodox statistical interpretation, their square magnitude gives the probability to find the energy to be the corresponding eigenvalue $E_{n_x n_y n_z}$. For example, $|c_{000}|^2$ gives the probability of finding that the oscillator is in the ground state of lowest energy.

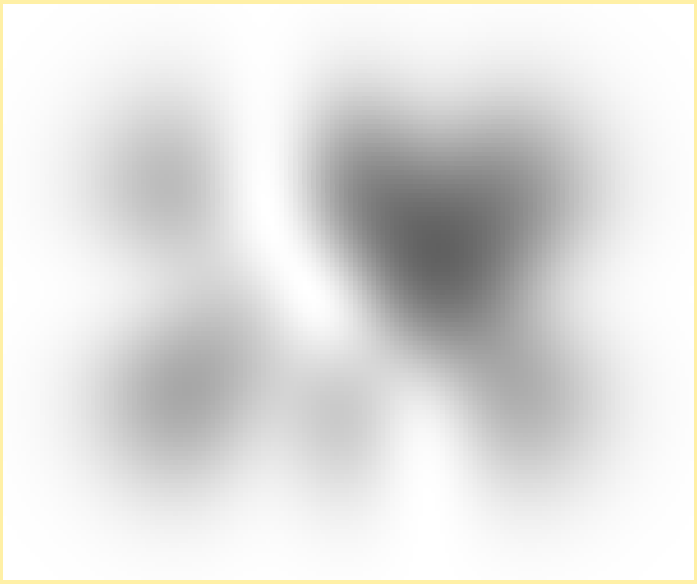


Figure 4.6: Arbitrary wave function (not an energy eigenfunction).

If the wave function Ψ is in a known state, (maybe because the position of the particle was fairly accurately measured), then each coefficient $c_{n_x n_y n_z}$ can be found by computing an inner product:

$$c_{n_x n_y n_z} = \langle \psi_{n_x n_y n_z} | \Psi \rangle \tag{4.17}$$

The reason this works is orthonormality of the eigenfunctions. As an example, consider the case of coefficient c_{100}:

$$c_{100} = \langle \psi_{100} | \Psi \rangle = \langle \psi_{100} | c_{000}\psi_{000} + c_{100}\psi_{100} + c_{010}\psi_{010} + c_{001}\psi_{001} + c_{200}\psi_{200} + \ldots \rangle$$

Now proper eigenfunctions of Hermitian operators are orthonormal; the inner product between different eigenfunctions is zero, and between identical eigenfunctions is one:

$$\langle \psi_{100} | \psi_{000} \rangle = 0 \quad \langle \psi_{100} | \psi_{100} \rangle = 1 \quad \langle \psi_{100} | \psi_{010} \rangle = 0 \quad \langle \psi_{100} | \psi_{001} \rangle = 0 \quad \ldots$$

So, the inner product above must indeed produce c_{100}.

Chapter 7.1 will discuss another reason why the coefficients are important: they determine the time evolution of the wave function. It may be recalled that the Hamiltonian, and hence the eigenfunctions derived from it, did not involve time. However, the coefficients do.

Even if the wave function is initially in a state involving many eigenfunctions, such as the one in figure 4.6, the orthodox interpretation says that energy "measurement" will collapse it into a single eigenfunction. For example, assume

that the energies in all three coordinate directions are measured and that they return the values:

$$E_{x2} = \tfrac{5}{2}\hbar\omega \quad E_{y1} = \tfrac{3}{2}\hbar\omega \quad E_{z3} = \tfrac{7}{2}\hbar\omega$$

for a total energy $E = \frac{15}{2}\hbar\omega$. Quantum mechanics could not exactly predict that this was going to happen, but it did predict that the energies had to be odd multiples of $\frac{1}{2}\hbar\omega$. Also, quantum mechanics gave the probability of measuring the given values to be whatever $|c_{213}|^2$ was. Or in other words, what $|\langle\psi_{213}|\Psi\rangle|^2$ was.

After the example measurement, the predictions become much more specific, because the wave function is now collapsed into the measured one:

$$\Psi^{\text{new}} = c_{213}^{\text{new}}\psi_{213}$$

This eigenfunction was shown earlier in figure 4.5.

If another measurement of the energies is now done, the only values that can come out are E_{x2}, E_{y1}, and E_{z3}, the same as in the first measurement. There is now certainty of getting those values; the probability $|c_{213}^{\text{new}}|^2 = 1$. This will continue to be true for energy measurements until the system is disturbed, maybe by a position measurement.

Key Points

- The basic ideas of quantum mechanics were illustrated using an example.
- The energy eigenfunctions are not the only game in town. Their seemingly lowly coefficients are important too.
- When the wave function is known, the coefficient of any eigenfunction can be found by taking an inner product of the wave function with that eigenfunction.

4.2 Angular Momentum

Before a solution can be found for the important electronic structure of the hydrogen atom, the basis for the description of all the other elements and chemical bonds, first angular momentum must be discussed. Like in the classical Newtonian case, angular momentum is essential for the analysis, and in quantum mechanics, angular momentum is also essential for describing the final solution. Moreover, the quantum properties of angular momentum turn out to be quite unexpected and important for practical applications.

4.2.1 Definition of angular momentum

The old Newtonian physics defines *angular* momentum $\vec{L}$ as the vectorial product $\vec{r} \times \vec{p}$, where $\vec{r}$ is the position of the particle in question and $\vec{p}$ is its *linear* momentum.

Following the Newtonian analogy, quantum mechanics substitutes the gradient operator $\hbar\nabla/\mathrm{i}$ for the linear momentum, so the angular momentum operator becomes:

$$\widehat{\vec{L}} = \frac{\hbar}{\mathrm{i}} \widehat{\vec{r}} \times \nabla \qquad \widehat{\vec{r}} \equiv (\widehat{x}, \widehat{y}, \widehat{z}) \quad \nabla \equiv \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \tag{4.18}$$

Unlike the Hamiltonian, the angular momentum operator is not specific to a given system. All observations about angular momentum will apply regardless of the physical system being studied.

Key Points

- The angular momentum operator (4.18) has been identified.

4.2.2 Angular momentum in an arbitrary direction

The intent in this subsection is to find the operator for the angular momentum in an arbitrary direction and its eigenfunctions and eigenvalues.

For convenience, the direction in which the angular momentum is desired will be taken as the z-axis of the coordinate system. In fact, much of the mathematics that you do in quantum mechanics requires you to select some arbitrary direction as your z-axis, even if the physics itself does not have any preferred direction. It is further conventional in the quantum mechanics of atoms and molecules to draw the chosen z-axis horizontal, (though not in [25] or [52]), and that is what will be done here.

Things further simplify greatly if you switch from Cartesian coordinates x, y, and z to "spherical coordinates" r, θ, and ϕ, as shown in figure 4.7. The coordinate r is the distance from the chosen origin, θ is the angular position away from the chosen z-axis, and ϕ is the angular position around the z-axis, measured from the chosen x-axis.

In terms of these spherical coordinates, the z-component of angular momentum simplifies to:

$$\boxed{\widehat{L}_z \equiv \frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial \phi}} \tag{4.19}$$

This can be verified by looking up the gradient operator ∇ in spherical coordinates in [41, pp. 124-126] and then taking the component of $\vec{r} \times \nabla$ in the z-direction.

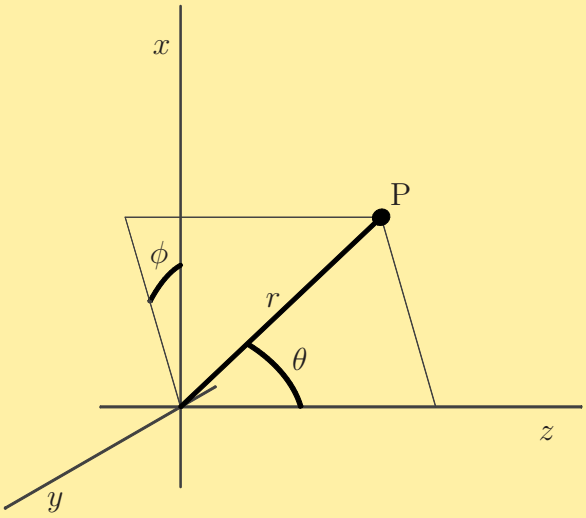


Figure 4.7: Spherical coordinates of an arbitrary point P.

In any case, with a bit of thought, it clearly makes sense: the z-component of linear momentum classically describes the motion in the *direction* of the z-axis, while the z-component of angular momentum describes the motion *around* the z-axis. So if in quantum mechanics the z linear momentum is $\hbar/\mathrm{i}$ times the derivative with respect the coordinate z along the z-axis, then surely the logical equivalent for z angular momentum is $\hbar/\mathrm{i}$ times the derivative with respect to the angle ϕ around the z-axis?

Anyway, the eigenfunctions of the operator $\widehat{L}_z$ above turn out to be exponentials in ϕ. More precisely, the eigenfunctions are of the form

$$C(r,\theta)e^{\mathrm{i}m\phi} \tag{4.20}$$

Here m is a constant and $C(r,\theta)$ can be any arbitrary function of r and θ. For historical reasons, the number m is called the "magnetic quantum number". Historically, physicists have never seen a need to get rid of obsolete and confusing terms. The magnetic quantum number must be an integer, one of $\ldots,-2,-1,0,1,2,3,\ldots$ The reason is that if you increase the angle ϕ by 2π, you make a complete circle around the z-axis and return to the same point. Then the eigenfunction (4.20) must again be the same, but that is only the case if m is an integer. To verify this, use the Euler formula (2.5).

Note further that the orbital momentum is associated with a particle whose mass is *also* indicated by m. This book will more specifically indicate the magnetic quantum number as m_l if confusion between the two is likely.

The above solution is easily verified directly, and the eigenvalue L_z identified, by substitution into the eigenvalue problem $\widehat{L}_z Ce^{\mathrm{i}m\phi} = L_z Ce^{\mathrm{i}m\phi}$ using the expression for $\widehat{L}_z$ above:

$$\frac{\hbar}{\mathrm{i}}\frac{\partial Ce^{\mathrm{i}m\phi}}{\partial\phi} = L_z Ce^{\mathrm{i}m\phi} \quad \Longrightarrow \quad \frac{\hbar}{\mathrm{i}}\mathrm{i}mCe^{\mathrm{i}m\phi} = L_z Ce^{\mathrm{i}m\phi}$$

It follows that every eigenvalue is of the form:

$$\boxed{L_z = m_l \hbar \text{ for } m_l \text{ an integer}} \tag{4.21}$$

So the angular momentum in a given direction cannot just take on any value: it must be a whole multiple m_l, (possibly negative), of Planck's constant $\hbar$.

Compare that with the linear momentum component p_z which can take on any value, within the accuracy that the uncertainty principle allows. L_z can only take discrete values, but they will be precise. And since the z-axis was arbitrary, this is true in any direction you choose.

It is important to keep in mind that if the surroundings of the particle has no preferred direction, the angular momentum in the arbitrarily chosen z-direction is physically irrelevant. For example, for the motion of the electron in an isolated hydrogen atom, no preferred direction of space can be identified. Therefore, the energy of the electron will only depend on its total angular momentum, not on the angular momentum in whatever is completely arbitrarily chosen to be the z-direction. In terms of quantum mechanics, that means that the value of m does not affect the energy. (Actually, this is not exactly true, although it is true to very high accuracy. The electron and nucleus have magnetic fields that give them inherent directionality. It remains true that the z-component of net angular momentum of the complete atom is not relevant. However, the space in which the electron moves has a preferred direction due to the magnetic field of the nucleus and vice-versa. It affects energy very slightly. Therefore the electron and nucleus must coordinate their angular momentum components, addendum {A.39}.)

Key Points

- Even if the physics that you want to describe has no preferred direction, you usually need to select some arbitrary z-axis to do the mathematics of quantum mechanics.
- Spherical coordinates based on the chosen z-axis are needed in this and subsequent analysis. They are defined in figure 4.7.
- The operator for the z-component of angular momentum is (4.19), where ϕ is the angle around the z-axis.
- The eigenvalues, or measurable values, of angular momentum in any arbitrary direction are whole multiples m, possibly negative, of $\hbar$.
- The whole multiple m is called the magnetic quantum number.

4.2.2 Review Questions

1. If the angular momentum in a given direction is a multiple of $\hbar = 1.054\,57\ 10^{-34}$ J s, then $\hbar$ should have units of angular momentum. Verify that.

   Solution angub-a

2. What is the magnetic quantum number of a macroscopic, 1 kg, particle that is encircling the z-axis at a distance of 1 m at a speed of 1 m/s? Write out as an integer, and show digits you are not sure about as a question mark.

   Solution angub-b

3. Actually, based on the derived eigenfunction, $C(r,\theta)e^{\mathrm{i}m\phi}$, would any macroscopic particle ever be at a single magnetic quantum number in the first place? In particular, what can you say about where the particle can be found in an eigenstate?

   Solution angub-c

4.2.3 Square angular momentum

Besides the angular momentum in an arbitrary direction, the other quantity of primary importance is the magnitude of the angular momentum. This is the length of the angular momentum vector, $\sqrt{\vec{L}\cdot\vec{L}}$. The square root is awkward, though; it is easier to work with the square angular momentum:

$$L^2 \equiv \vec{L}\cdot\vec{L}$$

This subsection discusses the $\widehat{L}^2$ operator and its eigenvalues.

Like the $\widehat{L}_z$ operator of the previous subsection, $\widehat{L}^2$ can be written in terms of spherical coordinates. To do so, note first that

$$\widehat{\vec{L}}\cdot\widehat{\vec{L}} = \frac{\hbar}{\mathrm{i}}(\vec{r}\times\nabla)\cdot\frac{\hbar}{\mathrm{i}}(\vec{r}\times\nabla) = -\hbar^2\vec{r}\cdot(\nabla\times(\vec{r}\times\nabla))$$

(That is the basic vector identity (D.2) with vectors $\vec{r}$, ∇, and $\vec{r}\times\nabla$.) Next look up the gradient and the curl in [41, pp. 124-126]. The result is:

$$\widehat{L}^2 \equiv -\frac{\hbar^2}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) - \frac{\hbar^2}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2} \tag{4.22}$$

Obviously, this result is not as intuitive as the $\widehat{L}_z$ operator of the previous subsection, but once again, it only involves the spherical coordinate angles. The measurable values of square angular momentum will be the eigenvalues of this operator. However, that eigenvalue problem is not easy to solve. In fact the solution is not even unique.

The solution to the problem may be summarized as follows. First, the nonuniqueness is removed by demanding that the eigenfunctions are *also* eigenfunctions of $\widehat{L}_z$, the operator of angular momentum in the z-direction. This makes the problem solvable, {D.14}, and the resulting eigenfunctions are called the "spherical harmonics" $Y_l^m(\theta,\phi)$. The first few are given explicitly in table 4.2. In case you need more of them for some reason, there is a generic expression (D.5) in derivation {D.14}.

$Y_0^0 = \sqrt{\frac{1}{4\pi}}$	$Y_1^0 = \sqrt{\frac{3}{4\pi}}\cos(\theta)$	$Y_2^0 = \sqrt{\frac{5}{16\pi}}(3\cos^2\theta - 1)$
	$Y_1^1 = -\sqrt{\frac{3}{8\pi}}\sin\theta\, e^{\mathrm{i}\phi}$	$Y_2^1 = -\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta\, e^{\mathrm{i}\phi}$
	$Y_1^{-1} = \sqrt{\frac{3}{8\pi}}\sin\theta\, e^{-\mathrm{i}\phi}$	$Y_2^{-1} = \sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta\, e^{-\mathrm{i}\phi}$
		$Y_2^2 = \sqrt{\frac{15}{32\pi}}\sin^2\theta\, e^{2\mathrm{i}\phi}$
		$Y_2^{-2} = \sqrt{\frac{15}{32\pi}}\sin^2\theta\, e^{-2\mathrm{i}\phi}$

Table 4.2: The first few spherical harmonics.

These eigenfunctions can additionally be multiplied by any arbitrary function of the distance from the origin r. They are normalized to be orthonormal integrated over the surface of the unit sphere:

$$\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} Y_l^m(\theta,\phi)^* Y_{\underline{l}}^{\underline{m}}(\theta,\phi)\sin\theta\,\mathrm{d}\theta\mathrm{d}\phi = \begin{cases} 1 & \text{if } l = \underline{l} \text{ and } m = \underline{m} \\ 0 & \text{otherwise} \end{cases} \tag{4.23}$$

The spherical harmonics Y_l^m are sometimes symbolically written in "ket notation" as $|l\, m\rangle$.

What to say about them, except that they are in general a mess? Well, at least every one is proportional to $e^{\mathrm{i}m\phi}$, as an eigenfunction of $\widehat{L}_z$ should be. More importantly, the very first one, Y_0^0 is independent of angular position compared to the origin (it is the same for all θ and ϕ angular positions.) This eigenfunction corresponds to the state in which there is no angular momentum around the origin at all. If a particle has no angular momentum around the origin, it can be found at all angular locations relative to it with equal probability.

There is a different way of looking at the angular momentum eigenfunctions. It is shown in table 4.3. It shows that $r^l Y_l^m$ is always a polynomial in the position component of degree l. Furthermore, you can check that $\nabla^2 r^l Y_l^m = 0$: the Laplacian of $r^l Y_l^m$ is always zero. This way of looking at the spherical harmonics is often very helpful in understanding more advanced quantum topics. These solutions may be indicated as

$$\mathcal{Y}_l^m \equiv r^l Y_l^m \tag{4.24}$$

$Y_0^0 = \sqrt{\frac{1}{4\pi}}$	$rY_1^0 = \sqrt{\frac{3}{4\pi}}z$	$r^2Y_2^0 = \sqrt{\frac{5}{16\pi}}(2z^2 - x^2 - y^2)$
	$rY_1^1 = -\sqrt{\frac{3}{8\pi}}(x + \mathrm{i}y)$	$r^2Y_2^1 = -\sqrt{\frac{15}{8\pi}}z(x + \mathrm{i}y)$
	$rY_1^{-1} = \sqrt{\frac{3}{8\pi}}(x - \mathrm{i}y)$	$r^2Y_2^{-1} = \sqrt{\frac{15}{8\pi}}z(x - \mathrm{i}y)$
		$r^2Y_2^2 = \sqrt{\frac{15}{32\pi}}(x + \mathrm{i}y)^2$
		$r^2Y_2^{-2} = \sqrt{\frac{15}{32\pi}}(x - \mathrm{i}y)^2$

Table 4.3: The first few spherical harmonics rewritten.

and referred to as the "harmonic polynomials." In general the term "harmonic" indicates a function whose Laplacian ∇^2 is zero.

Far more important than the details of the eigenfunctions themselves are the eigenvalues that come rolling out of the analysis. A spherical harmonic Y_l^m has an angular momentum in the z-direction

$$L_z = m\hbar \tag{4.25}$$

where the integer m is called the magnetic quantum number, as noted in the previous subsection. That is no surprise, because the analysis demanded that they take that form. The new result is that a spherical harmonic has a square angular momentum

$$\boxed{L^2 = l(l+1)\hbar^2} \tag{4.26}$$

where l is also an integer, and is called the "azimuthal quantum number" for reasons you do not want to know. It is maybe a weird result, (why not simply $l^2\hbar^2$?) but that is what square angular momentum turns out to be.

The azimuthal quantum number is at least as large as the magnitude of the magnetic quantum number m:

$$\boxed{l \geqslant |m|} \tag{4.27}$$

The reason is that $\widehat{L}^2 = \widehat{L}_x^2 + \widehat{L}_y^2 + \widehat{L}_z^2$ must be at least as large as $\widehat{L}_z^2$; in terms of eigenvalues, $l(l+1)\hbar^2$ must be at least as large as $m^2\hbar^2$. As it is, with $l \geqslant |m|$, either the angular momentum is completely zero, for $l = m = 0$, or L^2 is always greater than L_z^2.

Key Points

- The operator for square angular momentum is (4.22).
- The eigenfunctions of both square angular momentum and angular momentum in the chosen z-direction are called the spherical harmonics Y_l^m.
- If a particle has no angular momentum around the origin, it can be found at all angular locations relative to it with equal probability.
- The eigenvalues for square angular momentum take the counter-intuitive form $L^2 = l(l+1)\hbar^2$ where l is a nonnegative integer, one of 0, 1, 2, 3, ..., and is called the azimuthal quantum number.
- The azimuthal quantum number l is always at least as big as the absolute value of the magnetic quantum number m.

4.2.3 Review Questions

1. The general wave function of a state with azimuthal quantum number l and magnetic quantum number m is $\Psi = R(r)Y_l^m(\theta,\phi)$, where $R(r)$ is some further arbitrary function of r. Show that the condition for this wave function to be normalized, so that the total probability of finding the particle integrated over all possible positions is one, is that

$$\int_{r=0}^{\infty} R(r)^* R(r) r^2 \,\mathrm{d}r = 1.$$

   Solution anguc-a

2. Can you invert the statement about zero angular momentum and say: if a particle can be found at all angular positions compared to the origin with equal probability, it will have zero angular momentum?
   Solution anguc-b

3. What is the minimum amount that the total square angular momentum is larger than just the square angular momentum in the z-direction for a given value of l?
   Solution anguc-c

4.2.4 Angular momentum uncertainty

Rephrasing the final results of the previous subsection, if there is nonzero angular momentum, the angular momentum in the z-direction is always less than the total angular momentum. There is something funny going on here. The z-direction can be chosen arbitrarily, and if you choose it in the same direction as the angular momentum vector, then the z-component should be the entire vector. So, how can it always be less?

The answer of quantum mechanics is that the looked-for angular momentum vector *does not exist.* No axis, however arbitrarily chosen, can align with a nonexisting vector.

There is an uncertainty principle here, similar to the one of Heisenberg for position and linear momentum. For angular momentum, it turns out that if the component of angular momentum in a given direction, here taken to be z, has a definite value, then the components in both the x and y directions will be uncertain. (Details will be given in chapter 12.2). The wave function will be in a state where L_x and L_y have a range of possible values $m_1\hbar$, $m_2\hbar$, ..., each with some probability. Without definite x and y components, there simply is no angular momentum vector.

It is tempting to think of quantities that have not been measured, such as the angular momentum vector in this example, as being merely "hidden." However, the impossibility for the z-axis to ever align with any angular momentum vector shows that there is a fundamental difference between "being hidden" and "not existing".

Key Points

- According to quantum mechanics, an exact nonzero angular momentum vector will never exist. If one component of angular momentum has a definite value, then the other two components will be uncertain.

4.3 The Hydrogen Atom

This section examines the critically important case of the hydrogen atom. The hydrogen atom consists of a nucleus which is just a single proton, and an electron encircling that nucleus. The nucleus, being much heavier than the electron, can be assumed to be at rest, and only the motion of the electron is of concern.

The energy levels of the electron determine the photons that the atom will absorb or emit, allowing the powerful scientific tool of spectral analysis. The electronic structure is also essential for understanding the properties of the other elements and of chemical bonds.

4.3.1 The Hamiltonian

The first step is to find the Hamiltonian of the electron. The electron experiences an electrostatic Coulomb attraction to the oppositely charged nucleus. The corresponding potential energy is

$$V = -\frac{e^2}{4\pi\epsilon_0 r} \tag{4.28}$$

with r the distance from the nucleus. The constant

$$e \approx 1.602\,176\,6\ 10^{-19}\ \mathrm{C} \tag{4.29}$$

is the magnitude of the electric charges of the electron and proton, and the constant

$$\epsilon_0 \approx 8.854\,187\,817\ 10^{-12}\ \mathrm{C^2/J\ m} \tag{4.30}$$

is called the "permittivity of space."

Unlike for the harmonic oscillator discussed earlier, this potential energy cannot be split into separate parts for Cartesian coordinates x, y, and z. To do the analysis for the hydrogen atom, you must put the nucleus at the origin of the coordinate system and use spherical coordinates r (the distance from the nucleus), θ (the angle from an arbitrarily chosen z-axis), and ϕ (the angle around the z-axis); see figure 4.7. In terms of spherical coordinates, the potential energy above depends on just the single coordinate r.

To get the Hamiltonian, you need to add to this potential energy the kinetic energy operator $\widehat{T}$. Chapter 3.3 gave this operator as

$$\widehat{T} = -\frac{\hbar^2}{2m}\nabla^2$$

where ∇^2 is the Laplacian. The Laplacian in spherical coordinates is given in the notations, (N.5). Then the Hamiltonian is found to be:

$$H = -\frac{\hbar^2}{2m_e r^2}\left\{\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right\} - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r} \tag{4.31}$$

where

$$m_e \approx 9.109\ 10^{-31}\ \mathrm{kg} \tag{4.32}$$

is the mass of the electron.

It may be noted that the small proton motion can be corrected for by slightly adjusting the mass of the electron to be an effective $9.104\,4\ 10^{-31}$ kg, {A.5}. This makes the solution exact, except for extremely small errors due to relativistic effects. (These are discussed in addendum {A.39}.)

Key Points

- To analyze the hydrogen atom, you must use spherical coordinates.
- The Hamiltonian in spherical coordinates has been written down. It is (4.31).

4.3.2 Solution using separation of variables

This subsection describes in general lines how the eigenvalue problem for the electron of the hydrogen atom is solved. The basic ideas are like those used to solve the particle in a pipe and the harmonic oscillator, but in this case, they are used in spherical coordinates rather than Cartesian ones. Without getting too much caught up in the mathematical details, do not miss the opportunity of learning where the hydrogen energy eigenfunctions and eigenvalues come from. This is the crown jewel of quantum mechanics; brilliant, almost flawless, critically important; one of the greatest works of physical analysis ever.

The eigenvalue problem for the Hamiltonian, as formulated in the previous subsection, can be solved by searching for solutions ψ that take the form of a product of functions of each of the three coordinates: $\psi = R(r)\Theta(\theta)\Phi(\phi)$. More concisely, $\psi = R\Theta\Phi$. The problem now is to find separate equations for the individual functions R, Θ, and Φ from which they can then be identified. The arguments are similar as for the harmonic oscillator, but messier, since the coordinates are more entangled. First, substituting $\psi = R\Theta\Phi$ into the Hamiltonian eigenvalue problem $H\psi = E\psi$, with the Hamiltonian H as given in the previous subsection and E the energy eigenvalue, produces:

$$\left[-\frac{\hbar^2}{2m_{\rm e}r^2}\left\{\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right\}-\frac{e^2}{4\pi\epsilon_0}\frac{1}{r}\right]R\Theta\Phi = ER\Theta\Phi$$

To reduce this problem, premultiply by $2m_{\rm e}r^2/R\Theta\Phi$ and then separate the various terms:

$$-\frac{\hbar^2}{R}\frac{\partial}{\partial r}\left(r^2\frac{\partial R}{\partial r}\right)+\frac{1}{\Theta\Phi}\left\{-\frac{\hbar^2}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)-\frac{\hbar^2}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right\}\Theta\Phi - \frac{2m_{\rm e}r^2e^2}{4\pi\epsilon_0}\frac{1}{r} = 2m_{\rm e}r^2E \tag{4.33}$$

Next identify the terms involving the angular derivatives and name them $E_{\theta\phi}$. They are:

$$\frac{1}{\Theta\Phi}\left[-\frac{\hbar^2}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)-\frac{\hbar^2}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]\Theta\Phi = E_{\theta\phi}$$

By this definition, $E_{\theta\phi}$ only depends on θ and ϕ, not r. But it cannot depend on θ or ϕ either, since none of the other terms in the original equation (4.33) depends on them. So $E_{\theta\phi}$ must be a constant, independent of all three coordinates. Then multiplying the angular terms above by $\Theta\Phi$ produces a reduced eigenvalue problem involving $\Theta\Phi$ only, with eigenvalue $E_{\theta\phi}$:

$$\left[-\frac{\hbar^2}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)-\frac{\hbar^2}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]\Theta\Phi = E_{\theta\phi}\Theta\Phi \tag{4.34}$$

Repeat the game with this reduced eigenvalue problem. Multiply by $\sin^2\theta/\Theta\Phi$, and name the only ϕ-dependent term E_ϕ. It is:

$$-\frac{1}{\Phi}\hbar^2\left(\frac{\partial^2}{\partial\phi^2}\right)\Phi = E_\phi$$

By definition E_ϕ only depends on ϕ, but since the other two terms in the equation it came from did not depend on ϕ, E_ϕ cannot either, so it must be another constant. What is left is a simple eigenvalue problem just involving Φ:

$$-\hbar^2\left(\frac{\partial^2}{\partial\phi^2}\right)\Phi = E_\phi\Phi$$

And that is readily solvable.

In fact, the solution to this final problem has already been given, since the operator involved is just the square of the angular momentum operator $\widehat{L}_z$ of section 4.2.2:

$$-\hbar^2\left(\frac{\partial^2}{\partial\phi^2}\right)\Phi = \left(\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial\phi}\right)^2\Phi = \widehat{L}_z^2\Phi$$

So this equation must have the same eigenfunctions as the operator $\widehat{L}_z$,

$$\Phi_m = e^{\mathrm{i}m\phi}$$

and must have the square eigenvalues

$$E_\phi = (m\hbar)^2$$

(each application of $\widehat{L}_z$ multiplies the eigenfunction by $m\hbar$). It may be recalled that the magnetic quantum number m must be an integer.

The eigenvalue problem (4.34) for $\Theta\Phi$ is even easier; it is exactly the one for the square angular momentum L^2 of section 4.2.3. (So, no, there was not really a need to solve for Φ separately.) Its eigenfunctions are therefore the spherical harmonics,

$$\Theta\Phi = Y_l^m(\theta,\phi)$$

and its eigenvalues are

$$E_{\theta\phi} = l(l+1)\hbar^2$$

It may be recalled that the azimuthal quantum number l must be an integer greater than or equal to $|m|$.

Returning now to the solution of the original eigenvalue problem (4.33), replacement of the angular terms by $E_{\theta\phi} = l(l+1)\hbar^2$ turns it into an ordinary differential equation problem for the radial factor $R(r)$ in the energy eigenfunction. As usual, this problem is a pain to solve, so that is again shoved away in a note, {D.15}.

It turns out that the solutions of the radial problem can be numbered using a third quantum number, n, called the "principal quantum number". It is larger than the azimuthal quantum number l, which in turn must be at least as large as the absolute value of the magnetic quantum number:

$$\boxed{n > l \geqslant |m|} \tag{4.35}$$

so the principal quantum number must be at least 1. And if $n = 1$, then $l = m = 0$.

In terms of these three quantum numbers, the final energy eigenfunctions of the hydrogen atom are of the general form:

$$\boxed{\psi_{nlm} = R_{nl}(r) Y_l^m(\theta, \phi)} \tag{4.36}$$

where the spherical harmonics Y_l^m were described in section 4.2.3. The brand new radial wave functions R_{nl} can be found written out in table 4.4 for small values of n and l, or in derivation {D.15}, (D.8), for any n and l. They are usually written in terms of a scaled radial distance from the nucleus $\rho = r/a_0$, where the length a_0 is called the "Bohr radius" and has the value

$$\boxed{a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} \approx 0.529\,177\ 10^{-10}\ \text{m}} \tag{4.37}$$

or about half an Ångstrom. The Bohr radius is a really good length scale to describe atoms in terms of. The Ångstrom itself is a good choice too, it is 10^{-10} m, or one tenth of a nanometer.

$R_{10} = \dfrac{2}{\sqrt{a_0^3}}\, e^{-\rho}$	$R_{20} = \dfrac{2-\rho}{2\sqrt{2a_0^3}}\, e^{-\rho/2}$	$R_{30} = \dfrac{54 - 36\rho + 4\rho^2}{81\sqrt{3a_0^3}}\, e^{-\rho/3}$
	$R_{21} = \dfrac{\rho}{2\sqrt{6a_0^3}}\, e^{-\rho/2}$	$R_{31} = \dfrac{24\rho - 4\rho^2}{81\sqrt{6a_0^3}}\, e^{-\rho/3}$
		$R_{32} = \dfrac{4\rho^2}{81\sqrt{30a_0^3}}\, e^{-\rho/3}$
$a_0 = \dfrac{4\pi\epsilon_0\hbar^2}{m_e e^2}$		$\rho = \dfrac{r}{a_0}$

Table 4.4: The first few radial wave functions for hydrogen.

The energy eigenvalues are much simpler and more interesting than the eigenfunctions; they are

$$\boxed{E_n = -\frac{\hbar^2}{2m_e a_0^2}\frac{1}{n^2} = \frac{E_1}{n^2} \quad n = 1, 2, 3, \ldots \qquad E_1 = -\frac{\hbar^2}{2m_e a_0^2} = -13.605\,7 \text{ eV}} \tag{4.38}$$

where eV stands for electron volt, a unit of energy equal to $1.602\,18\ 10^{-19}$ J. It is the energy that an electron picks up during a 1 volt change in electric potential.

You may wonder why the energy only depends on the principal quantum number n, and not also on the azimuthal quantum number l and the magnetic quantum number m. Well, the choice of z-axis was arbitrary, so it should not seem that strange that the physics would not depend on the angular momentum in that direction. But that the energy does not depend on l is nontrivial: if you solve the simpler problem of a particle stuck inside an impenetrable spherical container, using procedures from {A.6}, the energy values depend on both n and l. So, that is just the way it is. (It stops being true anyway if you include relativistic effects in the Hamiltonian.)

Since the lowest possible value of the principal quantum number n is one, the ground state of lowest energy E_1 is eigenfunction ψ_{100}.

Key Points

- Skipping a lot of math, energy eigenfunctions ψ_{nlm} and their energy eigenvalues E_n have been found.
- There is one eigenfunction for each set of three integer quantum numbers n, l, and m satisfying $n > l \geqslant |m|$. The number n is called the principal quantum number.
- The typical length scale in the solution is called the Bohr radius a_0, which is about half an Ångstrom.
- The derived eigenfunctions ψ_{nlm} are eigenfunctions of
  - z angular momentum, with eigenvalue $L_z = m\hbar$;
  - square angular momentum, with eigenvalue $L^2 = l(l+1)\hbar^2$;
  - energy, with eigenvalue $E_n = -\hbar^2/2m_e a_0^2 n^2$.
- The energy values only depend on the principal quantum number n.
- The ground state is ψ_{100}.

4.3.2 Review Questions

1. Use the tables for the radial wave functions and the spherical harmonics to write down the wave function

$$\psi_{nlm} = R_{nl}(r)Y_l^m(\theta, \phi)$$

for the case of the ground state ψ_{100}.

Check that the state is normalized. Note: $\int_0^\infty e^{-2u} u^2 \, \mathrm{d}u = \frac{1}{4}$.

Solution hydb-a

2. Use the generic expression

$$\psi_{nlm} = -\frac{2}{n^2}\sqrt{\frac{(n-l-1)!}{[(n+l)!a_0]^3}}\left(\frac{2\rho}{n}\right)^l L_{n+l}^{2l+1}\left(\frac{2\rho}{n}\right) e^{-\rho/n} Y_l^m(\theta,\phi)$$

with $\rho = r/a_0$ and Y_l^m from the spherical harmonics table to find the ground state wave function ψ_{100}. Note: the Laguerre polynomial $L_1(x) = 1 - x$ and for any p, L_1^p is just its p-th derivative.

Solution hydb-b

3. Plug numbers into the generic expression for the energy eigenvalues,

$$E_n = -\frac{\hbar^2}{2m_e a_0^2}\frac{1}{n^2},$$

where $a_0 = 4\pi\epsilon_0\hbar^2/m_e e^2$, to find the ground state energy. Express in eV, where 1 eV equals $1.602\,2\ 10^{-19}$ J. Values for the physical constants can be found at the start of this section and in the notations section.

Solution hydb-c

4.3.3 Discussion of the eigenvalues

The only energy values that the electron in the hydrogen atom can have are the "Bohr energies" derived in the previous subsection:

$$E_n = -\frac{\hbar^2}{2m_e a_0^2}\frac{1}{n^2} \qquad n = 1, 2, 3, \ldots$$

This subsection discusses the physical consequences of this result.

To aid the discussion, the allowed energies are plotted in the form of an energy spectrum in figure 4.8. To the right of the lowest three energy levels the values of the quantum numbers that give rise to those energy levels are listed.

The first thing that the energy spectrum illustrates is that the energy levels are all negative, unlike the ones of the harmonic oscillator, which were all positive. However, that does not mean much; it results from defining the potential energy of the harmonic oscillator to be zero at the nominal position of the particle, while the hydrogen potential is instead defined to be zero at large distance from the nucleus. (It will be shown later, chapter 7.2, that the average potential energy is twice the value of the total energy, and the average kinetic energy is minus the total energy, making the average kinetic energy positive as it should be.)

A more profound difference is that the energy levels of the hydrogen atom have a maximum value, namely zero, while those of the harmonic oscillator

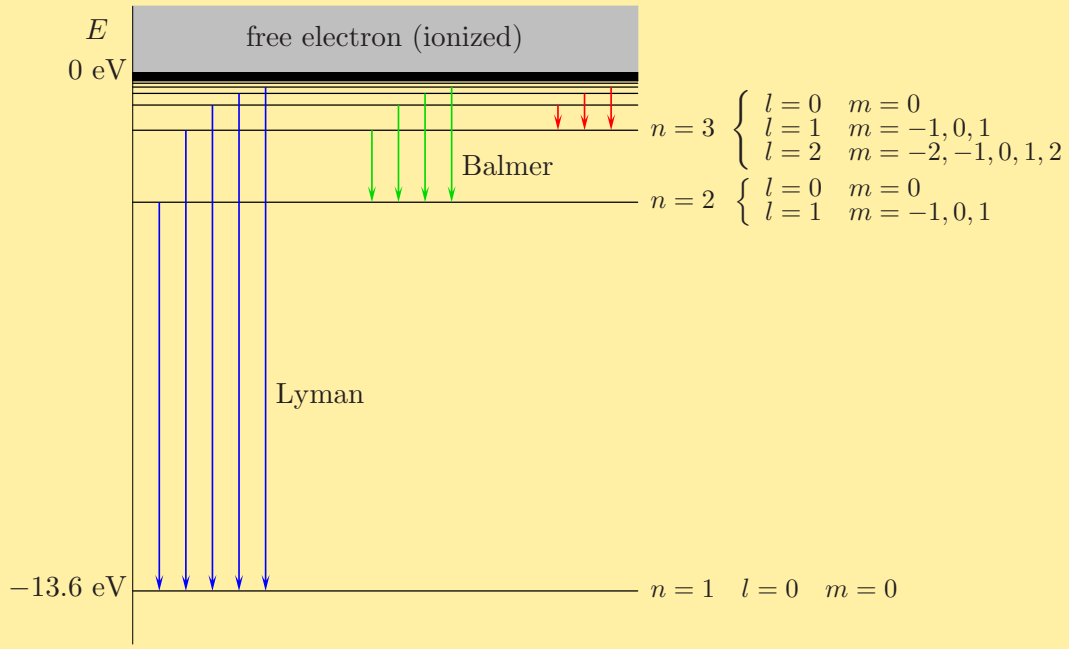


Figure 4.8: Spectrum of the hydrogen atom.

went all the way to infinity. It means physically that while the particle can never escape in a harmonic oscillator, in a hydrogen atom, the electron escapes if its total energy is greater than zero. Such a loss of the electron is called "ionization" of the atom.

There is again a ground state of lowest energy; it has total energy

$$E_1 = -13.6 \text{ eV} \tag{4.39}$$

(an eV or "electron volt" is $1.6\ 10^{-19}$ J). The ground state is the state in which the hydrogen atom will be at absolute zero temperature. In fact, it will still be in the ground state at room temperature, since even then the energy of heat motion is unlikely to raise the energy level of the electron to the next higher one, E_2.

The ionization energy of the hydrogen atom is 13.6 eV; this is the minimum amount of energy that must be added to raise the electron from the ground state to the state of a free electron.

If the electron is excited from the ground state to a higher but still bound energy level, (maybe by passing a spark through hydrogen gas), it will in time again transition back to a lower energy level. Discussion of the reasons and the time evolution of this process will have to wait until chapter 7. For now, it can be pointed out that different transitions are possible, as indicated by the arrows in figure 4.8. They are named by their final energy level to be Lyman, Balmer, or Paschen series transitions.

The energy lost by the electron during a transition is emitted as a quantum of electromagnetic radiation called a photon. The most energetic photons, in the ultraviolet range, are emitted by Lyman transitions. Balmer transitions emit visible light and Paschen ones infrared.

The photons emitted by isolated atoms at rest must have an energy very precisely equal to the difference in energy eigenvalues; anything else would violate the requirement of the orthodox interpretation that only the eigenvalues are observable. And according to the "Planck-Einstein relation," the photon's energy equals the angular frequency ω of its electromagnetic vibration times $\hbar$:

$$E_{n_1} - E_{n_2} = \hbar\omega.$$

Thus the spectrum of the light emitted by hydrogen atoms is very distinctive and can be identified to great accuracy. Different elements have different spectra, and so do molecules. It all allows atoms and molecules to be correctly recognized in a lab or out in space.

(To be sure, the spectral frequencies are not truly mathematically exact numbers. A slight "spectral broadening" is unavoidable because no atom is truly isolated as assumed here; there is always some radiation that perturbs it even in the most ideal empty space. In addition, thermal motion of the atom causes Doppler shifts. In short, only the energy eigenvalues are observable, but exactly what those eigenvalues are for a real-life atom can vary slightly.)

Atoms and molecules may also absorb electromagnetic energy of the same frequencies that they can emit. That allows them to enter an excited state. The excited state will eventually emit the absorbed energy again in a different direction, and possibly at different frequencies by using different transitions. In this way, in astronomy atoms can remove specific frequencies from light that passes them on its way to earth, resulting in an absorption spectrum. Or instead atoms may scatter specific frequencies of light in our direction that was originally not headed to earth, producing an emission spectrum. Doppler shifts can provide information about the thermal and average motion of the atoms. Since hydrogen is so prevalent in the universe, its energy levels as derived here are particularly important in astronomy. Chapter 7 will address the mechanisms of emission and absorption in much greater detail.

Key Points

- The energy levels of the electron in a hydrogen atom have a highest value. This energy is by convention taken to be the zero level.
- The ground state has a energy 13.6 eV below this zero level.
- If the electron in the ground state is given an additional amount of energy that exceeds the 13.6 eV, it has enough energy to escape from the nucleus. This is called ionization of the atom.

- If the electron transitions from a bound energy state with a higher principal quantum number n_1 to a lower one n_2, it emits radiation with an angular frequency ω given by

$$\hbar\omega = E_{n_1} - E_{n_2}$$

- Similarly, atoms with energy E_{n_2} may absorb electromagnetic energy of such a frequency.

4.3.3 Review Questions

1. If there are infinitely many energy levels E_1, E_2, E_3, E_4, E_5, E_6, ..., where did they all go in the energy spectrum?
   Solution hydc-a
2. What is the value of energy level E_2? And E_3?
   Solution hydc-b
3. Based on the results of the previous question, what is the color of the light emitted in a Balmer transition from energy E_3 to E_2? The Planck-Einstein relation says that the angular frequency ω of the emitted photon is its energy divided by $\hbar$, and the wave length of light is $2\pi c/\omega$ where c is the speed of light. Typical wave lengths of visible light are: violet 400 nm, indigo 445 nm, blue 475 nm, green 510 nm, yellow 570 nm, orange 590 nm, red 650 nm.
   Solution hydc-c
4. What is the color of the light emitted in a Balmer transition from an energy level E_n with a high value of n to E_2?
   Solution hydc-d

4.3.4 Discussion of the eigenfunctions

The appearance of the energy eigenstates will be of great interest in understanding the heavier elements and chemical bonds. This subsection describes the most important of them.

It may be recalled from subsection 4.3.2 that there is one eigenfunction ψ_{nlm} for each set of three integer quantum numbers. They are the principal quantum number n (determining the energy of the state), the azimuthal quantum number l (determining the square angular momentum), and the magnetic quantum number m (determining the angular momentum in the chosen z-direction.) They must satisfy the requirements that

$$n > l \geqslant |m|$$

For the ground state, with the lowest energy E_1, $n = 1$ and hence according to the conditions above both l and m must be zero. So the ground state eigenfunction is ψ_{100}; it is unique.

The expression for the wave function of the ground state is (from the results of subsection 4.3.2):

$$\psi_{100}(r) = \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0} \tag{4.40}$$

where a_0 is called the "Bohr radius",

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} \approx 0.529\,177\,210\,7\ 10^{-10}\ \text{m} \tag{4.41}$$

The square magnitude of the energy states will again be displayed as grey tones, darker regions corresponding to regions where the electron is more likely to be found. The ground state is shown this way in figure 4.9; the electron may be found within a blob size that is about three times the Bohr radius, or roughly an Ångstrom, (10^{-10} m), in diameter.

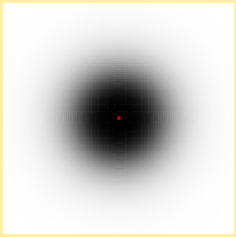


Figure 4.9: Ground state wave function of the hydrogen atom.

It is the quantum mechanical refusal of electrons to restrict themselves to a single location that gives atoms their size. If Planck's constant $\hbar$ would have been zero, so would have been the Bohr radius, and the electron would have been in the nucleus. It would have been a very different world.

The ground state probability distribution is spherically symmetric: the probability of finding the electron at a point depends on the distance from the nucleus, but not on the angular orientation relative to it.

The excited energy levels E_2, E_3, ... are all degenerate; as the spectrum figure 4.8 indicated, there is more than one eigenstate producing each level. Let's have a look at the states at energy level E_2 now.

Figure 4.10 shows energy eigenfunction ψ_{200}. Like ψ_{100}, it is spherically symmetric. In fact, all eigenfunctions ψ_{n00} are spherically symmetric. However, the wave function has blown up a lot, and now separates into a small, more or less spherical region in the center, surrounded by a second region that forms a spherical shell. Separating the two is a radius at which there is zero probability of finding the electron.

The state ψ_{200} is commonly referred to as the "2s" state. The 2 indicates that it is a state with energy E_2. The "s" indicates that the azimuthal quantum number is zero; just think "*s*pherically *s*ymmetric." Similarly, the ground state ψ_{100} is commonly indicated as "1s", having the lowest energy E_1.

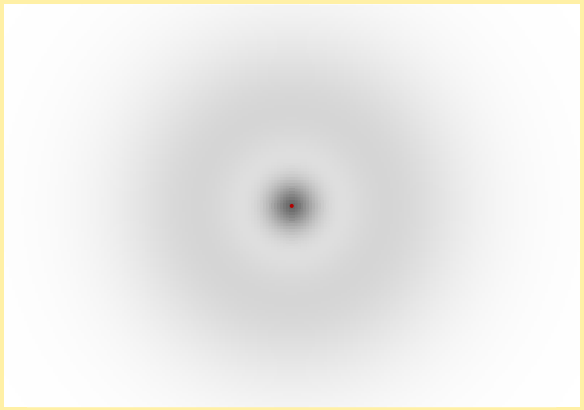


Figure 4.10: Eigenfunction ψ_{200}.

States which have azimuthal quantum number $l = 1$ are called "p" states, for some historical reason. Historically, physicists have always loved confusing and inconsistent notations. In particular, the ψ_{21m} states are called "2p" states. As first example of such a state, figure 4.11 shows ψ_{210}. This wave function squeezes itself close to the z-axis, which is plotted horizontally by convention. There is zero probability of finding the electron at the vertical x, y symmetry plane, and maximum probability at two symmetric points on the z-axis. Since the wave

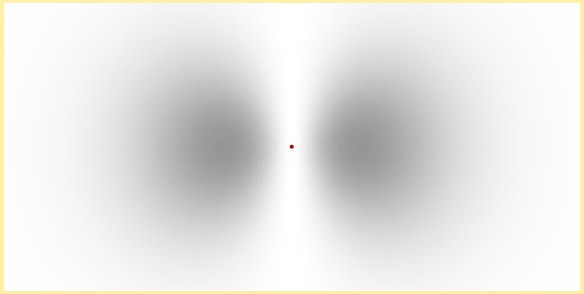


Figure 4.11: Eigenfunction ψ_{210}, or $2p_z$.

function squeezes close to the z-axis, this state is often more specifically referred to as the "$2p_z$" state. Think "*p*oints along the z-axis."

Figure 4.12 shows the other two "2p" states, ψ_{211} and ψ_{21-1}. These two states look exactly the same as far as the probability density is concerned. It is somewhat hard to see in the figure, but they really take the shape of a torus around the left-to-right z-axis.

Eigenfunctions ψ_{200}, ψ_{210}, ψ_{211}, and ψ_{21-1} are degenerate: they all four have the same energy $E_2 = -3.4$ eV. The consequence is that they are not unique. Combinations of them can be formed that have the same energy. These combination states may be more important physically than the original eigenfunctions.

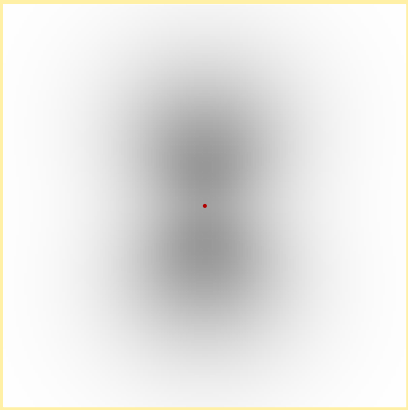


Figure 4.12: Eigenfunction ψ_{211} (and ψ_{21-1}).

In particular, the torus-shaped eigenfunctions ψ_{211} and ψ_{21-1} are often not very useful for descriptions of heavier elements and chemical bonds. Two states that are more likely to be relevant here are called $2\mathrm{p}_x$ and $2\mathrm{p}_y$; they are the combination states:

$$2\mathrm{p}_x\text{:}\ \frac{1}{\sqrt{2}}\left(-\psi_{211}+\psi_{21-1}\right) \qquad 2\mathrm{p}_y\text{:}\ \frac{\mathrm{i}}{\sqrt{2}}\left(\psi_{211}+\psi_{21-1}\right) \tag{4.42}$$

These two states are shown in figure 4.13; they look exactly like the "pointer" state $2\mathrm{p}_z$ of figure 4.11, except that they squeeze along the x-axis, respectively the y-axis, instead of along the z-axis. (Since the y-axis is pointing towards you, $2\mathrm{p}_y$ looks rotationally symmetric. Seen from the side, it would look like p_z in figure 4.11.)

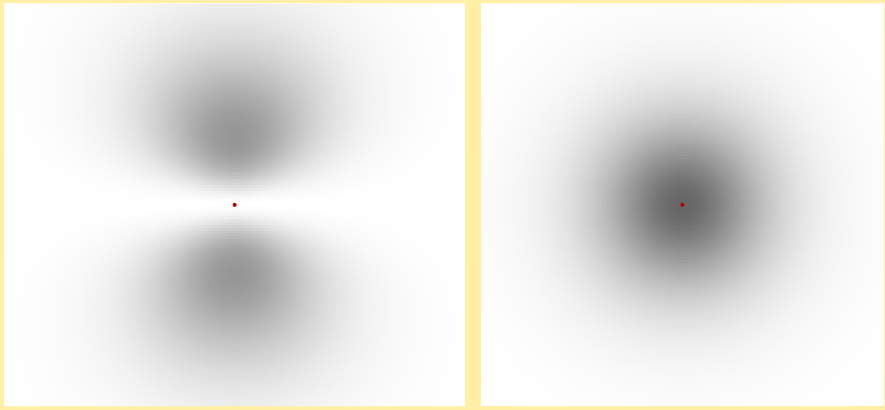


Figure 4.13: Eigenfunctions $2\mathrm{p}_x$, left, and $2\mathrm{p}_y$, right.

Note that unlike the two original states ψ_{211} and ψ_{21-1}, the states $2\mathrm{p}_x$ and $2\mathrm{p}_y$ do not have a definite value of the z-component of angular momentum; the z-component has a 50/50 uncertainty of being either $+\hbar$ or $-\hbar$. But that is

not important in most circumstances. What is important is that when multiple electrons occupy the p states, mutual repulsion effects tend to push them into the p_x, p_y, and p_z states.

So, the four independent eigenfunctions at energy level E_2 are best thought of as consisting of one spherically symmetrical 2s state, and three directional states, $2p_x$, $2p_y$, and $2p_z$, pointing along the three coordinate axes.

But even that is not always ideal; as discussed in chapter 5.11.4, for many chemical bonds, especially those involving the important element carbon, still different combination states called "hybrids" show up. They involve combinations of the 2s and the 2p states and therefore have uncertain square angular momentum as well.

Key Points

- The typical size of eigenstates is given by the Bohr radius, making the size of the atom of the order of an Å.
- The ground state ψ_{100}, or 1s state, is nondegenerate: no other set of quantum numbers n, l, m produces energy E_1.
- All higher energy levels are degenerate, there is more than one eigenstate producing that energy.
- All states of the form ψ_{n00}, including the ground state, are spherically symmetric, and are called s states. The ground state ψ_{100} is the 1s state, ψ_{200} is the 2s state, etcetera.
- States of the form ψ_{n1m} are called p states. The basic 2p states are ψ_{21-1}, ψ_{210}, and ψ_{211}.
- The state ψ_{210} is also called the $2p_z$ state, since it squeezes itself around the z-axis.
- There are similar $2p_x$ and $2p_y$ states that squeeze around the x and y axes. Each is a combination of ψ_{21-1} and ψ_{211}.
- The four spatial states at the E_2 energy level can therefore be thought of as one spherically symmetric 2s state and three 2p pointer states along the axes.
- However, since the E_2 energy level is degenerate, eigenstates of still different shapes are likely to show up in applications.

4.3.4 Review Questions

1. At what distance r from the nucleus does the square of the ground state wave function become less than one percent of its value at the nucleus? Express it both as a multiple of the Bohr radius a_0 and in Å.

   Solution hydd-a

2. Check from the conditions

$$n > l \geqslant |m|$$

that ψ_{200}, ψ_{211}, ψ_{210}, and ψ_{21-1} are the only states of the form ψ_{nlm} that have energy E_2. (Of course, all their combinations, like $2\mathrm{p}_x$ and $2\mathrm{p}_y$, have energy E_2 too, but they are not simply of the form ψ_{nlm}, but combinations of the "basic" solutions ψ_{200}, ψ_{211}, ψ_{210}, and ψ_{21-1}.)

Solution hydd-b

3. Check that the states

$$2\mathrm{p}_x = \frac{1}{\sqrt{2}}\,(-\psi_{211} + \psi_{21-1}) \qquad 2\mathrm{p}_y = \frac{\mathrm{i}}{\sqrt{2}}\,(\psi_{211} + \psi_{21-1})$$

are properly normalized.

Solution hydd-c

4.4 Expectation Value and Standard Deviation

It is a striking consequence of quantum mechanics that physical quantities may not have a value. This occurs whenever the wave function is not an eigenfunction of the quantity of interest. For example, the ground state of the hydrogen atom is not an eigenfunction of the position operator $\widehat{x}$, so the x-position of the electron does not have a value. According to the orthodox interpretation, it cannot be predicted with certainty what a measurement of such a quantity will produce.

However, it is possible to say something if the same measurement is done on a large number of systems that are all the same before the measurement. An example would be x-position measurements on a large number of hydrogen atoms that are all in the ground state before the measurement. In that case, it is relatively straightforward to predict what the average, or "expectation value," of all the measurements will be.

The expectation value is certainly not a replacement for the classical value of physical quantities. For example, for the hydrogen atom in the ground state, the expectation position of the electron is in the nucleus by symmetry. Yet because the nucleus is so small, measurements will never find it there! (The typical measurement will find it a distance comparable to the Bohr radius away.) Actually, that is good news, because if the electron would be in the nucleus as a classical particle, its potential energy would be almost minus infinity instead of the correct value of about -27 eV. It would be a very different universe. Still, having an expectation value is of course better than having no information at all.

The average discrepancy between the expectation value and the actual measurements is called the "standard deviation.". In the hydrogen atom example, where typically the electron is found a distance comparable to the Bohr radius

away from the nucleus, the standard deviation in the x-position turns out to be exactly one Bohr radius. (The same of course for the standard deviations in the y and z positions away from the nucleus.)

In general, the standard deviation is the quantitative measure for how much uncertainty there is in a physical value. If the standard deviation is very small compared to what you are interested in, it is probably OK to use the expectation value as a classical value. It is perfectly fine to say that the electron of the hydrogen atom that you are measuring is in your lab but it is not OK to say that it has countless electron volts of negative potential energy because it is in the nucleus.

This section discusses how to find expectation values and standard deviations after a brief introduction to the underlying ideas of statistics.

Key Points

- The expectation value is the average value obtained when doing measurements on a large number of initially identical systems. It is as close as quantum mechanics can come to having classical values for uncertain physical quantities.
- The standard deviation is how far the individual measurements on average deviate from the expectation value. It is the quantitative measure of uncertainty in quantum mechanics.

4.4.1 Statistics of a die

Since it seems to us humans as if, in Einstein's words, God is playing dice with the universe, it may be a worthwhile idea to examine the statistics of a die first.

For a fair die, each of the six numbers will, on average, show up a fraction 1/6 of the number of throws. In other words, each face has a probability of 1/6.

The average value of a large number of throws is called the expectation value. For a fair die, the expectation value is 3.5. After all, number 1 will show up in about 1/6 of the throws, as will numbers 2 through 6, so the average is

$$\frac{(\text{number of throws}) \times (\frac{1}{6}\,1 + \frac{1}{6}\,2 + \frac{1}{6}\,3 + \frac{1}{6}\,4 + \frac{1}{6}\,5 + \frac{1}{6}\,6)}{\text{number of throws}} = 3.5$$

The general rule to get the expectation value is to sum the probability for each value times the value. In this example:

$$\frac{1}{6}\,1 + \frac{1}{6}\,2 + \frac{1}{6}\,3 + \frac{1}{6}\,4 + \frac{1}{6}\,5 + \frac{1}{6}\,6 = 3.5$$

Note that the name "expectation value" is very poorly chosen. Even though the *average* value of a lot of throws will be 3.5, you would surely not *expect* to throw 3.5. But it is probably too late to change the name now.

The maximum possible deviation from the expectation value does of course occur when you throw a 1 or a 6; the absolute deviation is then $|1 - 3.5| = |6 - 3.5| = 2.5$. It means that the possible values produced by a throw can deviate as much as 2.5 from the expectation value.

However, the maximum possible deviation from the average is not a useful concept for quantities like position, or for the energy levels of the harmonic oscillator, where the possible values extend all the way to infinity. So, instead of the *maximum* deviation from the expectation value, some *average* deviation is better. The most useful of those is called the "standard deviation", denoted by σ. It is found in two steps: first the average *square* deviation from the expectation value is computed, and then a square root is taken of that. For the die that works out to be:

$$\begin{aligned} \sigma &= [\tfrac{1}{6}(1-3.5)^2 + \tfrac{1}{6}(2-3.5)^2 + \tfrac{1}{6}(3-3.5)^2 + \\ &\qquad \tfrac{1}{6}(4-3.5)^2 + \tfrac{1}{6}(5-3.5)^2 + \tfrac{1}{6}(6-3.5)^2]^{1/2} \\ &= 1.71 \end{aligned}$$

On average then, the throws are 1.71 points off from 3.5.

Key Points

- The expectation value is obtained by summing the possible values times their probabilities.
- To get the standard deviation, first find the average square deviation from the expectation value, then take a square root of that.

4.4.1 Review Questions

1. Suppose you toss a coin a large number of times, and count heads as one, tails as two. What will be the expectation value?
   Solution esda-a
2. Continuing this example, what will be the maximum deviation?
   Solution esda-b
3. Continuing this example, what will be the standard deviation?
   Solution esda-c
4. Have I got a die for you! By means of a small piece of lead integrated into its light-weight structure, it does away with that old-fashioned uncertainty. It comes up six every time! What will be the expectation value of your throws? What will be the standard deviation?
   Solution esda-d

4.4.2 Statistics of quantum operators

The expectation values of the operators of quantum mechanics are defined in the same way as those for the die.

Consider an arbitrary physical quantity, call it a, and assume it has an associated operator A. For example, if the physical quantity a is the total energy E, A will be the Hamiltonian H.

The equivalent of the face values of the die are the values that the quantity a can take, and according to the orthodox interpretation, that are the eigenvalues

$$a_1,\ a_2,\ a_3,\ \ldots$$

of the operator A.

Next, the probabilities of getting those values are according to quantum mechanics the square magnitudes of the coefficients when the wave function is written in terms of the eigenfunctions of A. In other words, if α_1, α_2, α_3, ... are the eigenfunctions of operator A, and the wave function is

$$\Psi = c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \ldots$$

then $|c_1|^2$ is the probability of value a_1, $|c_2|^2$ the probability of value a_2, etcetera.

The expectation value is written as $\langle a \rangle$, or as $\langle A \rangle$, whatever is more appealing. Like for the die, it is found as the sum of the probability of each value times the value:

$$\langle a \rangle = |c_1|^2 a_1 + |c_2|^2 a_2 + |c_3|^2 a_3 + \ldots$$

Of course, the eigenfunctions might be numbered using multiple indices; that does not really make a difference. For example, the eigenfunctions ψ_{nlm} of the hydrogen atom are numbered with three indices. In that case, if the wave function of the hydrogen atom is

$$\Psi = c_{100}\psi_{100} + c_{200}\psi_{200} + c_{210}\psi_{210} + c_{211}\psi_{211} + c_{21-1}\psi_{21-1} + c_{300}\psi_{300} + \ldots$$

then the expectation value for energy will be, noting that $E_1 = -13.6$ eV, $E_2 = -3.4$ eV, ...:

$$\langle E \rangle = -|c_{100}|^2 13.6 \text{ eV} - |c_{200}|^2 3.4 \text{ eV} - |c_{210}|^2 3.4 \text{ eV} - |c_{211}|^2 3.4 \text{ eV} - \ldots$$

Also, the expectation value of the square angular momentum will be, recalling that its eigenvalues are $l(l+1)\hbar^2$,

$$\langle L^2 \rangle = |c_{100}|^2 0 + |c_{200}|^2 0 + |c_{210}|^2 2\hbar^2 + |c_{211}|^2 2\hbar^2 + |c_{21-1}|^2 2\hbar^2 + |c_{300}|^2 0 + \ldots$$

Also, the expectation value of the z-component of angular momentum will be, recalling that its eigenvalues are $m\hbar$,

$$\langle L_z \rangle = |c_{100}|^2 0 + |c_{200}|^2 0 + |c_{210}|^2 0 + |c_{211}|^2 \hbar - |c_{21-1}|^2 \hbar + |c_{300}|^2 0 + \ldots$$

Key Points

- The expectation value of a physical quantity is found by summing its eigenvalues times the probability of measuring that eigenvalue.

- To find the probabilities of the eigenvalues, the wave function Ψ can be written in terms of the eigenfunctions of the physical quantity. The probabilities will be the square magnitudes of the coefficients of the eigenfunctions.

4.4.2 Review Questions

1. The $2p_x$ pointer state of the hydrogen atom was defined as

$$\frac{1}{\sqrt{2}}\left(-\psi_{211}+\psi_{21-1}\right).$$

What are the expectation values of energy, square angular momentum, and z angular momentum for this state?

Solution esdb-a

2. Continuing the previous question, what are the standard deviations in energy, square angular momentum, and z angular momentum?

Solution esdb-b

4.4.3 Simplified expressions

The procedure described in the previous section to find the expectation value of a quantity is unwieldy: it requires that first the eigenfunctions of the quantity are found, and next that the wave function is written in terms of those eigenfunctions. There is a quicker way.

Assume that you want to find the expectation value, $\langle a \rangle$ or $\langle A \rangle$, of some quantity a with associated operator A. The simpler way to do it is as an inner product:

$$\boxed{\langle A \rangle = \langle \Psi | A | \Psi \rangle.} \tag{4.43}$$

(Recall that $\langle \Psi | A | \Psi \rangle$ is just the inner product $\langle \Psi | A\Psi \rangle$; the additional separating bar is often visually convenient, though.) This formula for the expectation value is easily remembered as "leaving out Ψ" from the inner product bracket. The reason that $\langle \Psi | A | \Psi \rangle$ works for getting the expectation value is given in derivation {D.17}.

The simplified expression for the expectation value can also be used to find the standard deviation, σ_A or σ_a:

$$\boxed{\sigma_A = \sqrt{\langle (A - \langle A \rangle)^2 \rangle}} \tag{4.44}$$

where $\langle (A - \langle A \rangle)^2 \rangle$ is the inner product $\langle \Psi | (A - \langle A \rangle)^2 \Psi \rangle$.

Key Points

- The expectation value of a quantity a with operator A can be found as $\langle A \rangle = \langle \Psi | A\Psi \rangle$.

Similarly, the standard deviation can be found using the expression $\sigma_A = \sqrt{\langle (A - \langle A \rangle)^2 \rangle}$.

4.4.3 Review Questions

1. The $2p_x$ pointer state of the hydrogen atom was defined as

$$\frac{1}{\sqrt{2}} (-\psi_{211} + \psi_{21-1}) .$$

where both ψ_{211} and ψ_{21-1} are eigenfunctions of the total energy Hamiltonian H with eigenvalue E_2 and of square angular momentum $\widehat{L}^2$ with eigenvalue $2\hbar^2$; however, ψ_{211} is an eigenfunction of z angular momentum $\widehat{L}_z$ with eigenvalue $\hbar$, while ψ_{21-1} is one with eigenvalue $-\hbar$. Evaluate the expectation values of energy, square angular momentum, and z angular momentum in the $2p_x$ state using inner products. (Of course, since $2p_x$ is already written out in terms of the eigenfunctions, there is no simplification in this case.)

Solution esdb2-a

2. Continuing the previous question, evaluate the standard deviations in energy, square angular momentum, and z angular momentum in the $2p_x$ state using inner products.

Solution esdb2-b

4.4.4 Some examples

This section gives some examples of expectation values and standard deviations for known wave functions.

First consider the expectation value of the energy of the hydrogen atom in its ground state ψ_{100}. The ground state is an energy eigenfunction with the lowest possible energy level $E_1 = -13.6$ eV as eigenvalue. So, according to the orthodox interpretation, energy measurements of the ground state can only return the value E_1, with 100% certainty.

Clearly, if all measurements return the value E_1, then the average value must be that value too. So the expectation value $\langle E \rangle$ should be E_1. In addition, the measurements will never deviate from the value E_1, so the standard deviation σ_E should be zero.

It is instructive to check those conclusions using the simplified expressions for expectation values and standard deviations from the previous subsection. The expectation value can be found as:

$$\langle E \rangle = \langle H \rangle = \langle \Psi | H | \Psi \rangle$$

In the ground state

$$\Psi = c_{100} \psi_{100}$$

where c_{100} is a constant of magnitude one, and ψ_{100} is the ground state eigenfunction of the Hamiltonian H with the lowest eigenvalue E_1. Substituting this Ψ, the expectation value of the energy becomes

$$\langle E \rangle = \langle c_{100}\psi_{100} | H c_{100}\psi_{100} \rangle = c_{100}^* c_{100} \langle \psi_{100} | E_1 \psi_{100} \rangle = c_{100}^* c_{100} E_1 \langle \psi_{100} | \psi_{100} \rangle$$

since $H\psi_{100} = E_1\psi_{100}$ by the definition of eigenfunction. Note that constants come out of the inner product bra as their complex conjugate, but unchanged out of the ket. The final expression shows that $\langle E \rangle = E_1$ as it should, since c_{100} has magnitude one, while $\langle \psi_{100} | \psi_{100} \rangle = 1$ because proper eigenfunctions are normalized to one. So the expectation value checks out OK.

The standard deviation

$$\sigma_E = \sqrt{\langle (H - \langle E \rangle)^2 \rangle}$$

checks out OK too:

$$\sigma_E = \sqrt{\langle \psi_{100} | (H - E_1)^2 \psi_{100} \rangle}$$

and since $H\psi_{100} = E_1\psi_{100}$, you have that $(H - E_1)\psi_{100}$ is zero, so σ_E is zero as it should be.

In general,

> *If the wave function is an eigenfunction of the measured variable, the expectation value will be the eigenvalue, and the standard deviation will be zero.*

To get uncertainty, in other words, a nonzero standard deviation, the wave function should not be an eigenfunction of the quantity being measured.

For example, the ground state of the hydrogen atom is an energy eigenfunction, but not an eigenfunction of the position operators. The expectation value for the position coordinate x can still be found as an inner product:

$$\langle x \rangle = \langle \psi_{100} | \widehat{x} \psi_{100} \rangle = \iiint x |\psi_{100}|^2 \, \mathrm{d}x \mathrm{d}y \mathrm{d}z.$$

This integral is zero. The reason is that $|\psi_{100}|^2$, shown as grey scale in figure 4.9, is symmetric around $x = 0$; it has the same value at a negative value of x as at the corresponding positive value. Since the factor x in the integrand changes sign, integration values at negative x cancel out against those at positive x. So $\langle x \rangle = 0$.

The position coordinates y and z go the same way, and it follows that the expectation value of position is at $(x, y, z) = (0,0,0)$; the expectation position of the electron is in nucleus.

In fact, all basic energy eigenfunctions ψ_{nlm} of the hydrogen atom, like figures 4.9, 4.10, 4.11, 4.12, as well as the combination states $2\mathrm{p}_x$ and $2\mathrm{p}_y$ of figure 4.13, have a symmetric probability distribution, and all have the expectation value of

position in the nucleus. (For the hybrid states discussed later, that is no longer true.)

But don't really expect to ever find the electron in the negligible small nucleus! You will find it at locations that are on average one standard deviation away from it. For example, in the ground state

$$\sigma_x = \sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\langle x^2 \rangle} = \sqrt{\iiint x^2 |\psi_{100}(x,y,z)|^2 \,\mathrm{d}x\mathrm{d}y\mathrm{d}z}$$

which is positive since the integrand is everywhere positive. So, the results of x-position measurements are uncertain, even though they average out to the nominal position $x = 0$. The negative experimental results for x average away against the positive ones. The same is true in the y and z directions. Thus the expectation position becomes the nucleus even though the electron will really never be found there.

If you actually do the integral above, (it is not difficult in spherical coordinates,) you find that the standard deviation in x equals the Bohr radius. So on average, the electron will be found at an x-distance equal to the Bohr radius away from the nucleus. Similar deviations will occur in the y and z directions.

The expectation value of linear momentum in the ground state can be found from the linear momentum operator $\widehat{p}_x = \hbar\partial/\mathrm{i}\partial x$:

$$\langle p_x \rangle = \langle \psi_{100} | \widehat{p}_x \psi_{100} \rangle = \iiint \psi_{100} \frac{\hbar}{\mathrm{i}} \frac{\partial \psi_{100}}{\partial x} \,\mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\hbar}{\mathrm{i}} \iiint \frac{\partial \frac{1}{2}\psi_{100}^2}{\partial x} \,\mathrm{d}x\mathrm{d}y\mathrm{d}z$$

This is again zero, since differentiation turns a symmetric function into an antisymmetric one, one which changes sign between negative and corresponding positive positions. Alternatively, just perform integration with respect to x, noting that the wave function is zero at infinity.

More generally, the expectation value for linear momentum is zero for all the energy eigenfunctions; that is a consequence of Ehrenfest's theorem covered in chapter 7.2.1. The standard deviations are again nonzero, so that linear momentum is uncertain like position is.

All these observations carry over in the same way to the eigenfunctions $\psi_{n_x n_y n_z}$ of the harmonic oscillator. They too all have the expectation values of position at the origin, in other words in the nucleus, and the expectation linear momenta equal to zero.

If combinations of energy eigenfunctions are considered, it changes. Such combinations may have nontrivial expectation positions and linear momenta. A discussion will have to wait until chapter 7.

Key Points

- Examples of definite and uncertain quantities were given for example wave functions.

o— A quantity has a definite value when the wave function is an eigenfunction of the operator corresponding to that quantity.

4.5 The Commutator

As the previous section discussed, the standard deviation σ is a measure of the uncertainty of a property of a quantum system. The larger the standard deviation, the farther typical measurements stray from the expected average value. Quantum mechanics often requires a minimum amount of uncertainty when more than one quantity is involved, like position and linear momentum in Heisenberg's uncertainty principle. In general, this amount of uncertainty is related to an important mathematical object called the "commutator", to be discussed in this section.

4.5.1 Commuting operators

First, note that there is no fundamental reason why several quantities cannot have a definite value at the same time. For example, if the electron of the hydrogen atom is in a ψ_{nlm} eigenstate, its total energy, square angular momentum, and z-component of angular momentum all have definite values, with zero uncertainty.

More generally, two different quantities with operators A and B have definite values if the wave function is an eigenfunction of both A and B. So, the question whether two quantities can be definite at the same time is really whether their operators A and B have common eigenfunctions. And it turns out that the answer has to do with whether these operators "commute", in other words, on whether their order can be reversed as in $AB = BA$.

In particular, {D.18}:

> *Iff two Hermitian operators commute, there is a complete set of eigenfunctions that is common to them both.*

(For more than two operators, each operator has to commute with all others.)

For example, the operators H_x and H_y of the harmonic oscillator of chapter 4.1.2 commute:

$$\begin{aligned} H_xH_y\Psi &= \left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \tfrac{1}{2}cx^2\right]\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial y^2} + \tfrac{1}{2}cy^2\right]\Psi \\ &= \left(\frac{\hbar^2}{2m}\right)^2 \frac{\partial^4\Psi}{\partial x^2\partial y^2} - \frac{\hbar^2}{2m}\frac{\partial^2 \tfrac{1}{2}cy^2\Psi}{\partial x^2} - \tfrac{1}{2}cx^2\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial y^2} + \tfrac{1}{2}cx^2\tfrac{1}{2}cy^2\Psi \\ &= H_yH_x\Psi \end{aligned}$$

This is true since it makes no difference whether you differentiate Ψ first with respect to x and then with respect to y or vice versa, and since the $\frac{1}{2}cy^2$ can be pulled in front of the x-differentiations and the $\frac{1}{2}cx^2$ can be pushed inside the y-differentiations, and since multiplications can always be done in any order.

The same way, H_z commutes with H_x and H_y, and that means that H commutes with them all, since H is just their sum. So, these four operators should have a common set of eigenfunctions, and they do: it is the set of eigenfunctions $\psi_{n_x n_y n_z}$ derived in chapter 4.1.2.

Similarly, for the hydrogen atom, the total energy Hamiltonian H, the square angular momentum operator $\widehat{L}^2$ and the z-component of angular momentum $\widehat{L}_z$ all commute, and they have the common set of eigenfunctions ψ_{nlm}.

Note that such eigenfunctions are not necessarily the only game in town. As a counter-example, for the hydrogen atom H, $\widehat{L}^2$, and the x-component of angular momentum $\widehat{L}_x$ also all commute, and they too have a common set of eigenfunctions. But that will *not* be the ψ_{nlm}, since $\widehat{L}_x$ and $\widehat{L}_z$ do not commute. (It will however be the ψ_{nlm} after you rotate them all 90 degrees around the y-axis.) It would certainly be simpler mathematically if each operator had just one unique set of eigenfunctions, but nature does not cooperate.

Key Points

- Operators commute if you can change their order, as in $AB = BA$.
- For commuting operators, a common set of eigenfunctions exists.
- For those eigenfunctions, the physical quantities corresponding to the commuting operators all have definite values at the same time.

4.5.1 Review Questions

1. The pointer state

$$2\mathrm{p}_x = \frac{1}{\sqrt{2}}\left(-\psi_{211} + \psi_{21-1}\right).$$

is one of the eigenstates that H, $\widehat{L}^2$, and $\widehat{L}_x$ have in common. Check that it is not an eigenstate that H, $\widehat{L}^2$, and $\widehat{L}_z$ have in common.

Solution commutea-a

4.5.2 Noncommuting operators and their commutator

Two quantities with operators that do not commute cannot in general have definite values at the same time. If one has a definite value, the other is in general uncertain.

The qualification "in general" is needed because there may be exceptions. The angular momentum operators do not commute, but it is still possible for the angular momentum to be zero in all three directions. But as soon as the

angular momentum in any direction is nonzero, only one component of angular momentum can have a definite value.

A measure for the amount to which two operators A and B do not commute is the difference between AB and BA; this difference is called their "commutator" $[A, B]$:

$$\boxed{[A, B] \equiv AB - BA} \tag{4.45}$$

A nonzero commutator $[A, B]$ demands a minimum amount of uncertainty in the corresponding quantities a and b. It can be shown, {D.19}, that the uncertainties, or standard deviations, σ_a in a and σ_b in b are at least so large that:

$$\boxed{\sigma_a \sigma_b \geqslant \tfrac{1}{2} |\langle [A, B] \rangle|} \tag{4.46}$$

This equation is called the "generalized uncertainty relationship".

Key Points

- The commutator of two operators A and B equals $AB - BA$ and is written as $[A, B]$.
- The product of the uncertainties in two quantities is at least one half the magnitude of the expectation value of their commutator.

4.5.3 The Heisenberg uncertainty relationship

This section will work out the uncertainty relationship (4.46) of the previous subsection for the position and linear momentum in an arbitrary direction. The result will be a precise mathematical statement of the Heisenberg uncertainty principle.

To be specific, the arbitrary direction will be taken as the x-axis, so the position operator will be $\widehat{x}$, and the linear momentum operator $\widehat{p}_x = \hbar\partial/\mathrm{i}\partial x$. These two operators do not commute, $\widehat{p}_x \widehat{x} \Psi$ is simply not the same as $\widehat{x}\widehat{p}_x \Psi$: $\widehat{p}_x \widehat{x} \Psi$ means multiply function Ψ by x to get the product function $x\Psi$ and then apply $\widehat{p}_x$ on that product, while $\widehat{x}\widehat{p}_x \Psi$ means apply $\widehat{p}_x$ on Ψ and then multiply the resulting function by x. The difference is found from writing it out:

$$\widehat{p}_x \widehat{x} \Psi = \frac{\hbar}{\mathrm{i}} \frac{\partial x \Psi}{\partial x} = \frac{\hbar}{\mathrm{i}} \Psi + \frac{\hbar}{\mathrm{i}} x \frac{\partial \Psi}{\partial x} = -\mathrm{i}\hbar \Psi + \widehat{x}\widehat{p}_x \Psi$$

the second equality resulting from differentiating out the product.

Comparing start and end shows that the difference between $\widehat{x}\widehat{p}_x$ and $\widehat{p}_x \widehat{x}$ is not zero, but $\mathrm{i}\hbar$. By definition, this difference is their commutator:

$$\boxed{[\widehat{x}, \widehat{p}_x] = \mathrm{i}\hbar} \tag{4.47}$$

This important result is called the "canonical commutation relation." The commutator of position and linear momentum in the same direction is the nonzero constant $\mathrm{i}\hbar$.

Because the commutator is nonzero, there must be nonzero uncertainty involved. Indeed, the generalized uncertainty relationship of the previous subsection becomes in this case:

$$\sigma_x \sigma_{p_x} \geqslant \tfrac{1}{2}\hbar \tag{4.48}$$

This is the uncertainty relationship as first formulated by Heisenberg.

It implies that when the uncertainty in position σ_x is narrowed down to zero, the uncertainty in momentum σ_{p_x} must become infinite to keep their product nonzero, and vice versa. More generally, you can narrow down the position of a particle and you can narrow down its momentum. But you can never reduce the product of the uncertainties σ_x and σ_{p_x} below $\frac{1}{2}\hbar$, whatever you do.

It should be noted that the uncertainty relationship is often written as $\Delta p_x \Delta x \geqslant \frac{1}{2}\hbar$ or even as $\Delta p_x \Delta x \approx \hbar$ where Δp and Δx are taken to be vaguely described "uncertainties" in momentum and position, rather than rigorously defined standard deviations. And people write a corresponding uncertainty relationship for time, $\Delta E \Delta t \geqslant \frac{1}{2}\hbar$, because relativity suggests that time should be treated just like space. But note that unlike the linear momentum operator, the Hamiltonian is not at all universal. So, you might guess that the definition of the "uncertainty" Δt in time would not be universal either, and you would be right, chapter 7.2.2.

Key Points

- The canonical commutator $[\widehat{x}, \widehat{p}_x]$ equals $\mathrm{i}\hbar$.
- If either the uncertainty in position in a given direction or the uncertainty in linear momentum in that direction is narrowed down to zero, the other uncertainty blows up.
- The product of the two uncertainties is at least the constant $\frac{1}{2}\hbar$.

4.5.3 Review Questions

1. This sounds serious! If I am driving my car, the police requires me to know my speed (linear momentum). Also, I would like to know where I am. But neither is possible according to quantum mechanics.

   Solution commutec-a

4.5.4 Commutator reference

It is a fact of life in quantum mechanics that commutators pop up all over the place. Not just in uncertainty relations, but also in the time evolution of expectation values, in angular momentum, and in quantum field theory, the advanced

theory of quantum mechanics used in solids and relativistic applications. This section can make your life easier dealing with them. Browse through it to see what is there. Then come back when you need it.

Recall the definition of the commutator $[A, B]$ of any two operators A and B:

$$[A, B] = AB - BA \tag{4.49}$$

By this very definition , the commutator is zero for any two operators A_1 and A_2 that commute, (whose order can be interchanged):

$$[A_1, A_2] = 0 \quad \text{if } A_1 \text{ and } A_2 \text{ commute; } A_1 A_2 = A_2 A_1. \tag{4.50}$$

If operators all commute, all their products commute too:

$$[A_1 A_2 \ldots A_k, A_{k+1} \ldots A_n] = 0 \quad \text{if } A_1, A_2, \ldots, A_k, A_{k+1}, \ldots, A_n \text{ all commute.} \tag{4.51}$$

Everything commutes with itself, of course:

$$[A, A] = 0, \tag{4.52}$$

and everything commutes with a numerical constant; if A is an operator and a is some number, then:

$$[A, a] = [a, A] = 0. \tag{4.53}$$

The commutator is "antisymmetric"; or in simpler words, if you interchange the sides; it will change the sign, {D.20}:

$$[B, A] = -[A, B]. \tag{4.54}$$

For the rest however, linear combinations multiply out just like you would expect:

$$[aA + bB, cC + dD] = ac[A, C] + ad[A, D] + bc[B, C] + bd[B, D], \tag{4.55}$$

(in which it is assumed that A, B, C, and D are operators, and a, b, c, and d numerical constants.)

To deal with commutators that involve products of operators, the rule to remember is: "the first factor comes out at the front of the commutator, the second at the back". More precisely:

$$\underleftarrow{[}\underleftarrow{A}\underrightarrow{B}, \ldots] = A[B, \ldots] + [A, \ldots]B, \qquad \underleftarrow{[}\ldots, \underleftarrow{A}\underrightarrow{B}\underrightarrow{]} = A[\ldots, B] + [\ldots, A]B. \tag{4.56}$$

So, if A or B commutes with the other side of the operator, it can simply be taken out at at its side; (the second commutator will be zero.) For example,

$$[A_1 B, A_2] = A_1[B, A_2], \qquad [BA_1, A_2] = [B, A_2]A_1$$

if A_1 and A_2 commute.

Now from the general to the specific. Because changing sides in a commutator merely changes its sign, from here on only one of the two possibilities will be shown. First the position operators all mutually commute:

$$[\widehat{x},\widehat{y}] = [\widehat{y},\widehat{z}] = [\widehat{z},\widehat{x}] = 0 \tag{4.57}$$

as do position-dependent operators such as a potential energy $V(x,y,z)$:

$$[\widehat{x},V(x,y,z)] = [\widehat{y},V(x,y,z)] = [\widehat{z},V(x,y,z)] = 0 \tag{4.58}$$

This illustrates that if a set of operators all commute, then all combinations of those operators commute too.

The linear momentum operators all mutually commute:

$$[\widehat{p}_x,\widehat{p}_y] = [\widehat{p}_y,\widehat{p}_z] = [\widehat{p}_z,\widehat{p}_x] = 0 \tag{4.59}$$

However, position operators and linear momentum operators in the same direction do *not* commute; instead:

$$[\widehat{x},\widehat{p}_x] = [\widehat{y},\widehat{p}_y] = [\widehat{z},\widehat{p}_z] = \mathrm{i}\hbar \tag{4.60}$$

As seen in the previous subsection, this lack of commutation causes the Heisenberg uncertainty principle. Position and linear momentum operators in different directions do commute:

$$[\widehat{x},\widehat{p}_y] = [\widehat{x},\widehat{p}_z] = [\widehat{y},\widehat{p}_z] = [\widehat{y},\widehat{p}_x] = [\widehat{z},\widehat{p}_x] = [\widehat{z},\widehat{p}_y] = 0 \tag{4.61}$$

A generalization that is frequently very helpful is:

$$[f,\widehat{p}_x] = \mathrm{i}\hbar\frac{\partial f}{\partial x} \quad [f,\widehat{p}_y] = \mathrm{i}\hbar\frac{\partial f}{\partial y} \quad [f,\widehat{p}_z] = \mathrm{i}\hbar\frac{\partial f}{\partial z} \tag{4.62}$$

where f is any function of x, y, and z.

Unlike linear momentum operators, angular momentum operators do *not* mutually commute. The commutators are given by the so-called " fundamental commutation relations:"

$$[\widehat{L}_x,\widehat{L}_y] = \mathrm{i}\hbar\widehat{L}_z \quad [\widehat{L}_y,\widehat{L}_z] = \mathrm{i}\hbar\widehat{L}_x \quad [\widehat{L}_z,\widehat{L}_x] = \mathrm{i}\hbar\widehat{L}_y \tag{4.63}$$

Note the $\ldots xyzxyz\ldots$ order of the indices that produces positive signs; a reversed $\ldots zyxzy\ldots$ order adds a minus sign. For example $[\widehat{L}_z,\widehat{L}_y] = -\mathrm{i}\hbar\widehat{L}_x$ because y following z is in reversed order.

The angular momentum components do all commute with the square angular momentum operator:

$$[\widehat{L}_x,\widehat{L}^2] = [\widehat{L}_y,\widehat{L}^2] = [\widehat{L}_z,\widehat{L}^2] = 0 \quad \text{where } \widehat{L}^2 = \widehat{L}_x^2 + \widehat{L}_y^2 + \widehat{L}_z^2 \tag{4.64}$$

Just the opposite of the situation for linear momentum, position and angular momentum operators in the same direction commute,

$$[\widehat{x}, \widehat{L}_x] = [\widehat{y}, \widehat{L}_y] = [\widehat{z}, \widehat{L}_z] = 0 \tag{4.65}$$

but those in different directions do not:

$$[\widehat{x}, \widehat{L}_y] = [\widehat{L}_x, \widehat{y}] = \mathrm{i}\hbar\widehat{z} \quad [\widehat{y}, \widehat{L}_z] = [\widehat{L}_y, \widehat{z}] = \mathrm{i}\hbar\widehat{x} \quad [\widehat{z}, \widehat{L}_x] = [\widehat{L}_z, \widehat{x}] = \mathrm{i}\hbar\widehat{y} \tag{4.66}$$

Square position commutes with all components of angular momentum,

$$[\widehat{r}^2, \widehat{L}_x] = [\widehat{r}^2, \widehat{L}_y] = [\widehat{r}^2, \widehat{L}_z] = [\widehat{r}^2, \widehat{L}^2] = 0 \tag{4.67}$$

The commutator between position and square angular momentum is, using vector notation for conciseness,

$$[\widehat{\vec{r}}, \widehat{L}^2] = -2\hbar^2\widehat{\vec{r}} - 2\mathrm{i}\hbar\widehat{\vec{r}} \times \widehat{\vec{L}} = -2\hbar^2\widehat{\vec{r}} + 2\mathrm{i}\hbar(\widehat{\vec{r}} \cdot \widehat{\vec{r}})\widehat{\vec{p}} - 2\mathrm{i}\hbar\widehat{\vec{r}}(\widehat{\vec{r}} \cdot \widehat{\vec{p}}) \tag{4.68}$$

The commutators between linear and angular momentum are very similar to the ones between position and angular momentum:

$$[\widehat{p}_x, \widehat{L}_x] = [\widehat{p}_y, \widehat{L}_y] = [\widehat{p}_z, \widehat{L}_z] = 0 \tag{4.69}$$

$$[\widehat{p}_x, \widehat{L}_y] = [\widehat{L}_x, \widehat{p}_y] = \mathrm{i}\hbar\widehat{p}_z \quad [\widehat{p}_y, \widehat{L}_z] = [\widehat{L}_y, \widehat{p}_z] = \mathrm{i}\hbar\widehat{p}_x \quad [\widehat{p}_z, \widehat{L}_x] = [\widehat{L}_z, \widehat{p}_x] = \mathrm{i}\hbar\widehat{p}_y \tag{4.70}$$

$$[\widehat{p}^2, \widehat{L}_x] = [\widehat{p}^2, \widehat{L}_y] = [\widehat{p}^2, \widehat{L}_z] = [\widehat{p}^2, \widehat{L}^2] = 0 \tag{4.71}$$

$$[\widehat{\vec{p}}, \widehat{L}^2] = -2\hbar^2\widehat{\vec{p}} - 2\mathrm{i}\hbar\widehat{\vec{p}} \times \widehat{\vec{L}} = 2\hbar^2\widehat{\vec{p}} + 2\mathrm{i}\hbar(\widehat{\vec{r}} \cdot \widehat{\vec{p}})\widehat{\vec{p}} - 2\mathrm{i}\hbar\widehat{\vec{r}}(\widehat{\vec{p}} \cdot \widehat{\vec{p}}) \tag{4.72}$$

The following commutators are also useful:

$$[\vec{r} \times \widehat{\vec{L}}, \widehat{L}^2] = 2\mathrm{i}\hbar\vec{r}\widehat{L}^2 \qquad [[\vec{r}, \widehat{L}^2], \widehat{L}^2] = 2\hbar^2(\vec{r}\widehat{L}^2 + \widehat{L}^2\vec{r}) \tag{4.73}$$

Commutators involving spin are discussed in a later chapter, 5.5.3.

Key Points

- Rules for evaluating commutators were given.
- Return to this subsection if you need to figure out some commutator or the other.

4.6 The Hydrogen Molecular Ion

The hydrogen atom studied earlier is where full theoretical analysis stops. Larger systems are just too difficult to solve analytically. Yet, it is often quite possible to understand the solution of such systems using approximate arguments. As an example, this section considers the H_2^+-ion. This ion consists of two protons and a single electron circling them. It will be shown that a chemical bond forms that holds the ion together. The bond is a "covalent" one, in which the protons share the electron.

The general approach will be to compute the energy of the ion, and to show that the energy is less when the protons are sharing the electron as a molecule than when they are far apart. This must mean that the molecule is stable: energy must be expended to take the protons apart.

The approximate technique to be used to find the state of lowest energy is a basic example of what is called a "variational method."

4.6.1 The Hamiltonian

First the Hamiltonian is needed. Since the protons are so much heavier than the electron, to good approximation they can be considered fixed points in the energy computation. That is called the "Born-Oppenheimer approximation". In this approximation, only the Hamiltonian of the electron is needed. It makes things a lot simpler, which is why the Born-Oppenheimer approximation is a common assumption in applications of quantum mechanics.

Compared to the Hamiltonian of the hydrogen atom of section 4.3.1, there are now two terms to the potential energy, the electron experiencing attraction to both protons:

$$H = -\frac{\hbar^2}{2m_{\rm e}}\nabla^2 - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_{\rm l}} - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_{\rm r}} \tag{4.74}$$

where $r_{\rm l}$ and $r_{\rm r}$ are the distances from the electron to the left and right protons,

$$r_{\rm l} \equiv |\vec{r} - \vec{r}_{\rm lp}| \qquad r_{\rm r} \equiv |\vec{r} - \vec{r}_{\rm rp}| \tag{4.75}$$

with $\vec{r}_{\rm lp}$ the position of the left proton and $\vec{r}_{\rm rp}$ that of the right one.

The hydrogen ion in the Born-Oppenheimer approximation can be solved analytically using "prolate spheroidal coordinates." However, approximations will be used here. For one thing, you learn more about the physics that way.

Key Points

- In the Born-Oppenheimer approximation, the electronic structure is computed assuming that the nuclei are at fixed positions.
- The Hamiltonian in the Born-Oppenheimer approximation has been found. It is above.

4.6.2 Energy when fully dissociated

The fully dissociated state is when the protons are very far apart and there is no coherent molecule, as in figure 4.14. The best the electron can do under those circumstances is to combine with either proton, say the left one, and form a hydrogen atom in the ground state of lowest energy. In that case the right proton will be alone. According to the solution for the hydrogen atom, the

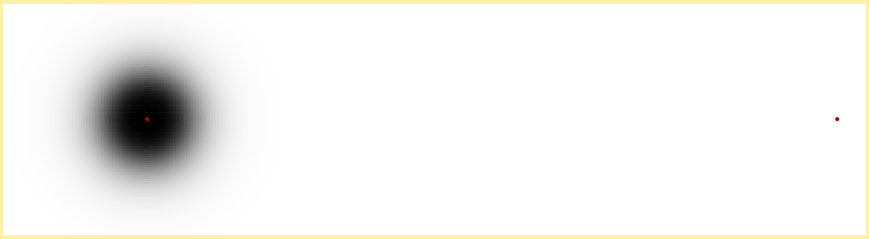


Figure 4.14: Hydrogen atom plus free proton far apart.

electron loses 13.6 eV of energy by going in the ground state around the left proton. Of course, it would lose the same energy going into the ground state around the right proton, but for now, assume that it is around the left proton.

The wave function describing this state is just the ground state ψ_{100} derived for the hydrogen atom, equation (4.40), but the distance should be measured from the position $\vec{r}_{\rm lp}$ of the left proton instead of from the origin:

$$\psi = \psi_{100}(|\vec{r} - \vec{r}_{\rm lp}|)$$

To shorten the notations, this wave function will be denoted by $\psi_{\rm l}$:

$$\psi_{\rm l}(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{\rm lp}|) \tag{4.76}$$

Similarly the wave function that would describe the electron as being in the ground state around the right proton will be denoted as $\psi_{\rm r}$, with

$$\psi_{\rm r}(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{\rm rp}|) \tag{4.77}$$

Key Points

- When the protons are far apart, there are two lowest energy states, $\psi_{\rm l}$ and $\psi_{\rm r}$, in which the electron is in the ground state around the left, respectively right, proton. In either case there is an hydrogen atom plus a free proton.

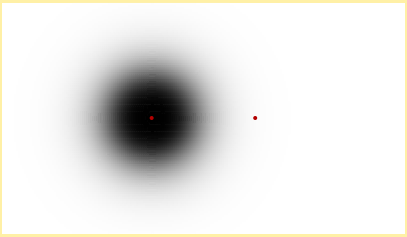


Figure 4.15: Hydrogen atom plus free proton closer together.

4.6.3 Energy when closer together

When the protons get a bit closer to each other, but still well apart, the distance r_r between the electron orbiting the left proton and the right proton decreases, as sketched in figure 4.15. The potential that the electron sees is now not just that of the left proton; the distance r_r is no longer so large that the $-e^2/4\pi\epsilon_0 r_r$ potential can be completely neglected.

However, assuming that the right proton stays sufficiently clear of the electron wave function, the distance r_r between electron and right proton can still be averaged out as being the same as the distance d between the two protons. Within that approximation, it simply adds the constant $-e^2/4\pi\epsilon_0 d$ to the Hamiltonian of the electron. And adding a constant to a Hamiltonian does not change the eigenfunction; it only changes the eigenvalue, the energy, by that constant. So the ground state ψ_l of the left proton remains a good approximation to the lowest energy wave function.

Moreover, the decrease in energy due to the electron/right proton attraction is balanced by an increase in energy of the protons by their mutual repulsion, so the total energy of the ion remains the same. In other words, the right proton is to first approximation neither attracted nor repelled by the neutral hydrogen atom on the left. To second approximation the right proton does change the wave function of the electron a bit, resulting in some attraction, but this effect will be ignored.

So far, it has been assumed that the electron is circling the left proton. But the case that the electron is circling the right proton is of course physically equivalent. In particular the energy must be exactly the same by symmetry.

Key Points

- To first approximation, there is no attraction between the free proton and the neutral hydrogen atom, even somewhat closer together.

4.6.4 States that share the electron

The approximate energy eigenfunction ψ_l that describes the electron as being around the left proton has the same energy as the eigenfunction ψ_r that describes the electron as being around the right one. Therefore any linear combination of the two,

$$\psi = a\psi_\mathrm{l} + b\psi_\mathrm{r} \tag{4.78}$$

is also an eigenfunction with the same energy. In such combinations, the electron is shared by the protons, in ways that depend on the chosen values of a and b.

Note that the constants a and b are not independent: the wave function should be normalized, $\langle\psi|\psi\rangle = 1$. Since ψ_l and ψ_r are already normalized, and assuming that a and b are real, this works out to

$$\langle a\psi_\mathrm{l} + b\psi_\mathrm{r}|a\psi_\mathrm{l} + b\psi_\mathrm{r}\rangle = a^2 + b^2 + 2ab\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle = 1 \tag{4.79}$$

As a consequence, only the ratio the coefficients a/b can be chosen freely.

A particularly interesting case is the "antisymmetric" one, $b = -a$. As figure 4.16 shows, in this state there is zero probability of finding the electron at the symmetry plane midway in between the protons. The reason is that ψ_l and ψ_r

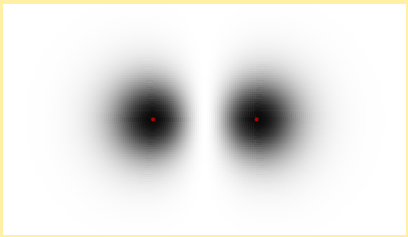


Figure 4.16: The electron being antisymmetrically shared.

are equal at the symmetry plane, making their difference zero.

This is actually a quite weird result. You combine two states, in both of which the electron has some probability of being at the symmetry plane, and in the combination the electron has *zero* probability of being there. The probability of finding the electron at any position, including the symmetry plane, in the first state is given by $|\psi_\mathrm{l}|^2$. Similarly, the probability of finding the electron in the second state is given by $|\psi_\mathrm{r}|^2$. But for the combined state nature does not do the logical thing of adding the two probabilities together to come up with $\frac{1}{2}|\psi_\mathrm{l}|^2 + \frac{1}{2}|\psi_\mathrm{r}|^2$.

Instead of adding physically *observable* probabilities, nature squares the *unobservable* wave function $a\psi_\mathrm{l} - a\psi_\mathrm{r}$ to find the new probability distribution. The squaring adds a cross term, $-2a^2\psi_\mathrm{l}\psi_\mathrm{r}$, that simply adding probabilities does not have. This term has the physical effect of preventing the electron to be at the symmetry plane, but it does not have a normal physical explanation. There is

no force repelling the electrons from the symmetry plane or anything like that. Yet it looks as if there is one in this state.

The most important combination of ψ_l and ψ_r is the "symmetric" one, $b = a$. The approximate wave function then takes the form $a(\psi_l + \psi_r)$. That can be written out fully in terms of the hydrogen ground state wave function as:

$$\boxed{\Psi \approx a\,[\psi_{100}(|\vec{r} - \vec{r}_{lp}|) + \psi_{100}(|\vec{r} - \vec{r}_{rp}|)] \qquad \psi_{100}(r) \equiv \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0}} \tag{4.80}$$

where $a_0 = 0.53$ Å is the Bohr radius and $\vec{r}$, $\vec{r}_{lp}$, and $\vec{r}_{rp}$ are again the position vectors of electron and protons. In this case, there is increased probability for the electron to be at the symmetry plane, as shown in figure 4.17.

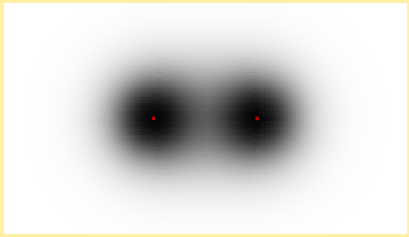


Figure 4.17: The electron being symmetrically shared.

A state in which the electron is shared is truly a case of the electron being in two different places at the same time. For if instead of sharing the electron, each proton would be given its own half electron, the expression for the Bohr radius, $a_0 = 4\pi\epsilon_0\hbar^2/m_e e^2$, shows that the eigenfunctions ψ_l and ψ_r would have to blow up in radius by a factor four. (Because of m_e and e; the second factor e is the proton charge.) The energy would then reduce by the same factor four. That is simply not what happens. You get the physics of a complete electron being present around each proton with 50% probability, not the physics of half an electron being present for sure.

Key Points

- This subsection brought home the physical weirdness arising from the mathematics of the unobservable wave function.
- In particular, within the approximations made, there exist states that all have the same ground state energy, but whose physical properties are dramatically different.
- The protons may "share the electron." In such states there is a probability of finding the electron around either proton.
- Even if the protons share the electron equally as far as the probability distribution is concerned, different physical states are still possible.

In the symmetric case that the wave functions around the protons have the same sign, there is increased probability of the electron being found in between the protons. In the antisymmetric case of opposite sign, there is decreased probability of the electron being found in between the protons.

4.6.5 Comparative energies of the states

The previous two subsections described states of the hydrogen molecular ion in which the electron is around a single proton, as well as states in which it is shared between protons. To the approximations made, all these states have the same energy. Yet, if the expectation energy of the states is more accurately examined, it turns out that increasingly large differences show up when the protons get closer together. The symmetric state has the least energy, the antisymmetric state the highest, and the states where the electron is around a single proton have something in between.

It is not that easy to see physically why the symmetric state has the lowest energy. An argument is often made that in the symmetric case, the electron has increased probability of being in between the protons, where it is most effective in pulling them together. However, actually the potential energy of the symmetric state is higher than for the other states: putting the electron midway in between the two protons means having to pull it away from one of them.

The Feynman lectures on physics, [22], argue instead that in the symmetric case, the electron is somewhat less constrained in position. According to the Heisenberg uncertainty relationship, that allows it to have less variation in momentum, hence less kinetic energy. Indeed the symmetric state does have less kinetic energy, but this is almost totally achieved at the cost of a corresponding increase in potential energy, rather than due to a larger area to move in at the same potential energy. And the kinetic energy is not really directly related to available area in any case. The argument is not incorrect, but in what sense it explains, rather than just summarizes, the answer is debatable.

Key Points

- The energies of the discussed states are not the same when examined more closely.
- The symmetric state has the lowest energy, the antisymmetric one the highest.

4.6.6 Variational approximation of the ground state

The objective of this subsection is to use the rough approximations of the previous subsections to get some very concrete data on the hydrogen molecular ion.

The idea is simple but powerful: since the true ground state is the state of lowest energy among *all* wave functions, the best among approximate wave functions is the one with the lowest energy. In the previous subsections, approximations to the ground state were discussed that took the form $a\psi_{\rm l} + b\psi_{\rm r}$, where $\psi_{\rm l}$ described the state where the electron was in the ground state around the left proton, and $\psi_{\rm r}$ where it was around the right proton. The wave function of this type with the lowest energy will produce the best possible data on the true ground state, {N.6}.

Note that all that can be changed in the approximation $a\psi_{\rm l} + b\psi_{\rm r}$ to the wave function is the ratio of the coefficients a/b, and the distance between the protons d. If the ratio a/b is fixed, a and b can be computed from it using the normalization condition (4.79), so there is no freedom to chose them individually. The basic idea is now to search through all possible values of a/b and d until you find the values that give the lowest energy.

This sort of method is called a "variational method" because at the minimum of energy, the derivatives of the energy must be zero. That in turn means that the energy does not vary with infinitesimally small changes in the parameters a/b and d.

To find the minimum energy is nothing that an engineering graduate student could not do, but it does take some effort. You cannot find the best values of a/b and d analytically; you have to have a computer find the energy at a lot of values of d and a/b and search through them to find the lowest energy. Or actually, simply having a computer print out a table of values of energy versus d for a few typical values of a/b, including $a/b = 1$ and $a/b = -1$, and looking at the print-out to see where the energy is most negative works fine too. That is what the numbers below came from.

You do want to evaluate the energy of the approximate states accurately as the expectation value. If you do not find the energy as the expectation value, the results may be less dependable. Fortunately, finding the expectation energy for the given approximate wave functions can be done exactly; the details are in derivation {D.21}.

If you actually go through the steps, your print-out should show that the minimum energy occurs when $a = b$, the symmetric state, and at a separation distance between the protons equal to about 1.3 Å. This separation distance is called the "bond length". The minimum energy is found to be about 1.8 eV *below* the energy of -13.6 eV when the protons are far apart. So it will take at least 1.8 eV to take the ground state with the protons at a distance of 1.3 Å completely apart into well separated protons. For that reason, the 1.8 eV is

called the "binding energy".

Key Points

- The best approximation to the ground state using approximate wave functions is the one with the lowest energy.
- Making such an approximation is called a variational method.
- The energy should be evaluated as the expectation value of the Hamiltonian.
- Using combinations of ψ_l and ψ_r as approximate wave functions, the approximate ground state turns out to be the one in which the electron is symmetrically shared between the protons.
- The binding energy is the energy required to take the molecule apart.
- The bond length is the distance between the nuclei.

4.6.6 Review Questions

1. The solution for the hydrogen molecular ion requires elaborate evaluations of inner product integrals and a computer evaluation of the state of lowest energy. As a much simpler example, you can try out the variational method on the one-dimensional case of a particle stuck inside a pipe, as discussed in chapter 3.5. Take the approximate wave function to be:

$$\psi = ax(\ell - x)$$

Find a from the normalization requirement that the total probability of finding the particle integrated over all possible x positions is one. Then evaluate the energy $\langle E \rangle$ as $\langle \psi | H | \psi \rangle$, where according to chapter 3.5.3, the Hamiltonian is

$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$$

Compare the ground state energy with the exact value,

$$E_1 = \hbar^2\pi^2/2m\ell^2$$

(Hints: $\int_0^\ell x(\ell - x)\,\mathrm{d}x = \ell^3/6$ and $\int_0^\ell x^2(\ell - x)^2\,\mathrm{d}x = \ell^5/30$)
Solution hione-a

4.6.7 Comparison with the exact ground state

The variational solution derived in the previous subsection is only a crude approximation of the true ground state of the hydrogen molecular ion. In particular, the assumption that the molecular wave function can be approximated using the individual atom ground states is only valid when the protons are far apart, and is inaccurate if they are 1.3 Å apart, as the solution says they are.

Yet, for such a poor wave function, the main results are surprisingly good. For one thing, it leaves no doubt that a bound state really exists. The reason is that the true ground state must always have a lower energy than any approximate one. So, the binding energy must be *at least* the 1.8 eV predicted by the approximation.

In fact, the experimental binding energy is 2.8 eV. The found approximate value is only a third less, pretty good for such a simplistic assumption for the wave function. It is really even better than that, since a fair comparison requires the absolute energies to be compared, rather than just the binding energy; the approximate solution has −15.4 eV, rather than −16.4. This high accuracy for the energy using only marginal wave functions is one of the advantages of variational methods {A.7}.

The estimated bond length is not too bad either; experimentally the protons are 1.06 Å apart instead of 1.3 Å. (The analytical solution using spheroidal coordinates mentioned earlier gives 2.79 eV and 1.06 Å, in good agreement with the experimental values. But even that solution is not really exact: the electron does not bind the nuclei together rigidly, but more like a spring force. As a result, the nuclei behave like a harmonic oscillator around their common center of gravity. Even in the ground state, they will retain some uncertainty around the 1.06 Å position of minimal energy, and a corresponding small amount of additional molecular kinetic and potential energy. The improved Born-Oppenheimer approximation of chapter 9.2.3 can be used to compute such effects.)

The qualitative properties of the approximate wave function are correct. For example, it can be seen that the exact ground state wave function must be real and positive {A.8}; the approximate wave function is real and positive too.

It can also be seen that the exact ground state must be symmetric around the symmetry plane midway between the protons, and rotationally symmetric around the line connecting the protons, {A.9}. The approximate wave function has both those properties too.

Incidentally, the fact that the ground state wave function must be real and positive is a much more solid reason that the protons must share the electron symmetrically than the physical arguments given in subsection 4.6.5, even though it is more mathematical.

Key Points

- The obtained approximate ground state is pretty good.
- The protons really share the electron symmetrically in the ground state.

Chapter 5

Multiple-Particle Systems

Abstract

So far, only wave functions for single particles have been discussed. This chapter explains how the ideas generalize to more particles. The basic idea is simple: you just keep adding more and more arguments to your wave function.

That simple idea will immediately be used to derive a solution for the hydrogen molecule. The chemical bond that keeps the molecule together is a two-electron one. It involves sharing the two electrons in a very weird way that can only be described in quantum terms.

Now it turns out that usually chemical bonds involve the sharing of two electrons like in the hydrogen molecule, not just one as in the hydrogen molecular ion. To understand the reason, simple approximate systems will be examined that have no more than two different states. It will then be seen that sharing lowers the energy due to "twilight" terms. These are usually more effective for two-electron bonds than for single electron-ones.

Before systems with more than two electrons can be discussed, a different issue must be addressed first. Electrons, as well as most other quantum particles, have intrinsic angular momentum called "spin". It is quantized much like orbital angular momentum. Electrons can either have spin angular momentum $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$ in a given direction. It is said that the electron has spin $\frac{1}{2}$. Photons can have angular momentum $\hbar$, 0, or $-\hbar$ in a given direction and have spin 1. Particles with half-integer spin like electrons are called fermions. Particles with integer spin like photons are called bosons.

For quantum mechanics there are two consequences. First, it means that spin must be added to the wave function as an uncertain quantity in addition to position. That can be done in various equivalent ways. Second, it turns out that there are requirements on the wave function depending on whether particles are bosons or fermions. In particular, wave func-

tions must stay the same if two identical bosons, say two photons, are interchanged. Wave functions must change sign when any two electrons, or any other two identical fermions, are interchanged.

This so-called antisymmetrization requirement is usually not such a big deal for two electron systems. Two electrons can satisfy the requirement by assuming a suitable combined spin state. However, for more than two electrons, the effects of the antisymmetrization requirement are dramatic. They determine the very nature of the chemical elements beyond helium. Without the antisymmetrization requirements on the electrons, chemistry would be something completely different. And therefore, so would all of nature be. Before that can be properly understood, first a better look is needed at the ways in which the symmetrization requirements can be satisfied. It is then seen that the requirement for fermions can be formulated as the so-called Pauli exclusion principle. The principle says that any number I of identical fermions must occupy I different quantum states. Fermions are excluded from entering the same quantum state.

At that point, the atoms heavier than hydrogen can be properly discussed. It can also be explained why atoms prevent each other from coming too close. Finally, the derived quantum properties of the atoms are used to describe the various types of chemical bonds.

5.1 Wave Function for Multiple Particles

While a single particle is described by a wave function $\Psi(\vec{r};t)$, a system of two particles, call them 1 and 2, is described by a wave function

$$\Psi(\vec{r}_1,\vec{r}_2;t) \tag{5.1}$$

depending on both particle positions. The value of $|\Psi(\vec{r}_1,\vec{r}_2;t)|^2\,\mathrm{d}^3\vec{r}_1\,\mathrm{d}^3\vec{r}_2$ gives the probability of simultaneously finding particle 1 within a vicinity $\mathrm{d}^3\vec{r}_1$ of $\vec{r}_1$ and particle 2 within a vicinity $\mathrm{d}^3\vec{r}_2$ of $\vec{r}_2$.

The wave function must be normalized to express that the electrons must be somewhere:

$$\langle\Psi|\Psi\rangle_6 = \iint |\Psi(\vec{r}_1,\vec{r}_2;t)|^2\,\mathrm{d}^3\vec{r}_1\mathrm{d}^3\vec{r}_2 = 1 \tag{5.2}$$

where the subscript 6 of the inner product is just a reminder that the integration is over all six scalar position coordinates of Ψ.

The underlying idea of increasing system size is "every possible combination:" allow for every possible combination of state for particle 1 and state for particle 2. For example, in one dimension, all possible x positions of particle 1 geometrically form an x_1-axis. Similarly all possible x positions of particle 2 form an x_2-axis. If every possible position x_1 is separately combined with every

possible position x_2, the result is an x_1, x_2-*plane* of possible positions of the combined system.

Similarly, in three dimensions the three-dimensional space of positions $\vec{r}_1$ combines with the three-dimensional space of positions $\vec{r}_2$ into a six-dimensional space having all possible combinations of values for $\vec{r}_1$ with all possible values for $\vec{r}_2$.

The increase in the number of dimensions when the system size increases is a major practical problem for quantum mechanics. For example, a *single* arsenic atom has 33 electrons, and each electron has 3 position coordinates. It follows that the wave function is a function of 99 scalar variables. (Not even counting the nucleus, spin, etcetera.) In a brute-force numerical solution of the wave function, maybe you could restrict each position coordinate to only ten computational values, if no very high accuracy is desired. Even then, Ψ values at 10^{99} different combined positions must be stored, requiring maybe 10^{91} Gigabytes of storage. To do a single multiplication on each of those those numbers within a few years would require a computer with a speed of 10^{82} gigaflops. No need to take any of that arsenic to be long dead before an answer is obtained. (Imagine what it would take to compute a microgram of arsenic instead of an atom.) Obviously, more clever numerical procedures are needed.

Sometimes the problem size can be reduced. In particular, the problem for a two-particle system like the proton-electron hydrogen atom can be reduced to that of a single particle using the concept of reduced mass. That is shown in addendum {A.5}.

Key Points

- To describe multiple-particle systems, just keep adding more independent variables to the wave function.
- Unfortunately, this makes many-particle problems impossible to solve by brute force.

5.1 Review Questions

1. A simple form that a six-dimensional wave function can take is a product of two three-dimensional ones, as in $\psi(\vec{r}_1, \vec{r}_2) = \psi_1(\vec{r}_1)\psi_2(\vec{r}_2)$. Show that if ψ_1 and ψ_2 are normalized, then so is ψ.

   Solution complex-a

2. Show that for a simple product wave function as in the previous question, the relative probabilities of finding particle 1 near a position $\vec{r}_a$ versus finding it near another position $\vec{r}_b$ is the same regardless where particle 2 is. (Or rather, where particle 2 is likely to be found.)

   Note: This is the reason that a simple product wave function is called "uncorrelated." For particles that interact with each other, an uncorrelated wave function is often not a good approximation. For example, two

electrons repel each other. All else being the same, the electrons would rather be at positions where the other electron is nowhere close. As a result, it really makes a difference for electron 1 where electron 2 is likely to be and vice-versa. To handle such situations, usually *sums* of product wave functions are used. However, for some cases, like for the helium atom, a single product wave function is a perfectly acceptable first approximation. Real-life electrons are crowded together around attracting nuclei and learn to live with each other.

Solution complex-b

5.2 The Hydrogen Molecule

This section uses similar approximations as for the hydrogen molecular ion of chapter 4.6 to examine the neutral H_2 hydrogen molecule. This molecule has two electrons circling two protons. It will turn out that in the ground state, the protons share the two electrons, rather than each being assigned one. This is typical of covalent bonds.

Of course, "share" is a vague term, but the discussion will show what it really means in terms of the six-dimensional electron wave function.

5.2.1 The Hamiltonian

Just like for the hydrogen molecular ion of chapter 4.6, for the neutral molecule the Born-Oppenheimer approximation will be made that the protons are at given fixed points. So the problem simplifies to just finding the wave function of the two electrons, $\Psi(\vec{r}_1,\vec{r}_2)$, where $\vec{r}_1$ and $\vec{r}_2$ are the positions of the two electrons 1 and 2. In terms of scalar arguments, the wave function can be written out further as $\Psi(x_1,y_1,z_1,x_2,y_2,z_2)$.

In the Hamiltonian, following the Newtonian analogy the kinetic and potential energy operators simply add:

$$H = -\frac{\hbar^2}{2m_{\rm e}}\left(\nabla_1^2+\nabla_2^2\right) - \frac{e^2}{4\pi\epsilon_0}\left(\frac{1}{r_{1\rm l}}+\frac{1}{r_{1\rm r}}+\frac{1}{r_{2\rm l}}+\frac{1}{r_{2\rm r}}-\frac{1}{|\vec{r}_1-\vec{r}_2|}\right) \tag{5.3}$$

In this expression, the Laplacians of the first two, kinetic energy, terms are with respect to the position coordinates of the two electrons:

$$\nabla_1^2 = \frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial y_1^2}+\frac{\partial^2}{\partial z_1^2} \qquad \nabla_2^2 = \frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial y_2^2}+\frac{\partial^2}{\partial z_z^2}.$$

The next four terms in the Hamiltonian (5.3) are the attractive potentials between the electrons and the protons, with $r_{1\rm l}$, $r_{2\rm l}$, $r_{1\rm r}$, and $r_{2\rm r}$ being the distances between electrons 1 and 2 and the left, respectively right proton. The final term represents the repulsive potential between the two electrons.

Key Points

- The Hamiltonian for the 6-dimensional electron wave function has been written down.

5.2.1 Review Questions

1. Verify that the repulsive potential between the electrons is infinitely large when the electrons are at the same position.

   Note: You might therefore think that the wave function needs to be zero at the locations in six-dimensional space where $\vec{r}_1 = \vec{r}_2$. Some authors refer to that as a "Coulomb hole." But the truth is that in quantum mechanics, electrons are smeared out due to uncertainty. That causes electron 1 to "see electron 2 at all sides", and vice-versa, and they do therefore not encounter any unusually large potential when the wave function is nonzero at $\vec{r}_1 = \vec{r}_2$. In general, it is just not worth the trouble for the electrons to stay away from the same position: that would reduce their uncertainty in position, increasing their uncertainty-demanded kinetic energy.

   Solution hmola-a

2. Note that the total kinetic energy term is simply a multiple of the six-dimensional Laplacian operator. It treats all Cartesian position coordinates exactly the same, regardless of which direction or which electron it is. Is this still the case if other particles are involved?

   Solution hmola-b

5.2.2 Initial approximation to the lowest energy state

The next step is to identify an approximate ground state for the hydrogen molecule. Following the same approach as in chapter 4.6, it will first be assumed that the protons are relatively far apart. One obvious approximate solution is then that of two neutral atoms, say the one in which electron 1 is around the left proton in its ground state and electron 2 is around the right one.

To formulate the wave function for that, the shorthand notation $\psi_{\rm l}$ will again be used for the wave function of a *single* electron that in the ground state around the left proton and $\psi_{\rm r}$ for one that is in the ground state around the right hand one:

$$\psi_{\rm l}(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{\rm lp}|) \quad \psi_{\rm r}(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{\rm rp}|)$$

where ψ_{100} is the hydrogen atom ground state (4.40), and $\vec{r}_{\rm lp}$ and $\vec{r}_{\rm rp}$ are the positions of the left and right protons.

The wave function that describes that electron 1 is in the ground state around the left proton and electron 2 around the right one will be approximated to be the product of the single electron states:

$$\psi(\vec{r}_1, \vec{r}_2) = \psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2)$$

Taking the combined wave function as a product of single electron states is really equivalent to an assumption that the two electrons are independent. Indeed, for the product state, the probability of finding electron 1 at position $\vec{r}_1$ and electron 2 at $\vec{r}_2$ is:

$$|\psi_{\rm l}(\vec{r}_1)|^2\,{\rm d}^3\vec{r}_1 \times |\psi_{\rm r}(\vec{r}_2)|^2\,{\rm d}^3\vec{r}_2$$

or in words:

[probability of finding 1 at $\vec{r}_1$ unaffected by where 2 is]
$\times$ [probability of finding 2 at $\vec{r}_2$ unaffected by where 1 is]

Such product probabilities are characteristic of statistically independent quantities. As a simple example, the chances of getting a three in the first throw of a die and a five in the second throw are $\frac{1}{6} \times \frac{1}{6}$ or 1 in 36. Throwing the three does not affect the chances of getting a five in the second throw.

Key Points

- When the protons are well apart, an approximate ground state is that of two neutral atoms.
- Single electron wave functions for that case are $\psi_{\rm l}$ and $\psi_{\rm r}$.
- The complete wave function for that case is $\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2)$, assuming that electron 1 is around the left proton and electron 2 around the right one.

5.2.2 Review Questions

1. If electron 2 does not affect where electron 1 is likely to be, how would a grey-scale picture of the probability of finding electron 1 look?
   Solution hmolb-a
2. When the protons are close to each other, the electrons do affect each other, and the wave function above is no longer valid. But suppose you were given the true wave function, and you were once again asked to draw the blob showing the probability of finding electron 1 (using a plotting package, say). What would the big problem be?
   Solution hmolb-b

5.2.3 The probability density

For multiple-particle systems like the electrons of the hydrogen molecule, showing the magnitude of the wave function as grey tones no longer works since it is a function in six-dimensional space. You cannot visualize six-dimensional space. However, at every spatial position $\vec{r}$ in normal space, you can instead show the

"probability density" $n(\vec{r})$, which is the probability per unit volume of finding *either* electron in a vicinity $\mathrm{d}^3\vec{r}$ of the point. This probability is found as

$$n(\vec{r}) = \int |\Psi(\vec{r},\vec{r}_2)|^2\,\mathrm{d}^3\vec{r}_2 + \int |\Psi(\vec{r}_1,\vec{r})|^2\,\mathrm{d}^3\vec{r}_1 \tag{5.4}$$

since the first integral gives the probability of finding electron 1 at $\vec{r}$ regardless of where electron 2 is, (i.e. integrated over all possible positions for electron 2), and the second gives the probability of finding 2 at $\vec{r}$ regardless of where 1 is. Since $\mathrm{d}^3\vec{r}$ is vanishingly small, the chances of finding both particles in it at the same time are zero.

The probability density $n(\vec{r})$ for state $\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2)$ with electron 1 around the left proton and electron 2 around the right one is shown in figure 5.1. Of course the probability density for the state $\psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)$ with the electrons exchanged would look exactly the same.

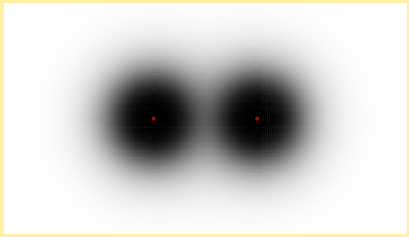


Figure 5.1: State with two neutral atoms.

Key Points

- The probability density is the probability per unit volume of finding an electron, whichever one, near a given point.

5.2.3 Review Questions

1. Suppose, given the wave function $\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2)$, that you found an electron near the left proton. What electron would it probably be? Suppose you found an electron at the point halfway in between the protons. What electron would that likely be?
   Solution hmolc-a

5.2.4 States that share the electrons

This section will examine the states where the protons share the two electrons.

The first thing is to shorten the notations a bit. So, the state $\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2)$ which describes that electron 1 is around the left proton and electron 2 around the right one will be indicated by $\psi_{\rm l}\psi_{\rm r}$, using the convention that the first factor

refers to electron 1 and the second to electron 2. In this convention, the state where electron 1 is around the right proton and electron 2 around the left one is $\psi_r\psi_l$, shorthand for $\psi_r(\vec{r}_1)\psi_l(\vec{r}_2)$. It is of course physically the same thing as $\psi_l\psi_r$; the two electrons are identical.

The "every possible combination" idea of combining every possible state for electron 1 with every possible state for electron 2 would suggest that the states $\psi_l\psi_l$ and $\psi_r\psi_r$ should also be included. But these states have the electrons around the same proton, and that is not going to be energetically favorable due to the mutual repulsion of the electrons. So they are not useful for finding a simple approximate ground state of lowest energy.

States where the electrons are no longer assigned to a particular proton can be found as linear combinations of $\psi_l\psi_r$ and $\psi_r\psi_l$:

$$\psi = a\psi_l\psi_r + b\psi_r\psi_l \tag{5.5}$$

In such a combination each electron has a probability of being found about either proton, but wherever it is found, the other electron will be around the other proton.

The eigenfunction must be normalized, which noting that ψ_l and ψ_r are real and normalized produces

$$\langle\psi|\psi\rangle_6 = \langle a\psi_l\psi_r + b\psi_r\psi_l|a\psi_l\psi_r + b\psi_r\psi_l\rangle = a^2 + b^2 + 2ab\langle\psi_l|\psi_r\rangle^2 = 1 \tag{5.6}$$

assuming that a and b are real. As a result, only the ratio a/b can be chosen freely. The probability density of the combination can be found to be:

$$n = \psi_l^2 + \psi_r^2 + 2ab\langle\psi_l|\psi_r\rangle\left\{2\psi_l\psi_r - \langle\psi_l|\psi_r\rangle(\psi_l^2 + \psi_r^2)\right\} \tag{5.7}$$

The most important combination state is the one with $b = a$:

$$\psi(\vec{r}_1, \vec{r}_2) = a\left[\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)\right] \tag{5.8}$$

This state is called "symmetric with respect to exchanging electron 1 with electron 2," or more precisely, with respect to replacing $\vec{r}_1$ by $\vec{r}_2$ and vice-versa. Such an exchange does not change this wave function at all. If you change $\vec{r}_1$ into $\vec{r}_2$ and vice-versa, you still end up with the same wave function. In terms of the hydrogen ground state wave function, it may be written out fully as

$$\boxed{\Psi \approx a\left[\psi_{100}(|\vec{r}_1 - \vec{r}_{\rm lp}|)\psi_{100}(|\vec{r}_2 - \vec{r}_{\rm rp}|) + \psi_{100}(|\vec{r}_1 - \vec{r}_{\rm rp}|)\psi_{100}(|\vec{r}_2 - \vec{r}_{\rm lp}|)\right]} \tag{5.9}$$

with $\psi_{100}(r) \equiv e^{-r/a_0}/\sqrt{\pi a_0^3}$, where $a_0 = 0.53$ Å is the Bohr radius, and $\vec{r}_1$, $\vec{r}_2$, $\vec{r}_{\rm lp}$, and $\vec{r}_{\rm rp}$ are again the position vectors of the electrons and protons.

The probability density of this wave function looks like figure 5.2. It has increased likelihood for electrons to be found in between the protons, compared to figure 5.1 in which each proton had its own electron.

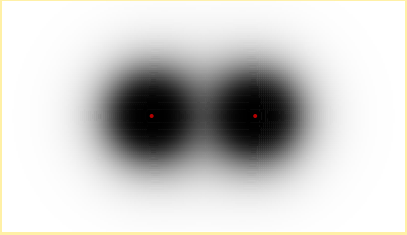


Figure 5.2: Symmetric sharing of the electrons.

The state with $b = -a$,

$$\psi(\vec{r}_1, \vec{r}_2) = a\left[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) - \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)\right] \tag{5.10}$$

is called "antisymmetric" with respect to exchanging electron 1 with electron 2: swapping $\vec{r}_1$ and $\vec{r}_2$ changes the sign of wave function, but leaves it further unchanged. As seen in figure 5.3, the antisymmetric state has decreased likelihood for electrons to be found in between the protons.

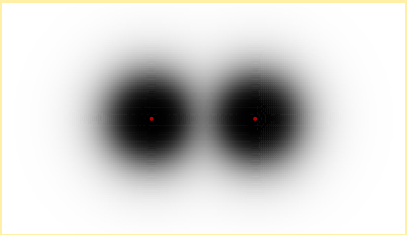


Figure 5.3: Antisymmetric sharing of the electrons.

Key Points

- In state $\psi_{\rm l}\psi_{\rm r}$, the electron numbered 1 is around the left proton and 2 around the right one.
- In state $\psi_{\rm r}\psi_{\rm l}$, the electron numbered 1 is around the right proton and 2 around the left one.
- In the symmetric state $a(\psi_{\rm l}\psi_{\rm r} + \psi_{\rm r}\psi_{\rm l})$ the protons share the electrons equally; each electron has an equal chance of being found around either proton. In this state there is increased probability of finding an electron somewhere in between the protons.
- In the antisymmetric state $a(\psi_{\rm l}\psi_{\rm r} - \psi_{\rm r}\psi_{\rm l})$ the protons also share the electrons equally; each electron has again an equal chance of being found around either proton. But in this state there is decreased probability of finding an electron somewhere in between the protons.

So, like for the molecular ion, at large proton separations the weird trick of shuffling unobservable wave functions around does again produce different physical states with pretty much the same energy.

5.2.4 Review Questions

1. Obviously, the visual difference between the various states is minor. It may even seem counter-intuitive that there is any difference at all: the states $\psi_l\psi_r$ and $\psi_r\psi_l$ are exactly the same physically, with one electron around each proton. So why would their combinations be any different?

   The quantum difference would be much more clear if you could see the full six-dimensional wave function, but visualizing six-dimensional space just does not work. However, if you restrict yourself to only looking on the z-axis through the nuclei, you get a drawable z_1, z_2-plane describing near what *axial* combinations of positions you are most likely to find the two electrons. In other words: what would be the chances of finding electron 1 near some axial position z_1 and electron 2 at the same time near some other axial position z_2?

   Try to guess these probabilities in the z_1, z_2-plane as grey tones, (darker if more likely), and then compare with the answer.

   Solution hmold-a
2. Based on the previous question, how would you think the probability density $n(z)$ would look on the axis through the nuclei, again ignoring the existence of positions beyond the axis?

   Solution hmold-b

5.2.5 Variational approximation of the ground state

The purpose of this section is to find an approximation to the ground state of the hydrogen molecule using the rough approximation of the wave function described in the previous subsections.

Like for the hydrogen molecular ion of chapter 4.6.6, the idea is that since the true ground state is the state of lowest energy among *all* wave functions, the best among approximate wave functions is the one with the lowest energy. The approximate wave functions are here of the form $a\psi_l\psi_r + b\psi_r\psi_l$; in these the protons share the electrons, but in such a way that when one electron is around the left proton, the other is around the right one, and vice-versa.

A computer program is again needed to print out the expectation value of the energy for various values of the ratio of coefficients a/b and proton-proton distance d. And worse, the expectation value of energy for given a/b and d is a six-dimensional integral, and parts of it cannot be done analytically; numerical integration must be used. That makes it a much more messy problem, {D.23}.

You might just want to take it on faith that the binding energy, at the state of lowest energy found, turns out to be 3.2 eV, at a proton to proton spacing of 0.87 Å, and that it occurs for the symmetric state $a = b$.

Key Points

- An approximate ground state can be found for the hydrogen molecule using a variational method much like that for the molecular ion.

5.2.6 Comparison with the exact ground state

The solution for the ground state of the hydrogen molecule obtained in the previous subsection is, like the one for the molecular ion, pretty good. The approximate binding energy, 3.2 eV, is not too much different from the experimental value of 4.52 eV. Similarly, the bond length of 0.87 Å is not too far from the experimental value of 0.74 Å.

Qualitatively, the exact ground state wave function is real, positive and symmetric with respect to reflection around the symmetry plane and to rotations around the line connecting the protons, and so is the approximate one. The reasons for these properties are similar as for the molecular ion; {A.8,A.9}.

One very important new symmetry for the neutral molecule is the effect of exchanging the electrons, replacing $\vec{r}_1$ by $\vec{r}_2$ and vice-versa. The approximate wave function is symmetric (unchanged) under such an exchange, and so is the exact wave function. To understand why, note that the operation of exchanging the electrons commutes with the Hamiltonian, (exchanging identical electrons physically does not do anything). So energy eigenfunctions can be taken to be also eigenfunctions of the "exchange operator." Furthermore, the exchange operator is a Hermitian one, (taking it to the other side in inner products is equivalent to a simple name change of integration variables,) so it has real eigenvalues. And more specifically, the eigenvalues can only be plus or minus one, since swapping electrons does not change the magnitude of the wave function. So the energy eigenfunctions, including the ground state, must be symmetric under electron exchange (eigenvalue one), or antisymmetric (eigenvalue minus one.) Since the ground state must be everywhere positive, (or more precisely, of a single sign), a sign change due to swapping electrons is not possible. So only the symmetric possibility exists for the ground state.

One issue that does not occur for the molecular ion, but only for the neutral molecule is the mutual repulsion between the two electrons. This repulsion is reduced when the electron clouds start to merge, compared to what it would be if the clouds were more compact. (A similar effect is that the gravity force of the earth decreases when you go down below the surface. To be sure, the potential energy keeps going down, or up for electron clouds, but not as much as it would otherwise. Compare figure 13.7.) Since the nuclei are compact, it gives an advantage to nucleus-electron attraction over electron-electron repulsion. This increases the binding energy significantly; in the approximate model from about 1.8 eV to 3.2 eV. It also allows the protons to approach more closely; {D.23}.

The question has been asked whether there should not be an "activation energy" involved in creating the hydrogen molecule from the hydrogen atoms. The answer is no, hydrogen atoms are radicals, not stable molecules that need to be taken apart before recombining. In fact, the hydrogen atoms attract each other even at large distances due to Van der Waals attraction, chapter 10.1, an effect lost in the approximate wave functions used in this section. But hydrogen atoms that fly into each other also have enough energy to fly apart again; some of the excess energy must be absorbed elsewhere to form a stable molecule. According to web sources, hydrogen molecule formation in the universe is believed to typically occur on dust specks.

Key Points

- The approximate ground state is pretty good, considering its simplicity.

5.3 Two-State Systems

Two-state systems are systems in which only two quantum states are of importance. That makes such systems the simplest nontrivial quantum systems. A lot of qualitative understanding can be obtained from them. Among others, this section will shed some light on the reason why chemical bonds tend to involve pairs of electrons.

As seen in chapter 4.6, the protons in the H_2^+ hydrogen molecular ion are held together by a single shared electron. However, in the H_2 neutral hydrogen molecule of the previous section, they are held together by a shared pair of electrons. In both cases a stable bond was formed. So why are chemical bonds involving a single electron relatively rare, while bonds involving pairs of shared electrons are common?

The unifying concept relating the two bonds is that of two-state systems. Such systems involve two intuitive basic states ψ_1 and ψ_2.

For the hydrogen molecular ion, one state, $\psi_1 = \psi_{\rm l}$, described that the electron was in the ground state around the left proton. A physically equivalent state, $\psi_2 = \psi_{\rm r}$, had the electron in the ground state around the right proton. For the hydrogen molecule, $\psi_1 = \psi_{\rm l}\psi_{\rm r}$ had electron 1 around the left proton and electron 2 around the right one. The other state $\psi_2 = \psi_{\rm r}\psi_{\rm l}$ was physically the same, but it had the electrons reversed.

There are many other physical situations that may be described as two state systems. Covalent chemical bonds involving atoms other than hydrogen would be an obvious example. Just substitute a positive ion for one or both protons.

As another example of a two-state system, consider the C_6H_6 "benzene molecular ring." This molecule consists of a hexagon of 6 carbon atoms that

are held together by 9 covalent bonds. The logical way that 9 bonds can be arranged between the atoms of a 6 atom ring is to make every second bond a double one. However, that still leaves two possibilities; the locations of the single and double bonds can be swapped. So there are once again two different but equivalent states ψ_1 and ψ_2.

The NH_3 "ammonia molecule" consists of an nitrogen atom bonded to three hydrogen atoms. By symmetry, the logical place for the nitrogen atom to sit would surely be in the center of the triangle formed by the three hydrogen atoms. But it does not sit there. If it was in the center of the triangle, the angles between the hydrogen atoms, measured from the nitrogen nucleus, should be 120° each. However, as discussed later in chapter 5.11.3, valence bond theory requires that the angles should be about 90°, not 120°. (The actual angles are about 108° because of reasons similar to those for water as discussed in chapter 5.11.3.) The key point here is that the nitrogen must sit to the side of the triangle, and there are two sides, producing once again two different but equivalent physical states ψ_1 and ψ_2.

In each case described above, there are two intuitive physical states ψ_1 and ψ_2. The peculiarities of the quantum mechanics of two-state systems arise from states that are combinations of these two states, as in

$$\psi = c_1\psi_1 + c_2\psi_2$$

Note that according to the ideas of quantum mechanics, the square magnitude of the first coefficient of the combined state, $|c_1|^2$, represents the probability of being in state ψ_1 and $|c_2|^2$ the probability of being in state ψ_2. Of course, the total probability of being in one of the states should be one:

$$|c_1|^2 + |c_2|^2 = 1$$

(This is only true if the ψ_1 and ψ_2 states are orthonormal. In the hydrogen molecule cases, orthonormalizing the basic states would change them a bit, but their physical nature would remain much the same, especially if the protons are not too close.)

The key question is now what combination of states has the lowest energy. That will be the ground state ψ_{gs} of the two-state system. The expectation value of energy is

$$\langle E \rangle = \langle c_1\psi_1 + c_2\psi_2 | H | c_1\psi_1 + c_2\psi_2 \rangle$$

This can be multiplied out, taking into account that numerical factors come out of the left of an inner product as complex conjugates. The result is

$$\langle E \rangle = |c_1|^2\langle E_1 \rangle + c_1^* c_2 H_{12} + c_2^* c_1 H_{21} + |c_2|^2\langle E_2 \rangle$$

using the shorthand notation

$$\langle E_1 \rangle = \langle \psi_1 | H\psi_1 \rangle, \quad H_{12} = \langle \psi_1 | H\psi_2 \rangle, \quad H_{21} = \langle \psi_2 | H\psi_1 \rangle, \quad \langle E_2 \rangle = \langle \psi_2 | H\psi_2 \rangle$$

Note that $\langle E_1 \rangle$ and $\langle E_2 \rangle$ are real, (2.16). They are the expectation energies of the states ψ_1 and ψ_2. The states will be ordered so that $\langle E_1 \rangle$ is less or equal to $\langle E_2 \rangle$. (In all the examples mentioned so far, $\langle E_1 \rangle$ and $\langle E_2 \rangle$ are equal because the two states are physically equivalent.) Normally, H_{12} and H_{21} are not real but complex conjugates, (2.16). However, you can always change the definition of, say, ψ_1 by a complex factor of magnitude one to make H_{12} equal to a real and negative number, and then H_{21} will be that same negative number.

The above expression for the expectation energy consists of two kinds of terms, which will be called:

$$\text{the averaged energy:} \quad |c_1|^2 \langle E_1 \rangle + |c_2|^2 \langle E_2 \rangle \tag{5.11}$$

$$\text{the twilight terms:} \quad (c_1^* c_2 + c_2^* c_1) H_{12} \tag{5.12}$$

Each of those contributions will be discussed in turn.

The averaged energy is the energy that you would intuitively expect the combined wave function to have. It is a straightforward sum of the expectation energies of the two component states ψ_1 and ψ_2 times the probabilities of being in those states. In particular, in the important case that the two states have the same energy, the averaged energy is that energy. What is more logical than that any mixture of two states with the same energy would have that energy too?

But the twilight terms throw a monkey wrench in this simplistic thinking. It can be seen that they will always make the ground state energy $E_{\rm gs}$ lower than the lowest energy of the component states $\langle E_1 \rangle$. (To see that, just take c_1 and c_2 positive real numbers and c_2 small enough that c_2^2 can be neglected.) This lowering of the energy below the lowest component state comes out of the mathematics of combining states; absolutely no new physical forces are added to produce it. But if you try to describe it in terms of classical physics, it really looks like a mysterious new "twilight force" is in operation here. It is no new force; it is the weird mathematics of quantum mechanics.

So, what *are* these twilight terms physically? If you mean, what are they in terms of *classical* physics, there is simply no answer. But if you mean, what are they in terms of normal language, rather than formulae, it is easy. Just have another look at the definition of the twilight terms; they are a measure of the inner product $\langle \psi_1 | H \psi_2 \rangle$. That is the energy you would get if nature was in state ψ_1 if nature was in state ψ_2. On quantum scales, nature can get really, really ethereal, where it moves beyond being describable by classical physics, and the result is very concrete, but weird, interactions. For, at these scales twilight is real, and classical physics is not.

For the twilight terms to be nonzero, there must be a region where the two states overlap, i.e. there must be a region where both ψ_1 and ψ_2 are nonzero. In the simplest case of the hydrogen molecular ion, if the atoms are far apart, the left and right wave functions do not overlap and the twilight terms will be

zero. For the hydrogen molecule, it gets a bit less intuitive, since the overlap should really be visualized in the six-dimensional space of those functions. But still, the terms are zero when the atoms are far apart.

The twilight terms are customarily referred to as "exchange terms," but everybody seems to have a different idea of what that is supposed to mean. The reason may be that these terms pop up all over the place, in all sorts of very different settings. This book prefers to call them twilight terms, since that most clearly expresses what they really are. Nature is in a twilight zone of ambiguity.

The lowering of the energy by the twilight terms produces more stable chemical bonds than you would expect. Typically, the effect of the terms is greatest if the two basic states ψ_1 and ψ_2 are physically equivalent, like for the mentioned examples. Then the two states have the same expectation energy, call it $\langle E \rangle_{1,2}$. For such symmetric systems, the ground state will occur for an *equal* mixture of the two states, $c_1 = c_2 = \sqrt{\frac{1}{2}}$, because then the twilight terms are most negative. (Complex coefficients do not really make a physical difference, so c_1 and c_2 can be assumed to be real numbers for convenience.) In the ground state, the lowest energy is then an amount $|H_{12}|$ below the energy of the component states:

$$\boxed{\text{Symmetric 2-state systems:} \quad \psi_{\rm gs} = \frac{\psi_1 + \psi_2}{\sqrt{2}} \quad E_{\rm gs} = \langle E \rangle_{1,2} - |H_{12}|} \tag{5.13}$$

On the other hand, if the lower energy state ψ_1 has significantly less energy than state ψ_2, then the minimum energy will occur near the lower energy state. That means that $|c_1| \approx 1$ and $|c_2| \approx 0$. (This assumes that the twilight terms are not big enough to dominate the energy.) In that case $c_1 c_2 \approx 0$ in the twilight terms (5.12), which pretty much takes the terms out of the picture completely.

This happens for the single-electron bond of the hydrogen molecular ion if the second proton is replaced by another ion, say a lithium ion. The energy in state ψ_1, where the electron is around the proton, will now be significantly less than that of state ψ_2, where it is around the lithium ion. For such asymmetrical single-electron bonds, the twilight terms are not likely to help much in forging a strong bond. While it turns out that the LiH^+ ion is stable, the binding energy is only 0.14 eV or so, compared to 2.8 eV for the H_2^+ ion. Also, the LiH^+ bond seems to be best described as a Van der Waals attraction, rather than a true chemical bond.

In contrast, for the two-electron bond of the neutral hydrogen molecule, if the second proton is replaced by a lithium ion, states ψ_1 and ψ_2 will still be the same: both states will have one electron around the proton and one around the lithium ion. The two states do have the electrons reversed, but the electrons are identical. Thus the twilight terms are still likely to be effective. Indeed neutral

LiH lithium hydride exists as a stable molecule with a binding energy of about 2.5 eV at low pressures.

(It should be noted that the LiH bond is very ionic, with the "shared" electrons mostly at the hydrogen side, so the actual ground state is quite different from the covalent hydrogen model. But the model should be better when the nuclei are farther apart, so the analysis can at least justify the existence of a significant bond.)

For the ammonia molecule, the two states ψ_1 and ψ_2 differ only in the side of the hydrogen triangle that the nitrogen atom is at. Since these two states are physically equivalent, there is again a significant lowering of the energy $E_{\rm gs}$ for the symmetric combination $c_1 = c_2$. Similarly, there is a significant raising of the energy $E_{\rm as}$ for the antisymmetric combination $c_1 = -c_2$. Transitions between these two energy states produce photons of a single energy in the microwave range. It allows a maser (microwave-range laser) to be constructed. The first maser was in fact an ammonia one. It gave rise to the subsequent development of optical-range versions. These were initially called "optical masers," but are now known as "lasers." Masers are important for providing a single frequency reference, like in some atomic clocks. See chapter 7.7 for the operating principle of masers and lasers.

The ammonia molecule may well be the best example of how weird these twilight effects are. Consider, there are two common-sense states in which the nitrogen is at one side of the hydrogen triangle. What physical reason could there possibly be that there is a state of lower energy in which the atom is at both sides at the same time with a 50/50 probability? Before you answer that, recall that it only works if you do the 50/50 case right. If you do it wrong, you end up raising the energy. And the only way to figure out whether you do it right is to look at the behavior of the sign of a physically unobservable wave function.

It may finally be noted that in the context of chemical bonds, the raised-energy antisymmetric state is often called an "antibonding" state.

Key Points

- In quantum mechanics, the energy of different but physically equivalent states can be lowered by mixing them together.
- This lowering of energy does not come from new physical forces, but from the weird mathematics of the wave function.
- The effect tends to be much less when the original states are physically very different.
- One important place where states are indeed physically the same is in chemical bonds involving pairs of electrons. Here the equivalent states merely have the identical electrons interchanged.

5.3 Review Questions

1. The effectiveness of mixing states was already shown by the hydrogen molecule and molecular ion examples. But the generalized story above restricts the "basis" states to be orthogonal, and the states used in the hydrogen examples were not.

   Show that if ψ_1 and ψ_2 are not orthogonal states, but are normalized and produce a real and positive value for $\langle\psi_1|\psi_2\rangle$, like in the hydrogen examples, then orthogonal states can be found in the form

   $$\bar{\psi}_1 = \alpha\,(\psi_1 - \varepsilon\psi_2) \qquad \bar{\psi}_2 = \alpha\,(\psi_2 - \varepsilon\psi_1)\,.$$

   For normalized ψ_1 and ψ_2 the Cauchy-Schwartz inequality implies that $\langle\psi_1|\psi_2\rangle$ will be less than one. If the states do not overlap much, it will be much less than one and ε will be small.

   (If ψ_1 and ψ_2 do not meet the stated requirements, you can always redefine them by factors $ae^{\mathrm{i}c}$ and $be^{-\mathrm{i}c}$, with a, b, and c real, to get states that do.)

   Solution 2state-a

2. Show that it does not have an effect on the solution whether or not the basic states ψ_1 and ψ_2 are normalized, like in the previous question, before the state of lowest energy is found.

   This requires no detailed analysis; just check that the same solution can be described using the nonorthogonal and orthogonal basis states. It is however an important observation for various numerical solution procedures: your set of basis functions can be cleaned up and simplified without affecting the solution you get.

   Solution 2state-b

5.4 Spin

At this stage, it becomes necessary to look somewhat closer at the various particles involved in quantum mechanics themselves. The analysis so far already used the fact that particles have a property called mass, a quantity that special relativity has identified as being an internal amount of energy. It turns out that in addition particles have a fixed amount of "build-in" angular momentum, called "spin." Spin reflects itself, for example, in how a charged particle such as an electron interacts with a magnetic field.

To keep it apart from spin, from now on the angular momentum of a particle due to its motion will on be referred to as "orbital" angular momentum. As was discussed in chapter 4.2, the square orbital angular momentum of a particle is given by

$$L^2 = l(l+1)\hbar^2$$

where the azimuthal quantum number l is a nonnegative integer.

The square spin angular momentum of a particle is given by a similar expression:

$$S^2 = s(s+1)\hbar^2 \tag{5.14}$$

but the "spin s" is a fixed number for a given type of particle. And while l can only be an integer, the spin s can be any multiple of one half.

Particles with half integer spin are called "fermions." For example, electrons, protons, and neutrons all three have spin $s = \frac{1}{2}$ and are fermions.

Particles with integer spin are called "bosons." For example, photons have spin $s = 1$. The π-mesons have spin $s = 0$ and gravitons, unobserved at the time of writing, should have spin $s = 2$.

The spin angular momentum in an arbitrarily chosen z-direction is

$$S_z = m\hbar \tag{5.15}$$

the same formula as for orbital angular momentum, and the values of m range again from $-s$ to $+s$ in integer steps. For example, photons can have spin in a given direction that is $\hbar$, 0, or $-\hbar$. (The photon, a relativistic particle with zero rest mass, has only two spin states along the direction of propagation; the zero value does not occur in this case. But photons radiated by atoms can still come off with zero angular momentum in a direction normal to the direction of propagation. A derivation is in addendum {A.21.6} and {A.21.7}.)

The common particles, (electrons, protons, neutrons), can only have spin angular momentum $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$ in any given direction. The positive sign state is called "spin up", the negative one "spin down".

It may be noted that the proton and neutron are not elementary particles, but are baryons, consisting of three quarks. Similarly, mesons consist of a quark and an anti-quark. Quarks have spin $1/2$, which allows baryons to have spin $3/2$ or $1/2$. (It is not self-evident, but spin values can be additive or subtractive within the confines of their discrete allowable values; see chapter 12.) The same way, mesons can have spin 1 or 0.

Spin states are commonly shown in "ket notation" as $|s\, m\rangle$. For example, the spin-up state for an electron is indicated by $|1/2\ 1/2\rangle$ and the spin-down state as $|1/2\ -1/2\rangle$. More informally, $\uparrow$ and $\downarrow$ are often used.

Key Points

- Most particles have internal angular momentum called spin.
- The square spin angular momentum and its quantum number s are always the same for a given particle.
- Electrons, protons and neutrons all have spin $1/2$. Their spin angular momentum in a given direction is either $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$.
- Photons have spin one. Possible values for their angular momentum in a given direction are $\hbar$, zero, or $-\hbar$, though zero does not occur in the direction of propagation.

- Particles with integer spin, like photons, are called bosons. Particles with half-integer spin, like electrons, protons, and neutrons, are called fermions.
- The spin-up state of a spin one-half particle like an electron is usually indicated by $|\frac{1}{2}\ \frac{1}{2}\rangle$ or $\uparrow$. Similarly, the spin-down state is indicated by $|\frac{1}{2}\ -\frac{1}{2}\rangle$ or $\downarrow$.

5.4 Review Questions

1. Delta particles have spin $\frac{3}{2}$. What values can their spin angular momentum in a given direction have?
   Solution spin-a
2. Delta particles have spin $\frac{3}{2}$. What is their square spin angular momentum?
   Solution spin-b

5.5 Multiple-Particle Systems Including Spin

Spin will turn out to have a major effect on how quantum particles behave. Therefore, quantum mechanics as discussed so far must be generalized to include spin. Just like there is a probability that a particle is at some position $\vec{r}$, there is the additional probability that it has spin angular momentum S_z in an arbitrarily chosen z-direction and this must be included in the wave function. This section discusses how.

5.5.1 Wave function for a single particle with spin

The first question is how spin should be included in the wave function of a single particle. If spin is ignored, a single particle has a wave function $\Psi(\vec{r};t)$, depending on position $\vec{r}$ and on time t. Now, the spin S_z is just some other scalar variable that describes the particle, in that respect no different from say the x-position of the particle. The "every possible combination" idea of allowing every possible combination of states to have its own probability indicates that S_z needs to be added to the list of variables. So the complete wave function Ψ of the particle can be written out fully as:

$$\boxed{\Psi \equiv \Psi(\vec{r}, S_z; t)} \tag{5.16}$$

The value of $|\Psi(\vec{r}, S_z; t)|^2\, \mathrm{d}^3\vec{r}$ gives the probability of finding the particle within a vicinity $\mathrm{d}^3\vec{r}$ of $\vec{r}$ and with spin angular momentum in the z-direction S_z.

But note that there is a big difference between the spin "coordinate" and the position coordinates: while the position variables can take on any value, the values of S_z are highly limited. In particular, for the electron, proton, and

neutron, S_z can only be $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$, nothing else. You do not really have a full S_z "axis", just two points.

As a result, there are other meaningful ways of writing the wave function. The full wave function $\Psi(\vec{r}, S_z; t)$ can be thought of as consisting of two parts Ψ_+ and Ψ_- that only depend on position:

$$\boxed{\Psi_+(\vec{r};t) \equiv \Psi(\vec{r}, \tfrac{1}{2}\hbar; t) \quad \text{and} \quad \Psi_-(\vec{r};t) \equiv \Psi(\vec{r}, -\tfrac{1}{2}\hbar; t)} \tag{5.17}$$

These two parts can in turn be thought of as being the components of a two-dimensional vector that only depends on position:

$$\vec{\Psi}(\vec{r};t) \equiv \begin{pmatrix} \Psi_+(\vec{r};t) \\ \Psi_-(\vec{r};t) \end{pmatrix}$$

Remarkably, Dirac found that the wave function for particles like electrons *has* to be a vector, if it is assumed that the relativistic equations take a guessed simple and beautiful form, like the Schrödinger and all other basic equations of physics are simple and beautiful. Just like relativity reveals that particles should have build-in energy, it also reveals that particles like electrons have build-in angular momentum. A description of the Dirac equation is in chapter 12.12 if you are curious.

The two-dimensional vector is called a "spinor" to indicate that its components do not change like those of ordinary physical vectors when the coordinate system is rotated. (How they do change is of no importance here, but will eventually be described in derivation {D.68}.) The spinor can also be written in terms of a magnitude times a unit vector:

$$\vec{\Psi}(\vec{r};t) = \Psi_m(\vec{r};t) \begin{pmatrix} \chi_1(\vec{r};t) \\ \chi_2(\vec{r};t) \end{pmatrix}.$$

This book will just use the scalar wave function $\Psi(\vec{r}, S_z; t)$; not a vector one. But it is often convenient to write the scalar wave function in a form equivalent to the vector one:

$$\Psi(\vec{r}, S_z; t) = \Psi_+(\vec{r};t)\uparrow(S_z) + \Psi_-(\vec{r};t)\downarrow(S_z). \tag{5.18}$$

The square magnitude of function Ψ_+ gives the probability of finding the particle near a position with spin-up. That of Ψ_- gives the probability of finding it with spin-down. The "spin-up" function $\uparrow(S_z)$ and the "spin-down" function $\downarrow(S_z)$ are in some sense the equivalent of the unit vectors $\hat{\imath}$ and $\hat{\jmath}$ in normal vector analysis; they have by definition the following values:

$$\boxed{\uparrow(\tfrac{1}{2}\hbar) = 1 \quad \uparrow(-\tfrac{1}{2}\hbar) = 0 \qquad\qquad \downarrow(\tfrac{1}{2}\hbar) = 0 \quad \downarrow(-\tfrac{1}{2}\hbar) = 1.}$$

The function arguments will usually be left away for conciseness, so that

$$\boxed{\Psi = \Psi_+\uparrow + \Psi_-\downarrow}$$

is the way the wave function of, say, an electron will normally be written out.

Key Points

- Spin must be included as an independent variable in the wave function of a particle with spin.
- Usually, the wave function $\Psi(\vec{r}, S_z; t)$ of a single particle with spin $1/2$ will be written as
$$\Psi = \Psi_+\uparrow + \Psi_-\downarrow$$
where $\Psi_+(\vec{r}; t)$ determines the probability of finding the particle near a given location $\vec{r}$ with spin up, and $\Psi_-(\vec{r}; t)$ the one for finding it spin down.
- The functions $\uparrow(S_z)$ and $\downarrow(S_z)$ have the values
$$\uparrow(\tfrac{1}{2}\hbar) = 1 \quad \uparrow(-\tfrac{1}{2}\hbar) = 0 \qquad\qquad \downarrow(\tfrac{1}{2}\hbar) = 0 \quad \downarrow(-\tfrac{1}{2}\hbar) = 1$$
and represent the pure spin-up, respectively spin-down states.

5.5.1 Review Questions

1. What is the normalization requirement of the wave function of a spin $1/2$ particle in terms of Ψ_+ and Ψ_-?
   Solution complexsa-a

5.5.2 Inner products including spin

Inner products are important: they are needed for finding normalization factors, expectation values, uncertainty, approximate ground states, etcetera. The additional spin coordinates add a new twist, since there is no way to integrate over the few discrete points on the spin "axis". Instead, you must sum over these points.

As an example, the inner product of two arbitrary electron wave functions $\Psi_1(\vec{r}, S_z; t)$ and $\Psi_2(\vec{r}, S_z; t)$ is

$$\langle\Psi_1|\Psi_2\rangle = \sum_{S_z=\pm\frac{1}{2}\hbar} \int_{\text{all } \vec{r}} \Psi_1^*(\vec{r}, S_z; t)\Psi_2(\vec{r}, S_z; t)\,\mathrm{d}^3\vec{r}$$

or writing out the two-term sum,

$$\langle\Psi_1|\Psi_2\rangle = \int_{\text{all } \vec{r}} \Psi_1^*(\vec{r}, \tfrac{1}{2}\hbar; t)\Psi_2(\vec{r}, \tfrac{1}{2}\hbar; t)\,\mathrm{d}^3\vec{r} + \int_{\text{all } \vec{r}} \Psi_1^*(\vec{r}, -\tfrac{1}{2}\hbar; t)\Psi_2(\vec{r}, -\tfrac{1}{2}\hbar; t)\,\mathrm{d}^3\vec{r}$$

The individual factors in the integrals are by definition the spin-up components Ψ_{1+} and Ψ_{2+} and the spin down components Ψ_{1-} and Ψ_{2-} of the wave functions, so:

$$\langle\Psi_1|\Psi_2\rangle = \int_{\text{all } \vec{r}} \Psi_{1+}^*(\vec{r}; t)\Psi_{2+}(\vec{r}; t)\,\mathrm{d}^3\vec{r} + \int_{\text{all } \vec{r}} \Psi_{1-}^*(\vec{r}; t)\Psi_{2-}(\vec{r}; t)\,\mathrm{d}^3\vec{r}$$

In other words, the inner product with spin evaluates as

$$\boxed{\langle \Psi_{1+}\uparrow + \Psi_{1-}\downarrow | \Psi_{2+}\uparrow + \Psi_{2-}\downarrow\rangle = \langle \Psi_{1+}|\Psi_{2+}\rangle + \langle \Psi_{1-}|\Psi_{2-}\rangle} \tag{5.19}$$

It is spin-up components together and spin-down components together.

Another way of looking at this, or maybe remembering it, is to note that the spin states are an orthonormal pair,

$$\boxed{\langle\uparrow|\uparrow\rangle = 1 \qquad \langle\uparrow|\downarrow\rangle = \langle\downarrow|\uparrow\rangle = 0 \qquad \langle\downarrow|\downarrow\rangle = 1} \tag{5.20}$$

as can be verified directly from the definitions of those functions as given in the previous subsection. Then you can think of an inner product with spin as multiplying out as:

$$\begin{aligned}
&\langle \Psi_{1+}\uparrow + \Psi_{1-}\downarrow | \Psi_{2+}\uparrow + \Psi_{2-}\downarrow\rangle \\
&\quad = \langle \Psi_{1+}|\Psi_{2+}\rangle\langle\uparrow|\uparrow\rangle + \langle \Psi_{1+}|\Psi_{2-}\rangle\langle\uparrow|\downarrow\rangle + \langle \Psi_{1-}|\Psi_{2+}\rangle\langle\downarrow|\uparrow\rangle + \langle \Psi_{1-}|\Psi_{2-}\rangle\langle\downarrow|\downarrow\rangle \\
&\quad = \langle \Psi_{1+}|\Psi_{2+}\rangle + \langle \Psi_{1-}|\Psi_{2-}\rangle
\end{aligned}$$

Key Points

- In inner products, you must sum over the spin states.
- For spin ½ particles:

  $$\langle \Psi_{1+}\uparrow + \Psi_{1-}\downarrow | \Psi_{2+}\uparrow + \Psi_{2-}\downarrow\rangle = \langle \Psi_{1+}|\Psi_{2+}\rangle + \langle \Psi_{1-}|\Psi_{2-}\rangle$$

  which is spin-up components together plus spin-down components together.
- The spin-up and spin-down states $\uparrow$ and $\downarrow$ are an orthonormal pair.

5.5.2 Review Questions

1. Show that the normalization requirement for the wave function of a spin ½ particle in terms of Ψ_+ and Ψ_- requires its norm $\sqrt{\langle\Psi|\Psi\rangle}$ to be one.
   Solution complexsai-a
2. Assume that ψ_l and ψ_r are normalized spatial wave functions. Now show that a combination of the two like $(\psi_l\uparrow + \psi_r\downarrow)/\sqrt{2}$ is a normalized wave function with spin.
   Solution complexsai-b

5.5.3 Commutators including spin

There is no known "internal physical mechanism" that gives rise to spin like there is for orbital angular momentum. Fortunately, this lack of detailed information about spin is to a considerable amount made less of an issue by knowledge about its commutators.

In particular, physicists have concluded that spin components satisfy the same commutation relations as the components of orbital angular momentum:

$$\boxed{[\widehat{S}_x, \widehat{S}_y] = \mathrm{i}\hbar\widehat{S}_z \quad [\widehat{S}_y, \widehat{S}_z] = \mathrm{i}\hbar\widehat{S}_x \quad [\widehat{S}_z, \widehat{S}_x] = \mathrm{i}\hbar\widehat{S}_y} \tag{5.21}$$

These equations are called the "fundamental commutation relations." As will be shown in chapter 12, a large amount of information about spin can be teased from them.

Further, spin operators commute with all functions of the spatial coordinates and with all spatial operators, including position, linear momentum, and orbital angular momentum. The reason why can be understood from the given description of the wave function with spin. First of all, the square spin operator $\widehat{S}^2$ just multiplies the entire wave function by the constant $\hbar^2 s(s+1)$, and everything commutes with a constant. And the operator $\widehat{S}_z$ of spin in an arbitrary z-direction commutes with spatial functions and operators in much the same way that an operator like $\partial/\partial x$ commutes with functions depending on y and with $\partial/\partial y$. The z-component of spin corresponds to an additional "axis" separate from the x, y, and z ones, and $\widehat{S}_z$ only affects the variation in this additional direction. For example, for a particle with spin one half, $\widehat{S}_z$ multiplies the spin-up part of the wave function Ψ_+ by the constant $\frac{1}{2}\hbar$ and Ψ_- by $-\frac{1}{2}\hbar$. Spatial functions and operators commute with these constants for both Ψ_+ and Ψ_- hence commute with $\widehat{S}_z$ for the entire wave function. Since the z-direction is arbitrary, this commutation applies for any spin component.

Key Points

- While a detailed mechanism of spin is missing, commutators with spin can be evaluated.
- The components of spin satisfy the same mutual commutation relations as the components of orbital angular momentum.
- Spin commutes with spatial functions and operators.

5.5.3 Review Questions

1. Are not some commutators missing from the fundamental commutation relationship? For example, what is the commutator $[\widehat{S}_y, \widehat{S}_x]$?
   Solution complexsac-a

5.5.4 Wave function for multiple particles with spin

The extension of the ideas of the previous subsections towards multiple particles is straightforward. For two particles, such as the two electrons of the hydrogen

molecule, the full wave function follows from the "every possible combination" idea as

$$\boxed{\Psi = \Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t)} \tag{5.22}$$

The value of $|\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t)|^2 \, \mathrm{d}^3\vec{r}_1 \mathrm{d}^3\vec{r}_2$ gives the probability of simultaneously finding particle 1 within a vicinity $\mathrm{d}^3\vec{r}_1$ of $\vec{r}_1$ with spin angular momentum in the z-direction S_{z1}, and particle 2 within a vicinity $\mathrm{d}^3\vec{r}_2$ of $\vec{r}_2$ with spin angular momentum in the z-direction S_{z2}.

Restricting the attention again to spin $\frac{1}{2}$ particles like electrons, protons and neutrons, there are now four possible spin states at any given point, with corresponding spatial wave functions

$$\boxed{\begin{aligned}
\Psi_{++}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, +\tfrac{1}{2}\hbar, \vec{r}_2, +\tfrac{1}{2}\hbar; t) \\
\Psi_{+-}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, +\tfrac{1}{2}\hbar, \vec{r}_2, -\tfrac{1}{2}\hbar; t) \\
\Psi_{-+}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, -\tfrac{1}{2}\hbar, \vec{r}_2, +\tfrac{1}{2}\hbar; t) \\
\Psi_{--}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, -\tfrac{1}{2}\hbar, \vec{r}_2, -\tfrac{1}{2}\hbar; t)
\end{aligned}} \tag{5.23}$$

For example, $|\Psi_{+-}(\vec{r}_1, \vec{r}_2; t)|^2 \, \mathrm{d}^3\vec{r}_1 \, \mathrm{d}^3\vec{r}_2$ gives the probability of finding particle 1 within a vicinity $\mathrm{d}^3\vec{r}_1$ of $\vec{r}_1$ with spin up, and particle 2 within a vicinity $\mathrm{d}^3\vec{r}_2$ of $\vec{r}_2$ with spin down.

The wave function can be written using purely spatial functions and purely spin functions as

$$\begin{aligned}
\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) &= \Psi_{++}(\vec{r}_1, \vec{r}_2; t)\uparrow(S_{z1})\uparrow(S_{z2}) + \Psi_{+-}(\vec{r}_1, \vec{r}_2; t)\uparrow(S_{z1})\downarrow(S_{z2}) \\
&+ \Psi_{-+}(\vec{r}_1, \vec{r}_2; t)\downarrow(S_{z1})\uparrow(S_{z2}) + \Psi_{--}(\vec{r}_1, \vec{r}_2; t)\downarrow(S_{z1})\downarrow(S_{z2})
\end{aligned}$$

As you might guess from this multi-line display, usually this will be written more concisely as

$$\boxed{\Psi = \Psi_{++}\uparrow\uparrow + \Psi_{+-}\uparrow\downarrow + \Psi_{-+}\downarrow\uparrow + \Psi_{--}\downarrow\downarrow}$$

by leaving out the arguments of the spatial and spin functions. The understanding is that the first of each pair of arrows refers to particle 1 and the second to particle 2.

The inner product now evaluates as

$$\langle\Psi_1|\Psi_2\rangle = \sum_{S_{z1}=\pm\frac{1}{2}\hbar} \sum_{S_{z2}=\pm\frac{1}{2}\hbar} \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} \Psi_1^*(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t)\Psi_2(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) \, \mathrm{d}^3\vec{r}_1 \, \mathrm{d}^3\vec{r}_2$$

This can be written in terms of the purely spatial components as

$$\boxed{\langle\Psi_1|\Psi_2\rangle = \langle\Psi_{1++}|\Psi_{2++}\rangle + \langle\Psi_{1+-}|\Psi_{2+-}\rangle + \langle\Psi_{1-+}|\Psi_{2-+}\rangle + \langle\Psi_{1--}|\Psi_{2--}\rangle} \tag{5.24}$$

It reflects the fact that the four spin basis states $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, and $\downarrow\downarrow$ are an orthonormal quartet.

Key Points

- The wave function of a single particle with spin generalizes in a straightforward way to multiple particles with spin.
- The wave function of two spin $\frac{1}{2}$ particles can be written in terms of spatial components multiplying pure spin states as

$$\Psi = \Psi_{++}\uparrow\uparrow + \Psi_{+-}\uparrow\downarrow + \Psi_{-+}\downarrow\uparrow + \Psi_{--}\downarrow\downarrow$$

  where the first arrow of each pair refers to particle 1 and the second to particle 2.
- In terms of spatial components, the inner product $\langle\Psi_1|\Psi_2\rangle$ evaluates as inner products of matching spin components:

$$\langle\Psi_{1++}|\Psi_{2++}\rangle + \langle\Psi_{1+-}|\Psi_{2+-}\rangle + \langle\Psi_{1-+}|\Psi_{2-+}\rangle + \langle\Psi_{1--}|\Psi_{2--}\rangle$$

- The four spin basis states $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, and $\downarrow\downarrow$ are an orthonormal quartet.

5.5.4 Review Questions

1. As an example of the orthonormality of the two-particle spin states, verify that $\langle\uparrow\uparrow|\downarrow\uparrow\rangle$ is zero, so that $\uparrow\uparrow$ and $\downarrow\uparrow$ are indeed orthogonal. Do so by explicitly writing out the sums over S_{z1} and S_{z2}.

   Solution complexsb-a

2. A more concise way of understanding the orthonormality of the two-particle spin states is to note that an inner product like $\langle\uparrow\uparrow|\downarrow\uparrow\rangle$ equals $\langle\uparrow|\downarrow\rangle\langle\uparrow|\uparrow\rangle$, where the first inner product refers to the spin states of particle 1 and the second to those of particle 2. The first inner product is zero because of the orthogonality of $\uparrow$ and $\downarrow$, making $\langle\uparrow\uparrow|\downarrow\uparrow\rangle$ zero too.

   To check this argument, write out the sums over S_{z1} and S_{z2} for $\langle\uparrow|\downarrow\rangle\langle\uparrow|\uparrow\rangle$ and verify that it is indeed the same as the written out sum for $\langle\uparrow\uparrow|\downarrow\uparrow\rangle$ given in the answer for the previous question.

   The underlying mathematical principle is that sums of products can be factored into separate sums as in:

$$\sum_{\text{all } S_{z1}} \sum_{\text{all } S_{z2}} f(S_{z1})g(S_{z2}) = \left[\sum_{\text{all } S_{z1}} f(S_{z1})\right]\left[\sum_{\text{all } S_{z2}} g(S_{z2})\right]$$

   This is similar to the observation in calculus that integrals of products can be factored into separate integrals:

$$\int_{\text{all } \vec{r}_1}\int_{\text{all } \vec{r}_2} f(\vec{r}_1)g(\vec{r}_2)\,\mathrm{d}^3\vec{r}_1\,\mathrm{d}^3\vec{r}_2 =$$

$$\left[\int_{\text{all } \vec{r}_1} f(\vec{r}_1)\,\mathrm{d}^3\vec{r}_1\right]\left[\int_{\text{all } \vec{r}_2} g(\vec{r}_2)\,\mathrm{d}^3\vec{r}_2\right]$$

Solution complexsb-b

5.5.5 Example: the hydrogen molecule

As an example, this section considers the ground state of the hydrogen molecule. It was found in section 5.2 that the ground state electron wave function must be of the approximate form

$$\psi_{\rm gs,0} = a\,[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)]$$

where $\psi_{\rm l}$ was the electron ground state of the left hydrogen atom, and $\psi_{\rm r}$ the one of the right one; a was just a normalization constant. This solution excluded all consideration of spin.

Including spin, the ground state wave function must be of the general form

$$\psi_{\rm gs} = \psi_{++}\uparrow\uparrow + \psi_{+-}\uparrow\downarrow + \psi_{-+}\downarrow\uparrow + \psi_{--}\downarrow\downarrow.$$

As you might guess, in the ground state, each of the four spatial functions ψ_{++}, ψ_{+-}, ψ_{-+}, and ψ_{--} must be proportional to the no-spin solution $\psi_{\rm gs,0}$ above. Anything else would have more than the lowest possible energy, {D.24}.

So the approximate ground state including spin must take the form

$$\psi_{\rm gs} = a\,[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)]\,[a_{++}\uparrow\uparrow + a_{+-}\uparrow\downarrow + a_{-+}\downarrow\uparrow + a_{--}\downarrow\downarrow] \quad (5.25)$$

where a_{++}, a_{+-}, a_{-+}, and a_{--} are constants.

Key Points

- The electron wave function $\psi_{\rm gs,0}$ for the hydrogen molecule derived previously ignored spin.
- In the full electron wave function, each spatial component must separately be proportional to $a(\psi_{\rm l}\psi_{\rm r} + \psi_{\rm r}\psi_{\rm l})$.

5.5.5 Review Questions

1. Show that the normalization requirement for $\psi_{\rm gs}$ means that

$$|a_{++}|^2 + |a_{+-}|^2 + |a_{-+}|^2 + |a_{--}|^2 = 1$$

Solution complexsc-a

5.5.6 Triplet and singlet states

In the case of two particles with spin $\frac{1}{2}$, it is often more convenient to use slightly different basis states to describe the spin states than the four arrow

combinations ↑↑, ↑↓, ↓↑, and ↓↓. The more convenient basis states can be written in $|s\,m\rangle$ ket notation, and they are:

$$\underbrace{|1\,1\rangle = \uparrow\uparrow \quad |1\,0\rangle = \frac{1}{\sqrt{2}}(\uparrow\downarrow + \downarrow\uparrow) \quad |1\,{-1}\rangle = \downarrow\downarrow}_{\text{the triplet states}} \qquad \underbrace{|0\,0\rangle = \frac{1}{\sqrt{2}}(\uparrow\downarrow - \downarrow\uparrow)}_{\text{the singlet state}} \tag{5.26}$$

A state $|s\,m\rangle$ has *net* spin s, giving a net square angular momentum $s(s+1)\hbar^2$, and has *net* angular momentum in the z-direction $m\hbar$. For example, if the two particles are in the state $|1\,1\rangle$, the net square angular momentum is $2\hbar^2$, and their net angular momentum in the z-direction is $\hbar$.

The ↑↓ and ↓↑ states can be written as

$$\uparrow\downarrow = \frac{1}{\sqrt{2}}(|1\,0\rangle + |0\,0\rangle) \qquad \downarrow\uparrow = \frac{1}{\sqrt{2}}(|1\,0\rangle - |0\,0\rangle)$$

This shows that while they have zero angular momentum in the z-direction; they *do not* have a value for the net spin: they have a 50/50 probability of net spin 1 and net spin 0. A consequence is that ↑↓ and ↓↑ cannot be written in $|s\,m\rangle$ ket notation; there is no value for s. (Related to that, these states also do not have a definite value for the dot product of the two spins, {A.10}.

Incidentally, note that z components of angular momentum simply add up, as the Newtonian analogy suggests. For example, for ↑↓, the $\frac{1}{2}\hbar$ spin angular momentum of the first electron adds to the $-\frac{1}{2}\hbar$ of the second electron to produce zero. But Newtonian analysis does not allow square angular momenta to be added together, and neither does quantum mechanics. In fact, it is quite a messy exercise to actually prove that the triplet and singlet states have the net spin values claimed above. (See chapter 12 if you want to see how it is done.)

The spin states $\uparrow = |\frac{1}{2}\,\frac{1}{2}\rangle$ and $\downarrow = |\frac{1}{2}\,{-\frac{1}{2}}\rangle$ that apply for a single spin-$\frac{1}{2}$ particle are often referred to as the "doublet" states, since there are two of them.

Key Points

- The set of spin states ↑↑, ↑↓, ↓↑, and ↓↓ are often better replaced by the triplet and singlet states $|1\,1\rangle$, $|1\,0\rangle$, $|1\,{-1}\rangle$, and $|0\,0\rangle$.
- The triplet and singlet states have definite values for the net square spin.

5.5.6 Review Questions

1. Like the states ↑↑, ↑↓, ↓↑, and ↓↓; the triplet and singlet states are an orthonormal quartet. For example, check that the inner product of $|1\,0\rangle$ and $|0\,0\rangle$ is zero.

   Solution complexse-a

5.6 Identical Particles

A number of the counter-intuitive features of quantum mechanics have already been discussed: Electrons being neither on Mars or on Venus until they pop up at either place. Superluminal interactions. The fundamental impossibility of improving the accuracy of both position and momentum beyond a given limit. Collapse of the wave function. A hidden random number generator. Quantized energies and angular momenta. Nonexisting angular momentum vectors. Intrinsic angular momentum. But nature has one more trick on its sleeve, and it is a big one.

Nature entangles all identical particles with each other. Specifically, it requires that the wave function remains unchanged if any two identical bosons are exchanged. If particles i and j are identical bosons, then:

$$\Psi\left(\vec{r}_1, S_{z1}, \ldots, \vec{r}_i, S_{zi}, \ldots, \vec{r}_j, S_{zj}, \ldots\right) = \Psi\left(\vec{r}_1, S_{z1}, \ldots, \vec{r}_j, S_{zj}, \ldots, \vec{r}_i, S_{zi}, \ldots\right) \tag{5.27}$$

On the other hand, nature requires that the wave function changes sign if any two identical fermions are exchanged. If particles i and j are identical fermions, (say, both electrons), then:

$$\Psi\left(\vec{r}_1, S_{z1}, \ldots, \vec{r}_i, S_{zi}, \ldots, \vec{r}_j, S_{zj}, \ldots\right) = -\Psi\left(\vec{r}_1, S_{z1}, \ldots, \vec{r}_j, S_{zj}, \ldots, \vec{r}_i, S_{zi}, \ldots\right) \tag{5.28}$$

In other words, the wave function must be symmetric with respect to exchange of identical bosons, and antisymmetric with respect to exchange of identical fermions. This greatly restricts what wave functions can be.

For example, consider what this means for the electron structure of the hydrogen molecule. The approximate ground state of lowest energy was in the previous section found to be

$$\psi_{\rm gs} = a\left[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)\right]\left[a_{++}\uparrow\uparrow + a_{+-}\uparrow\downarrow + a_{-+}\downarrow\uparrow + a_{--}\downarrow\downarrow\right] \tag{5.29}$$

were $\psi_{\rm l}$ was the ground state of the left hydrogen atom, $\psi_{\rm r}$ the one of the right one, first arrows indicate the spin of electron 1 and second arrows the one of electron 2, and a and the $a_{\pm\pm}$ are constants.

But since the two electrons are identical fermions, this wave function must turn into its negative under exchange of the two electrons. Exchanging the two electrons produces

$$-\psi_{\rm gs} = a\left[\psi_{\rm l}(\vec{r}_2)\psi_{\rm r}(\vec{r}_1) + \psi_{\rm r}(\vec{r}_2)\psi_{\rm l}(\vec{r}_1)\right]\left[a_{++}\uparrow\uparrow + a_{+-}\downarrow\uparrow + a_{-+}\uparrow\downarrow + a_{--}\downarrow\downarrow\right];$$

note in particular that since the first arrow of each pair is taken to refer to electron 1, exchanging the electrons means that the order of each pair of arrows must be inverted. To compare the above wave function with the nonexchanged version (5.29), reorder the terms back to the same order:

$$-\psi_{\rm gs} = a\left[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)\right]\left[a_{++}\uparrow\uparrow + a_{-+}\uparrow\downarrow + a_{+-}\downarrow\uparrow + a_{--}\downarrow\downarrow\right]$$

The spatial factor is seen to be the same as the nonexchanged version in (5.29); the spatial part is symmetric under particle exchange. The sign change will have to come from the spin part.

Since each of the four spin states is independent from the others, the coefficient of each of these states will have to be the negative of the one of the nonexchanged version. For example, the coefficient a_{++} of $\uparrow\uparrow$ must be the negative of the coefficient a_{++} of $\uparrow\uparrow$ in the nonexchanged version, otherwise there is a conflict at $S_{z1} = \frac{1}{2}\hbar$ and $S_{z2} = \frac{1}{2}\hbar$, where only the spin state $\uparrow\uparrow$ is nonzero. Something can only be the negative of itself if it is zero, so a_{++} must be zero to satisfy the antisymmetry requirement. The same way, $a_{--} = -a_{--}$, requiring a_{--} to be zero too. The remaining two spin states both require that $a_{+-} = -a_{-+}$, but this can be nonzero.

So, due to the antisymmetrization requirement, the full wave function of the ground state must be,

$$\psi_{\rm gs} = a\,[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)]\,a_{+-}\,[\uparrow\downarrow - \downarrow\uparrow]$$

or after normalization, noting that a factor of magnitude one is always arbitrary,

$$\psi_{\rm gs} = a\,[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)]\,\frac{\uparrow\downarrow - \downarrow\uparrow}{\sqrt{2}}$$

It is seen that the antisymmetrization requirement restricts the spin state to be the "singlet" one, as defined in the previous section. It is the singlet spin state that achieves the sign change when the two electrons are exchanged; the spatial part remains the same.

If the electrons would have been bosons, the spin state could have been any combination of the three triplet states. The symmetrization requirement for fermions is much more restrictive than the one for bosons.

Since there are a lot more electrons in the universe than just these two, you might rightly ask where antisymmetrization stops. The answer given in chapter 8.3 is: nowhere. But don't worry about it. The existence of electrons that are too far away to affect the system being studied can be ignored.

Key Points

- The wave function must be symmetric (must stay the same) under exchange of identical bosons.
- The wave function must be antisymmetric (must turn into its negative) under exchange of identical fermions (e.g., electrons.)
- Especially the antisymmetrization requirement greatly restricts what wave functions can be.
- The antisymmetrization requirement forces the electrons in the hydrogen molecule ground state to assume the singlet spin state.

5.6 Review Questions

1. Check that indeed any linear combination of the triplet states is unchanged under particle exchange.

   Solution ident-a

2. Suppose the electrons of the hydrogen molecule are in the excited antisymmetric spatial state

$$a\left[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) - \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)\right].$$

   In that case what can you say about the spin state?

   Yes, in this case the spin would be less restricted if the electrons were bosons. But antisymmetric spatial states themselves are pretty restrictive in general. The precise sense in which the antisymmetrization requirement is more restrictive than the symmetrization requirement will be explored in the next section.

   Solution ident-b

5.7 Ways to Symmetrize the Wave Function

This section discusses ways in which the symmetrization requirements for wave functions of systems of identical particles can be achieved in general. This is a key issue in the numerical solution of any nontrivial quantum system, so this section will examine it in some detail.

It will be assumed that the approximate description of the wave function is done using a set of chosen single-particle functions, or "states",

$$\psi^{\rm p}_1(\vec{r}, S_z), \psi^{\rm p}_2(\vec{r}, S_z), \ldots$$

An example is provided by the approximate ground state of the hydrogen molecule from the previous section,

$$a\left[\psi_{\rm l}(\vec{r}_1)\psi_{\rm r}(\vec{r}_2) + \psi_{\rm r}(\vec{r}_1)\psi_{\rm l}(\vec{r}_2)\right]\frac{\uparrow\downarrow - \downarrow\uparrow}{\sqrt{2}}.$$

This can be multiplied out to be

$$\frac{a}{\sqrt{2}}\Big[\psi_{\rm l}(\vec{r}_1)\uparrow(S_{z1})\psi_{\rm r}(\vec{r}_2)\downarrow(S_{z2}) + \psi_{\rm r}(\vec{r}_1)\uparrow(S_{z1})\psi_{\rm l}(\vec{r}_2)\downarrow(S_{z2})$$
$$-\psi_{\rm l}(\vec{r}_1)\downarrow(S_{z1})\psi_{\rm r}(\vec{r}_2)\uparrow(S_{z2}) - \psi_{\rm r}(\vec{r}_1)\downarrow(S_{z1})\psi_{\rm l}(\vec{r}_2)\uparrow(S_{z2})\Big]$$

and consists of four single-particle functions:

$$\psi^{\rm p}_1(\vec{r}, S_z) = \psi_{\rm l}(\vec{r})\uparrow(S_z) \qquad \psi^{\rm p}_2(\vec{r}, S_z) = \psi_{\rm l}(\vec{r})\downarrow(S_z)$$
$$\psi^{\rm p}_3(\vec{r}, S_z) = \psi_{\rm r}(\vec{r})\uparrow(S_z) \qquad \psi^{\rm p}_4(\vec{r}, S_z) = \psi_{\rm r}(\vec{r})\downarrow(S_z).$$

The first of the four functions represents a single electron in the ground state around the left proton with spin up, the second a single electron in the same spatial state with spin down, etcetera. For better accuracy, more single-particle functions could be included, say excited atomic states in addition to the ground states. In terms of the above four functions, the expression for the hydrogen molecule ground state is

$$\frac{a}{\sqrt{2}}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2})+\frac{a}{\sqrt{2}}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2})$$
$$-\frac{a}{\sqrt{2}}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2})-\frac{a}{\sqrt{2}}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2})$$

The issue in this section is that the above hydrogen ground state is just one special case of the most general wave function for the two particles that can be formed from four single-particle states:

$$\begin{aligned}\Psi(\vec{r}_1,S_{z1},\vec{r}_2,S_{z2};t)=&\\ &a_{11}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2})+a_{12}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{13}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2})+a_{14}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{21}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2})+a_{22}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{23}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2})+a_{24}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{31}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2})+a_{32}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{33}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2})+a_{34}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{41}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2})+a_{42}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2})+\\ &a_{43}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2})+a_{44}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2})\end{aligned}$$

This can be written much more concisely using summation indices as

$$\Psi(\vec{r}_1,S_{z1},\vec{r}_2,S_{z2};t)=\sum_{n_1=1}^{4}\sum_{n_2=1}^{4}a_{n_1n_2}\psi_{n_1}^{\rm p}(\vec{r}_1,S_{z1})\psi_{n_2}^{\rm p}(\vec{r}_2,S_{z2})$$

However, the individual terms will be fully written out for now to reduce the mathematical abstraction. The individual terms are sometimes called "Hartree products."

The antisymmetrization requirement says that the wave function must be antisymmetric under exchange of the two electrons. More concretely, it must turn into its negative when the arguments $\vec{r}_1, S_{z1}$ and $\vec{r}_2, S_{z2}$ are swapped. To

understand what that means, the various terms need to be arranged in groups:

$$\begin{aligned}
\text{I}: &\quad a_{11}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{II}: &\quad a_{22}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{III}: &\quad a_{33}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{IV}: &\quad a_{44}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{V}: &\quad a_{12}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2}) + a_{21}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{VI}: &\quad a_{13}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2}) + a_{31}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{VII}: &\quad a_{14}\psi_1^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2}) + a_{41}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_1^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{VIII}: &\quad a_{23}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2}) + a_{32}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{IX}: &\quad a_{24}\psi_2^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2}) + a_{42}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_2^{\rm p}(\vec{r}_2,S_{z2}) \\
\text{X}: &\quad a_{34}\psi_3^{\rm p}(\vec{r}_1,S_{z1})\psi_4^{\rm p}(\vec{r}_2,S_{z2}) + a_{43}\psi_4^{\rm p}(\vec{r}_1,S_{z1})\psi_3^{\rm p}(\vec{r}_2,S_{z2})
\end{aligned}$$

Within each group, all terms involve the *same* combination of functions, but in a different *order*. Different groups have a different combination of functions.

Now if the electrons are exchanged, it turns the terms in groups I through IV back into themselves. Since the wave function must change sign in the exchange, and something can only be its own negative if it is zero, the antisymmetrization requirement requires that the coefficients a_{11}, a_{22}, a_{33}, and a_{44} must all be zero. Four coefficients have been eliminated from the list of unknown quantities.

Further, in each of the groups V through X with two different states, exchange of the two electrons turn the terms into each other, except for their coefficients. If that is to achieve a change of sign, the coefficients must be each other's negatives; $a_{21} = -a_{12}$, $a_{31} = -a_{13}$, ... So only six coefficients a_{12}, a_{13}, ... still need to be found from other physical requirements, such as energy minimization for a ground state. Less than half of the original sixteen unknowns survive the antisymmetrization requirement, significantly reducing the problem size.

There is a very neat way of writing the antisymmetrized wave function of systems of fermions, which is especially convenient for larger numbers of particles. It is done using determinants. The antisymmetric wave function for the above example is:

$$\begin{aligned}
\Psi \;=\; & a_{12}\begin{vmatrix} \psi_1^{\rm p}(\vec{r}_1,S_{z1}) & \psi_2^{\rm p}(\vec{r}_1,S_{z1}) \\ \psi_1^{\rm p}(\vec{r}_2,S_{z2}) & \psi_2^{\rm p}(\vec{r}_2,S_{z2}) \end{vmatrix} + a_{13}\begin{vmatrix} \psi_1^{\rm p}(\vec{r}_1,S_{z1}) & \psi_3^{\rm p}(\vec{r}_1,S_{z1}) \\ \psi_1^{\rm p}(\vec{r}_2,S_{z2}) & \psi_3^{\rm p}(\vec{r}_2,S_{z2}) \end{vmatrix} + \\
& a_{14}\begin{vmatrix} \psi_1^{\rm p}(\vec{r}_1,S_{z1}) & \psi_4^{\rm p}(\vec{r}_1,S_{z1}) \\ \psi_1^{\rm p}(\vec{r}_2,S_{z2}) & \psi_4^{\rm p}(\vec{r}_2,S_{z2}) \end{vmatrix} + a_{23}\begin{vmatrix} \psi_2^{\rm p}(\vec{r}_1,S_{z1}) & \psi_3^{\rm p}(\vec{r}_1,S_{z1}) \\ \psi_2^{\rm p}(\vec{r}_2,S_{z2}) & \psi_3^{\rm p}(\vec{r}_2,S_{z2}) \end{vmatrix} + \\
& a_{24}\begin{vmatrix} \psi_2^{\rm p}(\vec{r}_1,S_{z1}) & \psi_4^{\rm p}(\vec{r}_1,S_{z1}) \\ \psi_2^{\rm p}(\vec{r}_2,S_{z2}) & \psi_4^{\rm p}(\vec{r}_2,S_{z2}) \end{vmatrix} + a_{34}\begin{vmatrix} \psi_3^{\rm p}(\vec{r}_1,S_{z1}) & \psi_4^{\rm p}(\vec{r}_1,S_{z1}) \\ \psi_3^{\rm p}(\vec{r}_2,S_{z2}) & \psi_4^{\rm p}(\vec{r}_2,S_{z2}) \end{vmatrix}
\end{aligned}$$

These determinants are called "Slater determinants".

To find the actual hydrogen molecule ground state from the above expression, additional physical requirements have to be imposed. For example, the

coefficients a_{12} and a_{34} can reasonably be ignored for the ground state, because according to the given definition of the states, their Slater determinants have the electrons around the same nucleus, and that produces elevated energy due to the mutual repulsion of the electrons. Also, following the arguments of section 5.2, the coefficients a_{13} and a_{24} must be zero since their Slater determinants produce the excited antisymmetric spatial state $\psi_l\psi_r - \psi_r\psi_l$ times the $\uparrow\uparrow$, respectively $\downarrow\downarrow$ spin states. Finally, the coefficients a_{14} and a_{23} must be opposite in order that their Slater determinants combine into the lowest-energy symmetric spatial state $\psi_l\psi_r + \psi_r\psi_l$ times the $\uparrow\downarrow$ and $\downarrow\uparrow$ spin states. That leaves the single coefficient a_{14} that can be found from the normalization requirement, taking it real and positive for convenience.

But the issue in this section is what the symmetrization requirements say about wave functions in general, whether they are some ground state or not. And for four single-particle states for two identical fermions, the conclusion is that the wave function must be some combination of the six Slater determinants, regardless of what other physics may be relevant.

The next question is how that conclusion changes if the two particles involved are not fermions, but identical bosons. The symmetrization requirement is then that exchanging the particles must leave the wave function unchanged. Since the terms in groups I through IV do remain the same under particle exchange, their coefficients a_{11} through a_{44} can have any nonzero value. This is the sense in which the antisymmetrization requirement for fermions is much more restrictive than the one for bosons: groups involving a duplicated state must be zero for fermions, but not for bosons.

In groups V through X, where particle exchange turns each of the two terms into the other one, the coefficients must now be equal instead of negatives; $a_{21} = a_{12}$, $a_{31} = a_{13}$, ... That eliminates six coefficients from the original sixteen unknowns, leaving ten coefficients that must be determined by other physical requirements on the wave function.

(The equivalent of Slater determinants for bosons are "permanents," basically determinants with all minus signs in their definition replaced by plus signs. Unfortunately, many of the helpful properties of determinants do not apply to permanents.)

All of the above arguments can be extended to the general case that N, instead of 4, single-particle functions $\psi_1^{\rm p}(\vec{r}, S_z)$, $\psi_2^{\rm p}(\vec{r}, S_z)$, ..., $\psi_N^{\rm p}(\vec{r}, S_z)$ are used to describe I, instead of 2, particles. Then the most general possible wave function assumes the form:

$$\Psi = \sum_{n_1=1}^{N} \sum_{n_2=1}^{N} \cdots \sum_{n_I=1}^{N} a_{n_1 n_2 \ldots n_I} \psi_{n_1}^{\rm p}(\vec{r}_1, S_{z1}) \psi_{n_2}^{\rm p}(\vec{r}_2, S_{z2}) \ldots \psi_{n_I}^{\rm p}(\vec{r}_I, S_{zI}) \qquad (5.30)$$

where the $a_{n_1 n_2 \ldots n_I}$ are numerical coefficients that are to be chosen to satisfy the physical constraints on the wave function, including the (anti) symmetrization requirements.

This summation is again the "every possible combination" idea of combining every possible state for particle 1 with every possible state for particle 2, etcetera. So the total sum above contains N^I terms: there are N possibilities for the function number n_1 of particle 1, times N possibilities for the function number n_2 of particle 2, ... In general then, a corresponding total of N^I unknown coefficients $a_{n_1 n_2 \ldots n_I}$ must be determined to find out the precise wave function.

But for identical particles, the number that must be determined is much less. That number can again be determined by dividing the terms into groups in which the terms all involve the same combination of I single-particle functions, just in a different order. The simplest groups are those that involve just a single single-particle function, generalizing the groups I through IV in the earlier example. Such groups consist of only a single term; for example, the group that only involves $\psi_1^{\rm p}$ consists of the single term

$$a_{11\ldots1}\psi_1^{\rm p}(\vec{r}_1, S_{z1})\psi_1^{\rm p}(\vec{r}_2, S_{z2}) \ldots \psi_1^{\rm p}(\vec{r}_I, S_{zI}).$$

At the other extreme, groups in which every single-particle function is different have as many as $I!$ terms, since $I!$ is the number of ways that I different items can be ordered. In the earlier example, that were groups V through X, each having $2! = 2$ terms. If there are more than two particles, there will also be groups in which some states are the same and some are different.

For identical bosons, the symmetrization requirement says that all the coefficients within a group must be equal. Any term in a group can be turned into any other by particle exchanges; so, if they would not all have the same coefficients, the wave function could be changed by particle exchanges. As a result, for identical bosons the number of unknown coefficients reduces to the number of groups.

For identical fermions, only groups in which all single-particle functions are different can be nonzero. That follows because if a term has a duplicated single-particle function, it turns into itself without the required sign change under an exchange of the particles of the duplicated function.

So there is no way to describe a system of I identical fermions with anything less than I different single-particle functions $\psi_n^{\rm p}$. This critically important observation is known as the "Pauli exclusion principle:" $I-1$ fermions occupying $I-1$ single-particle functions exclude a I-th fermion from simply entering the same $I-1$ functions; a new function must be added to the mix for each additional fermion. The more identical fermions there are in a system, the more different single-particle functions are required to describe it.

Each group involving I different single-particle functions $\psi_{n_1}^{\rm p}, \psi_{n_2}^{\rm p}, \ldots \psi_{n_I}^{\rm p}$ reduces under the antisymmetrization requirement to a single Slater determinant

of the form

$$\frac{1}{\sqrt{I!}} \begin{vmatrix} \psi^{\rm p}_{n_1}(\vec{r}_1,S_{z1}) & \psi^{\rm p}_{n_2}(\vec{r}_1,S_{z1}) & \psi^{\rm p}_{n_3}(\vec{r}_1,S_{z1}) & \cdots & \psi^{\rm p}_{n_I}(\vec{r}_1,S_{z1}) \\ \psi^{\rm p}_{n_1}(\vec{r}_2,S_{z2}) & \psi^{\rm p}_{n_2}(\vec{r}_2,S_{z2}) & \psi^{\rm p}_{n_3}(\vec{r}_2,S_{z2}) & \cdots & \psi^{\rm p}_{n_I}(\vec{r}_2,S_{z2}) \\ \psi^{\rm p}_{n_1}(\vec{r}_3,S_{z3}) & \psi^{\rm p}_{n_2}(\vec{r}_3,S_{z3}) & \psi^{\rm p}_{n_3}(\vec{r}_3,S_{z3}) & \cdots & \psi^{\rm p}_{n_I}(\vec{r}_3,S_{z3}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi^{\rm p}_{n_1}(\vec{r}_I,S_{zI}) & \psi^{\rm p}_{n_2}(\vec{r}_I,S_{zI}) & \psi^{\rm p}_{n_3}(\vec{r}_I,S_{zI}) & \cdots & \psi^{\rm p}_{n_I}(\vec{r}_I,S_{zI}) \end{vmatrix} \tag{5.31}$$

multiplied by a single unknown coefficient. The normalization factor $1/\sqrt{I!}$ has been thrown in merely to ensure that if the functions $\psi^{\rm p}_n$ are orthonormal, then so are the Slater determinants. Using Slater determinants ensures the required sign changes of fermion systems automatically, because determinants change sign if two rows are exchanged.

In the case that the bare minimum of I functions is used to describe I identical fermions, only one Slater determinant can be formed. Then the antisymmetrization requirement reduces the I^I unknown coefficients $a_{n_1n_2\ldots n_I}$ to just one, $a_{12\ldots I}$; obviously a tremendous reduction.

At the other extreme, when the number of functions N is very large, much larger than I^2 to be precise, most terms have all indices different and the reduction is "only" from N^I to about $N^I/I!$ terms. The latter would also be true for identical bosons.

The functions better be chosen to produce a good approximation to the wave function with a small number of terms. As an arbitrary example to focus the thoughts, if $N = 100$ functions are used to describe an arsenic atom, with $I = 33$ electrons, there would be a prohibitive 10^{66} terms in the sum (5.30). Even after reduction to Slater determinants, there would still be a prohibitive $3\ 10^{26}$ or so unknown coefficients left. The precise expression for the number of Slater determinants is called "N choose I;" it is given by

$$\left(\begin{array}{c} N \\ I \end{array} \right) = \frac{N!}{(N-I)!I!} = \frac{N(N-1)(N-2)\ldots(N-I+1)}{I!},$$

since the top gives the total number of terms that have all functions different, (N possible functions for particle 1, times $N-1$ possible functions left for particle 2, etcetera,) and the bottom reflects that it takes $I!$ of them to form a single Slater determinant. {D.25}.

The basic "Hartree-Fock" approach, discussed in chapter 9.3, goes to the extreme in reducing the number of functions: it uses the very minimum of I single-particle functions. However, rather than choosing these functions a priori, they are adjusted to give the best approximation that is possible with a single Slater determinant. Unfortunately, if a single determinant still turns out to be not accurate enough, adding a few more functions quickly blows up in your face. Adding just one more function gives I more determinants; adding another function gives another $I(I+1)/2$ more determinants, etcetera.

Key Points

- Wave functions for multiple-particle systems can be formed using sums of products of single-particle wave functions.
- The coefficients of these products are constrained by the symmetrization requirements.
- In particular, for identical fermions such as electrons, the single-particle wave functions must combine into Slater determinants.
- Systems of identical fermions require at least as many single-particle states as there are particles. This is known as the Pauli exclusion principle.
- If more single-particle states are used to describe a system, the problem size increases rapidly.

5.7 Review Questions

1. How many single-particle states would a basic Hartree-Fock approximation use to compute the electron structure of an arsenic atom? How many Slater determinants would that involve?
   Solution symways-a
2. If two more single-particle states would be used to improve the accuracy for the arsenic atom, (one more normally does not help), how many Slater determinants could be formed with those states?
   Solution symways-b

5.8 Matrix Formulation

When the number of unknowns in a quantum mechanical problem has been reduced to a finite number, the problem can be reduced to a linear algebra one. This allows the problem to be solved using standard analytical or numerical techniques. This section describes how the linear algebra problem can be obtained.

Typically, quantum mechanical problems can be reduced to a finite number of unknowns using some finite set of chosen wave functions, as in the previous section. There are other ways to make the problems finite, it does not really make a difference here. But in general some simplification will still be needed afterwards. A multiple sum like equation (5.30) for distinguishable particles is awkward to work with, and when various coefficients drop out for identical particles, its gets even messier. So as a first step, it is best to order the terms involved in some way; any ordering will in principle do. Ordering allows each term to be indexed by a single counter q, being the place of the term in the ordering.

Using an ordering, the wave function for a total of I particles can be written more simply as

$$\Psi = a_1\psi_1^{\rm S}(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_I, S_{zI}) + a_2\psi_2^{\rm S}(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_I, S_{zI}) + \ldots$$

or in index notation:

$$\Psi = \sum_{q=1}^{Q} a_q\psi_q^{\rm S}(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_I, S_{zI}). \tag{5.32}$$

where Q is the total count of the chosen I-particle wave functions and the single counter q in a_q replaces a set of I indices in the description used in the previous section. The I-particle functions $\psi_q^{\rm S}$ are allowed to be anything; individual (Hartree) products of single-particle wave functions for distinguishable particles as in (5.30), Slater determinants for identical fermions, permanents for identical bosons, or whatever. The only thing that will be assumed is that they are mutually orthonormal. (Which means that any underlying set of single-particle functions $\psi_n^{\rm p}(\vec{r})$ as described in the previous section should be orthonormal. If they are not, there are procedures like Gram-Schmidt to make them so. Or you can just put in some correction terms.)

Under those conditions, the energy eigenvalue problem $H\psi = E\psi$ takes the form:

$$\sum_{q=1}^{Q} Ha_q\psi_q^{\rm S} = \sum_{q=1}^{Q} Ea_q\psi_q^{\rm S}$$

The trick is now to take the inner product of both sides of this equation with each function $\psi_{\underline{q}}^{\rm S}$ in the set of wave functions in turn. In other words, take an inner product with $\langle\psi_1^{\rm S}|$ to get one equation, then take an inner product with $\langle\psi_2^{\rm S}|$ to get a second equation, and so on. This produces, using the fact that the functions are orthonormal to clean up the right-hand side,

$$\begin{array}{ccccccccccc}
H_{11}a_1 & + & H_{12}a_2 & + & \ldots & + & H_{1Q}a_Q & = & Ea_1 \\
H_{21}a_1 & + & H_{22}a_2 & + & \ldots & + & H_{2Q}a_Q & = & Ea_2 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
H_{\underline{q}1}a_1 & + & H_{\underline{q}2}a_2 & + & \ldots & + & H_{\underline{q}Q}a_Q & = & Ea_{\underline{q}} \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
H_{Q1}a_1 & + & H_{Q2}a_2 & + & \ldots & + & H_{QQ}a_Q & = & Ea_Q
\end{array}$$

where

$$H_{11} = \langle\psi_1^{\rm S}|H\psi_1^{\rm S}\rangle,\ H_{12} = \langle\psi_1^{\rm S}|H\psi_2^{\rm S}\rangle,\ \ldots,\ H_{QQ} = \langle\psi_Q^{\rm S}|H\psi_Q^{\rm S}\rangle.$$

are the matrix coefficients, or Hamiltonian coefficients.

This can again be written more compactly in index notation:

$$\sum_{q=1}^{Q} H_{\underline{q}q} a_q = E a_{\underline{q}} \quad \text{for } \underline{q} = 1, 2, \ldots, Q \qquad \text{with } H_{\underline{q}q} = \langle \psi^{\mathrm{S}}_{\underline{q}} | H \psi^{\mathrm{S}}_{q} \rangle \tag{5.33}$$

which is just a finite-size matrix eigenvalue problem.

Since the functions ψ^{S}_q are known, chosen, functions, and the Hamiltonian H is also known, the matrix coefficients $H_{\underline{q}q}$ can be determined. The eigenvalues E and corresponding eigenvectors $(a_1, a_2, \ldots)$ can then be found using linear algebra procedures. Each eigenvector produces a corresponding approximate eigenfunction $a_1\psi^{\mathrm{S}}_1 + a_2\psi^{\mathrm{S}}_2 + \ldots$ with an energy equal to the eigenvalue E.

Key Points

- Operator eigenvalue problems can be approximated by the matrix eigenvalue problems of linear algebra.
- That allows standard analytical or numerical techniques to be used in their solution.

5.8 Review Questions

1. As a relatively simple example, work out the above ideas for the $Q = 2$ hydrogen molecule spatial states $\psi^{\mathrm{S}}_1 = \psi_{\mathrm{l}}\psi_{\mathrm{r}}$ and $\psi^{\mathrm{S}}_2 = \psi_{\mathrm{l}}\psi_{\mathrm{r}}$. Write the matrix eigenvalue problem and identify the two eigenvalues and eigenvectors. Compare with the results of section 5.3.

   Assume that ψ_{l} and ψ_{r} have been slightly adjusted to be orthonormal. Then so are ψ^{S}_1 and ψ^{S}_2 orthonormal, since the various six-dimensional inner product integrals, like

   $$\langle \psi^{\mathrm{S}}_1 | \psi^{\mathrm{S}}_2 \rangle \equiv \langle \psi_{\mathrm{l}}\psi_{\mathrm{r}} | \psi_{\mathrm{r}}\psi_{\mathrm{l}} \rangle \equiv$$
   $$\int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} \psi_{\mathrm{l}}(\vec{r}_1)\psi_{\mathrm{r}}(\vec{r}_2)\; \psi_{\mathrm{r}}(\vec{r}_1)\psi_{\mathrm{l}}(\vec{r}_2)\, \mathrm{d}^3\vec{r}_1\, \mathrm{d}^3\vec{r}_2$$

   can according to the rules of calculus be factored into three-dimensional integrals as

   $$\begin{aligned} &\langle \psi^{\mathrm{S}}_1 | \psi^{\mathrm{S}}_2 \rangle \\ &\quad = \left[\int_{\text{all } \vec{r}_1} \psi_{\mathrm{l}}(\vec{r}_1)\; \psi_{\mathrm{r}}(\vec{r}_1)\, \mathrm{d}^3\vec{r}_1 \right] \left[\int_{\text{all } \vec{r}_2} \psi_{\mathrm{r}}(\vec{r}_2)\; \psi_{\mathrm{l}}(\vec{r}_2)\, \mathrm{d}^3\vec{r}_2 \right] \\ &\quad = \langle \psi_{\mathrm{l}} | \psi_{\mathrm{r}} \rangle \langle \psi_{\mathrm{r}} | \psi_{\mathrm{l}} \rangle \end{aligned}$$

   which is zero if ψ_{l} and ψ_{r} are orthonormal.

Also, do not try to find actual values for H_{11}, H_{12}, H_{21}, and H_{22}. As section 5.2 noted, that can only be done numerically. Instead just refer to H_{11} as J and to H_{12} as $-L$:

$$H_{11} \equiv \langle \psi_1^{\rm S} | H \psi_1^{\rm S} \rangle \equiv \langle \psi_{\rm l} \psi_{\rm r} | H \psi_{\rm l} \psi_{\rm r} \rangle \equiv J$$

$$H_{12} \equiv \langle \psi_1^{\rm S} | H \psi_2^{\rm S} \rangle \equiv \langle \psi_{\rm l} \psi_{\rm r} | H \psi_{\rm r} \psi_{\rm l} \rangle \equiv -L.$$

Next note that you also have

$$H_{22} \equiv \langle \psi_2^{\rm S} | H \psi_2^{\rm S} \rangle \equiv \langle \psi_{\rm r} \psi_{\rm l} | H \psi_{\rm r} \psi_{\rm l} \rangle = J$$

$$H_{21} \equiv \langle \psi_2^{\rm S} | H \psi_1^{\rm S} \rangle \equiv \langle \psi_{\rm r} \psi_{\rm l} | H \psi_{\rm l} \psi_{\rm r} \rangle = -L$$

because they are the exact same inner product integrals; the difference is just which electron you number 1 and which one you number 2 that determines whether the wave functions are listed as $\psi_{\rm l}\psi_{\rm r}$ or $\psi_{\rm r}\psi_{\rm l}$.

Solution matfor-a

2. Find the eigenstates for the same problem, but now including spin.

As section 5.7 showed, the antisymmetric wave function with spin consists of a sum of six Slater determinants. Ignoring the highly excited first and sixth determinants that have the electrons around the same nucleus, the remaining $C = 4$ Slater determinants can be written out explicitly to give the two-particle states

$$\psi_1^{\rm S} = \frac{\psi_{\rm l}\psi_{\rm r}\uparrow\uparrow - \psi_{\rm r}\psi_{\rm l}\uparrow\uparrow}{\sqrt{2}} \qquad \psi_2^{\rm S} = \frac{\psi_{\rm l}\psi_{\rm r}\uparrow\downarrow - \psi_{\rm r}\psi_{\rm l}\downarrow\uparrow}{\sqrt{2}}$$

$$\psi_3^{\rm S} = \frac{\psi_{\rm l}\psi_{\rm r}\downarrow\uparrow - \psi_{\rm r}\psi_{\rm l}\uparrow\downarrow}{\sqrt{2}} \qquad \psi_4^{\rm S} = \frac{\psi_{\rm l}\psi_{\rm r}\downarrow\downarrow - \psi_{\rm r}\psi_{\rm l}\downarrow\downarrow}{\sqrt{2}}$$

Note that the Hamiltonian does not involve spin, to the approximation used in most of this book, so that, following the techniques of section 5.5, an inner product like $H_{23} = \langle \psi_2^{\rm S} | H \psi_3^{\rm S} \rangle$ can be written out like

$$\begin{aligned} H_{23} &= \frac{1}{2}\langle \psi_{\rm l}\psi_{\rm r}\uparrow\downarrow - \psi_{\rm r}\psi_{\rm l}\downarrow\uparrow | H(\psi_{\rm l}\psi_{\rm r}\downarrow\uparrow - \psi_{\rm r}\psi_{\rm l}\uparrow\downarrow)\rangle \\ &= \frac{1}{2}\langle \psi_{\rm l}\psi_{\rm r}\uparrow\downarrow - \psi_{\rm r}\psi_{\rm l}\downarrow\uparrow | (H\psi_{\rm l}\psi_{\rm r})\downarrow\uparrow - (H\psi_{\rm r}\psi_{\rm l})\uparrow\downarrow\rangle \end{aligned}$$

and then multiplied out into inner products of matching spin components to give

$$H_{23} = -\frac{1}{2}\langle \psi_{\rm l}\psi_{\rm r} | H \psi_{\rm r}\psi_{\rm l}\rangle - \frac{1}{2}\langle \psi_{\rm r}\psi_{\rm l} | H \psi_{\rm l}\psi_{\rm r}\rangle = L.$$

The other 15 matrix coefficients can be found similarly, and most will be zero.

If you do not have experience with linear algebra, you may want to skip this question, or better, just read the solution. However, the four eigenvectors are not that hard to guess; maybe easier to guess than correctly derive.

Solution matfor-b

5.9 Heavier Atoms

This section solves the ground state electron configuration of the atoms of elements heavier than hydrogen. The atoms of the elements are distinguished by their "atomic number" Z, which is the number of protons in the nucleus. For the neutral atoms considered in this section, Z is also the number of electrons circling the nucleus.

A crude approximation will be made to deal with the mutual interactions between the electrons. Still, many properties of the elements can be understood using this crude model, such as their geometry and chemical properties, and how the Pauli exclusion principle raises the energy of the electrons.

This is a descriptive section, in which no new analytical procedures are taught. However, it is a very important section to read, and reread, because much of our qualitative understanding of nature is based on the ideas in this section.

5.9.1 The Hamiltonian eigenvalue problem

The procedure to find the ground state of the heavier atoms is similar to the one for the hydrogen atom of chapter 4.3. The total energy Hamiltonian for the electrons of an element with atomic number Z with is:

$$H = \sum_{i=1}^{Z} \left[-\frac{\hbar^2}{2m_e}\nabla_i^2 - \frac{e^2}{4\pi\epsilon_0}\frac{Z}{r_i} + \tfrac{1}{2}\sum_{\substack{\underline{i}=1 \\ \underline{i}\neq i}}^{Z} \frac{e^2}{4\pi\epsilon_0}\frac{1}{|\vec{r}_i - \vec{r}_{\underline{i}}|} \right] \tag{5.34}$$

Within the brackets, the first term represents the kinetic energy of electron number i out of Z, the second the attractive potential due to the nuclear charge Ze, and the final term is the repulsion by all the other electrons. In the Hamiltonian as written, it is assumed that half of the energy of a repulsion is credited to each of the two electrons involved, accounting for the factor $\frac{1}{2}$.

The Hamiltonian eigenvalue problem for the energy states takes the form:

$$H\psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_Z, S_{zZ}) = E\psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_Z, S_{zZ})$$

Key Points

- The Hamiltonian for the electron structure has been written down.

5.9.2 Approximate solution using separation of variables

The Hamiltonian eigenvalue problem of the previous subsection cannot be solved exactly. The repulsive interactions between the electrons, given by the last term in the Hamiltonian are too complex.

More can be said under the, really poor, approximation that each electron "sees" a repulsion by the other $Z-1$ electrons that averages out as if the other electrons are located in the nucleus. The other $Z-1$ electrons then reduce the net charge of the nucleus from Ze to e. An other way of saying this is that each of the $Z-1$ other electrons "shields" one proton in the nucleus, allowing only a single remaining proton charge to filter through.

In this crude approximation, the electrons do not notice each other at all; they see only a single charge *hydrogen* nucleus. Obviously then, the wave function solutions for each electron should be the ψ_{nlm} eigenfunctions of the hydrogen atom, which were found in chapter 4.3.

To verify this explicitly, the approximate Hamiltonian is

$$H = \sum_{i=1}^{Z} \left\{ -\frac{\hbar^2}{2m}\nabla_i^2 - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_i} \right\}$$

since this represents a system of noninteracting electrons in which each experiences an hydrogen nucleus potential. This can be written more concisely as

$$H = \sum_{i=1}^{Z} h_i$$

where h_i is the hydrogen-atom Hamiltonian for electron number i,

$$h_i = -\frac{\hbar^2}{2m}\nabla_i^2 - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_i}.$$

The approximate Hamiltonian eigenvalue problem can now be solved using a method of separation of variables in which solutions are sought that take the form of products of single-electron wave functions:

$$\psi^Z = \psi_1^{\rm p}(\vec{r}_1, S_{z1})\psi_2^{\rm p}(\vec{r}_2, S_{z2}) \ldots \psi_Z^{\rm p}(\vec{r}_Z, S_{zZ}).$$

Substitution of this assumption into the eigenvalue problem $\sum_i h_i\psi^Z = E\psi^Z$ and dividing by ψ^Z produces

$$\frac{1}{\psi_1^{\rm p}(\vec{r}_1, S_{z1})}h_1\psi_1^{\rm p}(\vec{r}_1, S_{z1}) + \frac{1}{\psi_2^{\rm p}(\vec{r}_2, S_{z2})}h_2\psi_2^{\rm p}(\vec{r}_2, S_{z2}) + \ldots = E$$

since h_1 only does anything to the factor $\psi_1^{\rm p}(\vec{r}_1, S_{z1})$, h_2 only does anything to the factor $\psi_2^{\rm p}(\vec{r}_2, S_{z2})$, etcetera.

The first term in the equation above must be some constant ϵ_1; it cannot vary with $\vec{r}_1$ or S_{z1} as $\psi_1^{\rm p}(\vec{r}_1, S_{z1})$ itself does, since none of the other terms in the equation varies with those variables. That means that

$$h_1\psi_1^{\rm p}(\vec{r}_1, S_{z1}) = \epsilon_1\psi_1^{\rm p}(\vec{r}_1, S_{z1}),$$

which is an hydrogen atom eigenvalue problem for the single-electron wave function of electron 1. So, the single-electron wave function of electron 1 can be any one of the hydrogen atom wave functions from chapter 4.3; allowing for spin, the possible solutions are,

$$\psi_{100}(\vec{r}_1)\uparrow(S_{z1}),\ \psi_{100}(\vec{r}_1)\downarrow(S_{z1}),\ \psi_{200}(\vec{r}_1)\uparrow(S_{z1}),\ \psi_{200}(\vec{r}_1)\downarrow(S_{z1}),\ \ldots$$

The energy ϵ_1 is the corresponding hydrogen atom energy level, E_1 for $\psi_{100}\uparrow$ or $\psi_{100}\downarrow$, E_2 for any of the eight states $\psi_{200}\uparrow$, $\psi_{200}\downarrow$, $\psi_{211}\uparrow$, $\psi_{211}\downarrow$, $\psi_{210}\uparrow$, $\psi_{210}\downarrow$, $\psi_{21-1}\uparrow$, $\psi_{21-1}\downarrow$, etcetera.

The same observations hold for the other electrons; their single-electron eigenfunctions are $\psi_{nlm}\updownarrow$ hydrogen atom ones, (where $\updownarrow$ can be either $\uparrow$ or $\downarrow$.) Their individual energies must be the corresponding hydrogen atom energy levels.

The final wave functions for all Z electrons are then each a product of Z hydrogen-atom wave functions,

$$\psi_{n_1l_1m_1}(\vec{r}_1)\updownarrow(S_{z1})\psi_{n_2l_2m_2}(\vec{r}_2)\updownarrow(S_{z2})\ldots\psi_{n_Zl_Zm_Z}(\vec{r}_Z)\updownarrow(S_{zZ})$$

and the total energy is the sum of all the corresponding hydrogen atom energy levels,

$$E_{n_1} + E_{n_2} + \ldots + E_{n_Z}.$$

This solves the Hamiltonian eigenvalue problem under the shielding approximation. The bottom line is: just multiply Z hydrogen energy eigenfunctions together to get an energy eigenfunction for an heavier atom. The energy is the sum of the Z hydrogen energy levels. However, the electrons are identical fermions, so different eigenfunctions must still be combined together in Slater determinants to satisfy the antisymmetrization requirements for electron exchange, as discussed in section 5.7. That will be done during the discussion of the different atoms that is next.

Key Points

- The Hamiltonian eigenvalue problem is too difficult to solve analytically.
- To simplify the problem, the detailed interactions between electrons are ignored. For each electron, it is assumed that the only effect of the other electrons is to cancel, or "shield," that many protons in the nucleus, leaving only a hydrogen nucleus strength.
- This is a very crude approximation.
- It implies that the Z-electron wave functions are products of the single-electron hydrogen atom wave functions. Their energy is the sum of the corresponding single-electron hydrogen energy levels.
- These wave functions must still be combined together to satisfy the antisymmetrization requirement (Pauli exclusion principle).

5.9.3 Hydrogen and helium

This subsection starts off the discussion of the approximate ground states of the elements. Atomic number $Z = 1$ corresponds to hydrogen, which was already discussed in chapter 4.3. The lowest energy state, or ground state, is ψ_{100}, (4.40), also called the "1s" state, and the single electron can be in the spin-up or spin-down versions of that state, or in any combination of the two. The most general ground state wave function is therefore:

$$\begin{aligned}\Psi(\vec{r}_1, S_{z1}) &= a_1\psi_{100}(\vec{r}_1)\uparrow(S_{z1}) + a_2\psi_{100}(\vec{r}_1)\downarrow(S_{z1}) \\ &= \psi_{100}(\vec{r}_1)\Big(a_1\uparrow(S_{z1}) + a_2\downarrow(S_{z1})\Big)\end{aligned}$$

The "ionization energy" that would be needed to remove the electron from the atom is the absolute value of the energy eigenvalue E_1, or 13.6 eV, as derived in chapter 4.3.

For helium, with $Z = 2$, in the ground state both electrons are in the lowest possible energy state ψ_{100}. But since electrons are identical fermions, the antisymmetrization requirement now rears its head. It requires that the two states $\psi_{100}(\vec{r})\uparrow(S_z)$ and $\psi_{100}(\vec{r})\downarrow(S_z)$ appear together in the form of a Slater determinant (chapter 5.7):

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) = \frac{a}{\sqrt{2}} \begin{vmatrix} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) \end{vmatrix} \tag{5.35}$$

or, writing out the Slater determinant:

$$a\psi_{100}(\vec{r}_1)\psi_{100}(\vec{r}_2)\frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}.$$

The spatial part is symmetric with respect to exchange of the two electrons. The spin state is antisymmetric; it is the singlet configuration with zero net spin of section 5.5.6.

Figure 5.4 shows the approximate probability density for the first two elements, indicating where electrons are most likely to be found. In reality, the shielding approximation underestimates the nuclear attraction and the shown helium atom is much too big.

It is good to remember that the $\psi_{100}\uparrow$ and $\psi_{100}\downarrow$ states are commonly indicated as the "K shell" after the first initial of the airline of the Netherlands.

The analysis predicts that the ionization energy to remove *one* electron from helium would be 13.6 eV, the same as for the hydrogen atom. This is a very bad approximation indeed; the truth is almost double, 24.6 eV.

The problem is the made assumption that the repulsion by the other electron "shields" one of the two protons in the helium nucleus, so that only a single-proton hydrogen nucleus is seen. When electron wave functions overlap

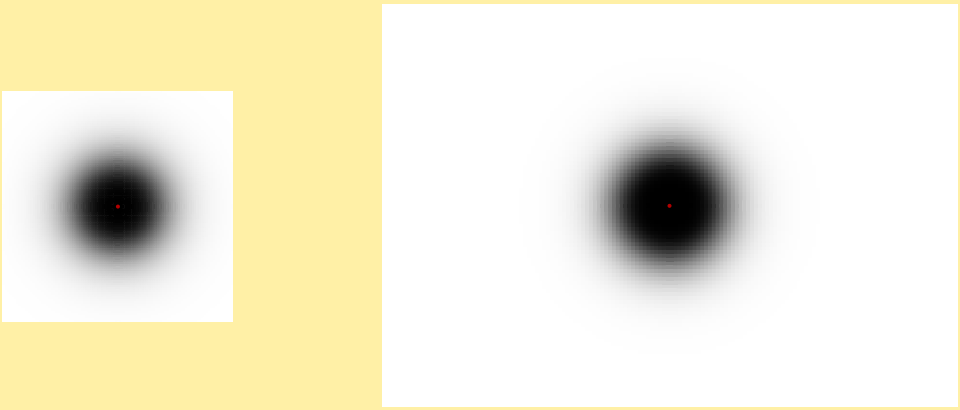


Figure 5.4: Approximate solutions for the hydrogen (left) and helium (right) atoms.

significantly as they do here, their mutual repulsion is a lot less than you would naively expect, (compare figure 13.7). As a result, the second proton is only partly shielded, and the electron is held much more tightly than the analysis predicts. See addendum {A.38.2} for better estimates of the helium atom size and ionization energy.

However, despite the inaccuracy of the approximation chosen, it is probably best to stay consistent, and not fool around at random. It must just be accepted that the theoretical energy levels will be too small in magnitude {N.7}.

The large ionization energy of helium is one reason that it is chemically inert. Helium is called a "noble" gas, presumably because nobody expects nobility to do anything.

Key Points

- The ground states of the atoms of the elements are to be discussed.
- Element one is hydrogen, solved before. Its ground state is ψ_{100} with arbitrary spin. Its ionization energy is 13.6 eV.
- Element two is helium. Its ground state has both electrons in the lowest-energy spatial state ψ_{100}, and locked into the singlet spin state. Its ionization energy is 24.6 eV.
- The large ionization energy of helium means it holds onto its two electrons tightly. Helium is an inert noble gas.
- The two "1s" states $\psi_{100}\uparrow$ and $\psi_{100}\downarrow$ are called the "K shell."

5.9.4 Lithium to neon

The next element is lithium, with three electrons. This is the first element for which the antisymmetrization requirement forces the theoretical energy to go

above the hydrogen ground state level E_1. The reason is that there is no way to create an antisymmetric wave function for three electrons using only the two lowest energy states $\psi_{100}\uparrow$ and $\psi_{100}\downarrow$. A Slater determinant for three electrons must have three different states. One of the eight $\psi_{2lm}\updownarrow$ states with energy E_2 will have to be thrown into the mix.

This effect of the antisymmetrization requirement, that a new state must become "occupied" every time an electron is added is known as the Pauli exclusion principle. It causes the energy values to become larger and larger as the supply of low energy states runs out.

The transition to the higher energy level E_2 is reflected in the fact that in the so-called "periodic table" of the elements, figure 5.5, lithium starts a new row.

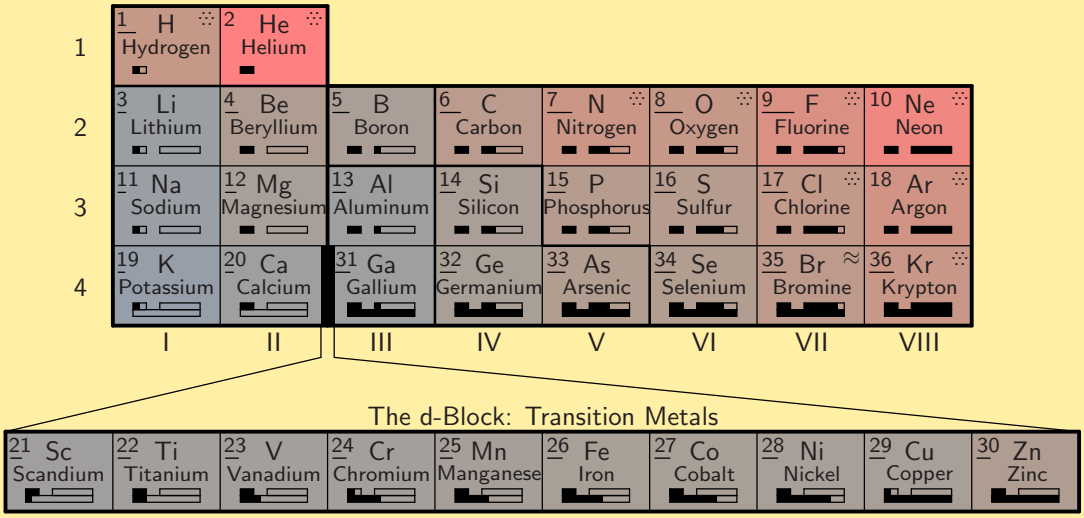


Figure 5.5: Abbreviated periodic table of the elements. Boxes below the element names indicate the quantum states being filled with electrons in that row. Cell color indicates ionization energy. The length of a bar below an atomic number indicates electronegativity. A dot pattern indicates that the element is a gas under normal conditions and wavy lines a liquid.

For the third electron of the lithium atom, the available states with theoretical energy E_2 are the $\psi_{200}\updownarrow$ "2s" states and the $\psi_{211}\updownarrow$, $\psi_{210}\updownarrow$, and $\psi_{21-1}\updownarrow$ "2p" states, a total of eight possible states. These states are, of course, commonly called the "L shell."

Within the crude nuclear shielding approximation made, all eight states have the same energy. However, on closer examination, the spherically symmetric 2s states really have less energy than the 2p ones. Very close to the nucleus, shielding is not a factor and the full attractive nuclear force is felt. So a state in which the electron is more likely to be close to the nucleus has less energy. Those are the 2s states; in the 2p states, which have nonzero orbital angular

momentum, the electron tends to stay away from the immediate vicinity of the nucleus {N.8}.

Within the assumptions made, there is no preference with regard to the spin direction of the 2s state, allowing two Slater determinants to be formed.

$$\frac{a_1}{\sqrt{6}} \begin{vmatrix} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) & \psi_{200}(\vec{r}_1)\uparrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) & \psi_{200}(\vec{r}_2)\uparrow(S_{z2}) \\ \psi_{100}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{100}(\vec{r}_3)\downarrow(S_{z3}) & \psi_{200}(\vec{r}_3)\uparrow(S_{z3}) \end{vmatrix}$$

$$+ \frac{a_2}{\sqrt{6}} \begin{vmatrix} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) & \psi_{200}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) & \psi_{200}(\vec{r}_2)\downarrow(S_{z2}) \\ \psi_{100}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{100}(\vec{r}_3)\downarrow(S_{z3}) & \psi_{200}(\vec{r}_3)\downarrow(S_{z3}) \end{vmatrix} \qquad (5.36)$$

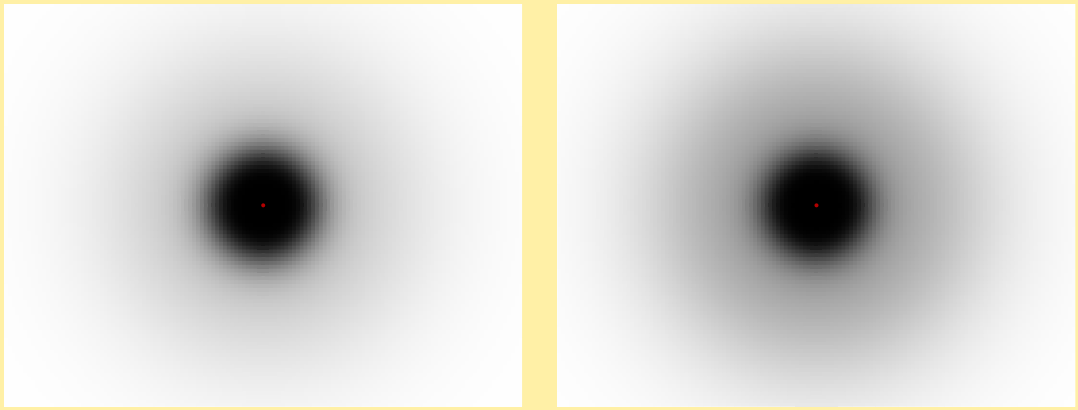


Figure 5.6: Approximate solutions for lithium (left) and beryllium (right).

It is common to say that the "third electron goes into a ψ_{200}" state. Of course that is not quite precise; the Slater determinants above have the first two electrons in ψ_{200} states too. But the third electron adds the third state to the mix, so in that sense it more or less "owns" the state. For the same reason, the Pauli exclusion principle is commonly phrased as "no two electrons may occupy the same state", even though the Slater determinants imply that all electrons share all states equally.

Since the third electron is bound with the much lower energy $|E_2|$ instead of $|E_1|$, it is rather easily given up. Despite the fact that the lithium ion has a nucleus that is 50% stronger than the one of helium, it only takes a ionization energy of 5.4 eV to remove an electron from lithium, versus 24.6 eV for helium. The theory would predict a ionization energy $|E_2| = 3.4$ eV for lithium, which is close, so it appears that the two 1s electrons shield their protons quite well from the 2s one. This is in fact what one would expect, since the 1s electrons are quite close to the nucleus compared to the large radial extent of the 2s state.

Lithium will readily give up its loosely bound third electron in chemical reactions. Conversely, helium would have even less hold on a third electron

than lithium, because it has only two protons in its nucleus. Helium simply does not have what it takes to seduce an electron away from another atom. This is the second part of the reason that helium is chemically inert: it neither will give up its electrons nor take on additional ones.

Thus the Pauli exclusion principle causes different elements to behave chemically in very different ways. Even elements that are just one unit apart in atomic number such as helium (inert) and lithium (very active).

For beryllium, with four electrons, the same four states as for lithium combine in a single 4×4 Slater determinant;

$$\frac{a}{\sqrt{24}} \begin{vmatrix} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) & \psi_{200}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{200}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) & \psi_{200}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{200}(\vec{r}_2)\downarrow(S_{z2}) \\ \psi_{100}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{100}(\vec{r}_3)\downarrow(S_{z3}) & \psi_{200}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{200}(\vec{r}_3)\downarrow(S_{z3}) \\ \psi_{100}(\vec{r}_4)\uparrow(S_{z4}) & \psi_{100}(\vec{r}_4)\downarrow(S_{z4}) & \psi_{200}(\vec{r}_4)\uparrow(S_{z4}) & \psi_{200}(\vec{r}_4)\downarrow(S_{z4}) \end{vmatrix} \tag{5.37}$$

The ionization energy jumps up to 9.3 eV, due to the increased nuclear strength and the fact that the fellow 2s electron does not shield its proton as well as the two 1s electrons do theirs.

For boron, one of the ψ_{21m} "2p" states will need to be occupied. Within the approximations made, there is no preference for any particular state. As an example, figure 5.7 shows the approximate solution in which the ψ_{210}, or "$2p_z$" state is occupied. It may be recalled from figure 4.11 that this state remains close to the z-axis (which is horizontal in the figure.) As a result, the wave function becomes directional. The ionization energy decreases a bit to 8.3 eV,

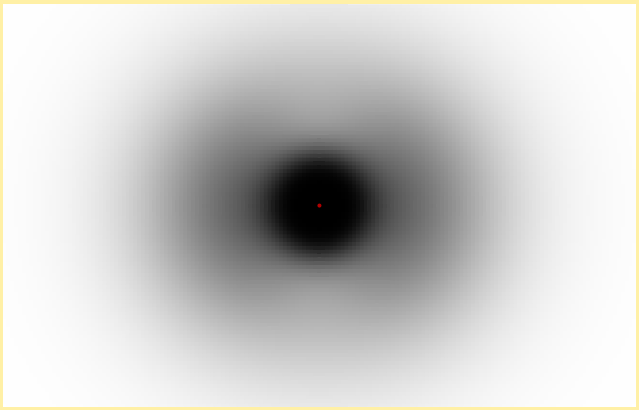


Figure 5.7: Example approximate solution for boron.

indicating that indeed the 2p states have higher energy than the 2s ones.

For carbon, a second ψ_{21m} state needs to be occupied. Within the made approximations, the second 2p electron could also go into the $2p_z$ state. However, in reality, repulsion by the electron already in the $2p_z$ state makes it preferable for the new electron to stay away from the z-axis, which it can do by going

into say the $2p_x$ state. This state is around the vertical x-axis instead of the horizontal z-axis. As noted in chapter 4.3, $2p_x$ is a ψ_{21m} combination state.

For nitrogen, the third 2p electron can go into the $2p_y$ state, which is around the y-axis. There are now three 2p electrons, each in a different spatial state.

However, for oxygen the game is up. There are no more free spatial states in the L shell. The new electron will have to go, say, into the p_y state, pairing up with the electron already there in an opposite-spin singlet state. The repulsion by the fellow electron in the same state reflects in an decrease in ionization energy compared to nitrogen.

For fluorine, the next electron goes into the $2p_x$ state, leaving only the $2p_z$ state unpaired.

For neon, all 2p electrons are paired, and the L shell is full. This makes neon an inert noble gas like helium: it cannot accommodate any more electrons at the E_2 energy level, and, with the strongest nucleus among the L-shell elements, it holds tightly onto the electrons it has.

On the other hand, the previous element, fluorine, has a nucleus that is almost as strong, and it can accommodate an additional electron in its unpaired $2p_z$ state. So fluorine is very willing to steal an electron if it can get away with it. The capability to draw electrons from other elements is called "electronegativity," and fluorine is the most electronegative of them all.

Neighboring elements oxygen and nitrogen are less electronegative, but oxygen can accommodate two additional electrons rather than one, and nitrogen will even accommodate three.

Key Points

- The Pauli exclusion principle forces states of higher energy to become occupied when the number of electrons increases. This raises the energy levels greatly above what they would be otherwise.
- With the third element, lithium, one of the $\psi_{200}\updownarrow$ "2s" states becomes occupied. Because of the higher energy of those states, the third electron is readily given up; the ionization energy is only 5.4 eV.
- Conversely, helium will not take on a third electron.
- The fourth element is beryllium, with both 2s states occupied.
- For boron, carbon, nitrogen, oxygen, fluorine, and neon, the successive ψ_{21m} "2p" states become occupied.
- Neon is a noble gas like helium: it holds onto its electrons tightly, and will not accommodate any additional electrons since they would have to enter the E_3 energy level states.
- Fluorine, oxygen, and nitrogen, however, are very willing to accommodate additional electrons in their vacant 2p states.
- The eight states $\psi_{2lm}\updownarrow$ are called the "L shell."

5.9.5 Sodium to argon

Starting with sodium (natrium), the E_3, or "M shell" begins to be filled. Sodium has a single 3s electron in the outermost shell, which makes it much like lithium, with a single 2s electron in its outermost shell. Since the outermost electrons are the critical ones in chemical behavior, sodium is chemically much like lithium. Both are metals with a "valence" of one; they are willing to sacrifice one electron.

Similarly, the elements following sodium in the third row of the periodic figure 5.5 mirror the corresponding elements in the previous row. Near the end of the row, the elements are again eager to accept additional electrons in the still vacant 3p states.

Finally argon, with no 3s and 3p vacancies left, is again inert. This is actually somewhat of a surprise, because the E_3 M-shell also includes 10 $\psi_{32m}\updownarrow$ states. These states of increased angular momentum are called the "3d" states. (What else?) According to the approximations made, the 3s, 3p, and 3d states would all have the same energy. So it might seem that argon could accept additional electrons into the 3d states.

But it was already noted that the p states in reality have more energy than the s states at the same theoretical energy level, and the d states have even more. The reason is the same: the d states stay even further away from the nucleus than the p states. Because of the higher energy of the d states, argon is really not willing to accept additional electrons.

Key Points

- The next eight elements mirror the properties of the previous eight, from the metal sodium to the highly electronegative chlorine and the noble gas argon.
- The states $\psi_{3lm}\updownarrow$ are called the "M shell."

5.9.6 Potassium to krypton

The logical continuation of the story so far would be that the potassium (kalium) atom would be the first one to put an electron into a 3d state. However, by now the shielding approximation starts to fail not just quantitatively, but qualitatively. The 3d states actually have so much more energy than the 3s states that they even exceed the energy of the 4s states. Potassium puts its last electron into a 4s state, not a 3d one. This makes its outer shell much like the ones of lithium and sodium, so it starts a new row in the periodic table.

The next element, calcium, fills the 4s shell, putting an end to that game. Since the six 4p states have more energy, the next ten elements now start filling the skipped 3d states with electrons, leaving the N-shell with 2 electrons in it. (Actually, this is not quite precise; the 3d and 4s energies are closely together,

and for chromium and copper one of the two 4s electrons turns out to switch to a 3d state.) In any case, it takes until gallium until the six 4p states start filling, which is fully accomplished at krypton. Krypton is again a noble gas, though it can form a weak bond with chlorine.

Continuing to still heavier elements, the energy levels get even more confused. This discussion will stop while it is still ahead.

Key Points

- Unlike what the approximate theory says, in real life the 4s states $\psi_{400}\updownarrow$ have less energy than the $\psi_{32m}\updownarrow$ 3d states, and are filled first.
- After that, the transition metals fill the skipped 3d states before the old logic resumes.
- The states $\psi_{4lm}\updownarrow$ are called the "N shell." It all spells KLM Netherlands.
- The substates are of course called "s," "p," "d," "f," ...

5.9.7 Full periodic table

A complete periodic table of the elements is shown in figure 5.8. The number in the top left corner of each cell is the atomic number Z of the element. The numbers to the left of the table indicate the periods. The length of the periods expands from 2 to 8 elements in period 2 when $l = 2$, p, states must be filled with electrons. Then in period 4, a delayed filling of $l = 2$, d, states expands the periods by another 10 elements. Finally, in period 6, a delayed filling of $l = 3$, f, states adds another 14 elements per period.

The top part of the shown table is called the main group. For some reason however, hydrogen is not included in this term. Note that compared to the previous abbreviated periodic table, hydrogen and helium have been moved to the final columns. The idea is to combine elements with similar properties together into the same columns. Helium is a noble gas like the group VIII elements with filled electron shells. Helium has absolutely nothing in common with the group II alkaline metals that have two electrons in an otherwise empty shell. Similarly, hydrogen behaves much more like a halogen with 1 electron missing from a filled shell than like an alkali metal with 1 electron in an otherwise empty shell. However, hydrogen is still sufficiently different that it should not be considered an actual halogen.

The elements in the periodic table are classified as metals, metalloids, and nonmetals. Metalloids have chemical properties intermediate between metals and nonmetals. The band of metalloids is indicated by dark red cell boundaries in figure 5.8. It extends from boron to polonium. The metals are found to the

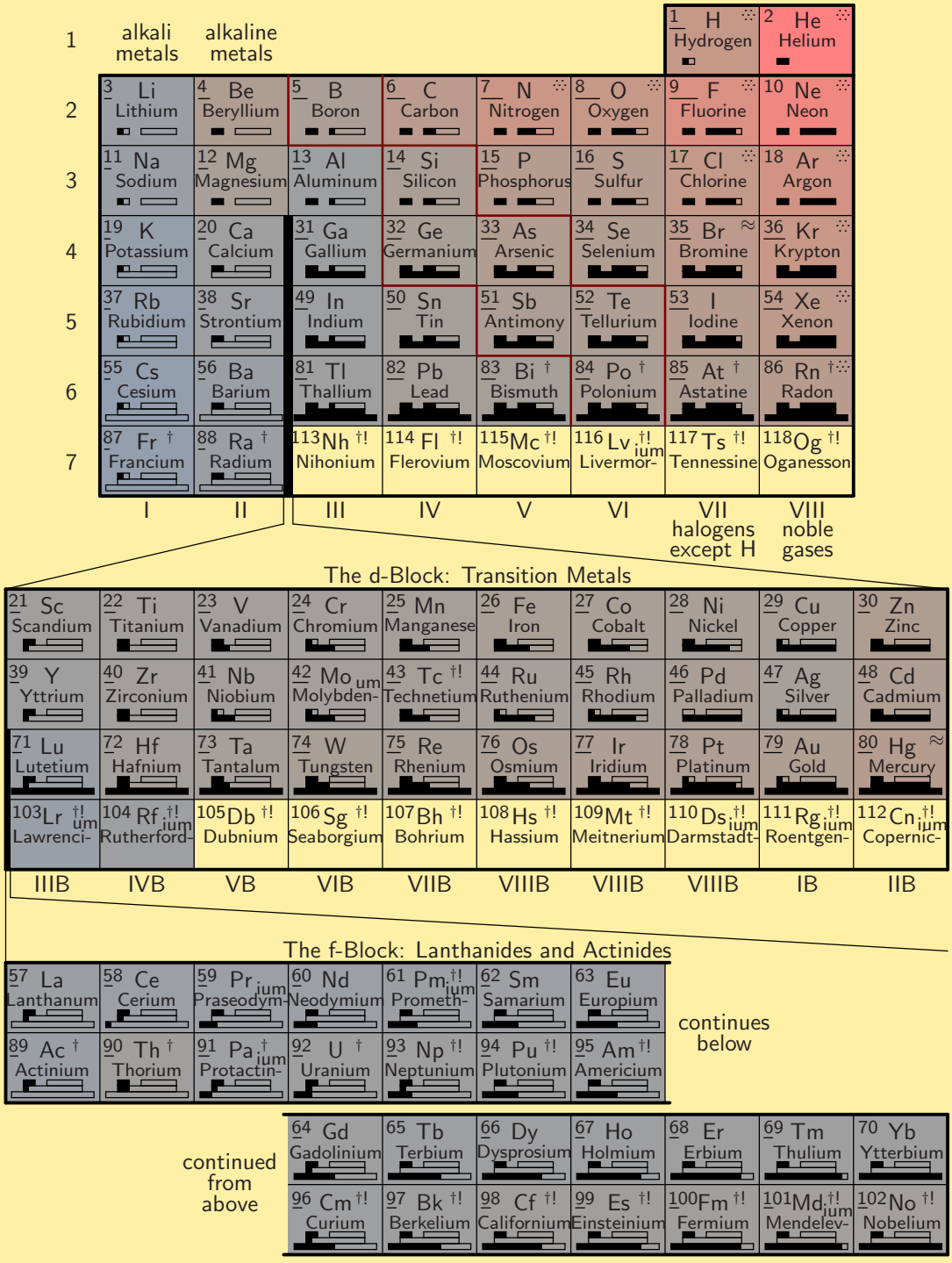


Figure 5.8: Periodic table of the elements. Cell color indicates ionization energy. Boxes indicate the outer electron structure. See the text for more information. [pdf]

left of this band and the nonmetals to the right. Hydrogen and helium are most definitely nonmetals and their shown position in the table reflects that.

The color of each cell indicates the ionization energy, increasing from bluish to reddish. The length of the bar below the atomic number gives the electronegativity. In the top right corner wavy lines indicate that the element is a liquid under normal conditions, and dots that it is a gas. A dagger indicates that the atomic nucleus is radioactive (for every isotope, chapter 14). If the dagger is followed by an exclamation mark, the radioactivity causes the nucleus to decay fast enough that there are no usable quantities of the element found in nature. These elements must be artificially prepared in a lab.

The boxes below the element names indicate the s, p, d, and f shells being filled in that period of the table. The shells already filled in the noble gas at the end of the previous period remain filled and are not shown. Note that the filling of nd states is delayed one period, to period $n + 1$, and the filing of nf states is delayed two periods, to period $n + 2$.

Besides element name, symbol, and radioactivity, periodic table figure 5.8 limits itself to data for which the periodic table arrangement is meaningful. Many other periodic tables also list the average atomic mass for the isotopic composition found on earth. However, for purposes of understanding atomic masses physically, graphs in chapter 14 on nuclei, like figures 14.2 and 14.4, are much more useful.

It should be noted that periodic table figure 5.8 deviates in a number of aspects from the normal conventions. Figure 5.8 is what seems the simplest and most logical. If you put historical oddities and a few committees in charge, you get something different.

Most prominently, most periodic tables leave hydrogen in group I instead of moving it to the top of group VII. But if you move helium to group VIII because of its similarity with the other noble gases in that group, then it is ludicrous to leave hydrogen in group I. Hydrogen has virtually nothing in common with the alkali metals in group I. Like the light halogens, it is a diatomic gas under normal conditions, not a solid metal. Even at the extremely low temperatures at which hydrogen solidifies, it is a nonconducting molecular solid, not a metal. The melting, boiling, and critical points of hydrogen form a logical sequence with those of the halogens. They are totally inconsistent with those of the alkali metals. Hydrogen has the ionization energy of oxygen and the electronegativity of phosphorus. A ionic compound like NaH is a direct equivalent of NaCl, salt, with the hydrogen as the negative ion.

It is true that hydrogen can also form positive ions in chemical reactions, more or less, something that the halogens simply do not do. But do not actually expect to find bare protons when other atoms are around. Also the ionization energy and electronegativity of hydrogen are quite a bit out of line with those of the other halogens. Hydrogen is certainly not a true halogen. But if you order the elements by properties, there is no doubt that hydrogen belongs in group

VII, not I. If you want to refer to the quantum-mechanical shell structure, the term "s block" can still be used to indicate the alkali and alkaline metals along with hydrogen and helium. The remainder of the main group is the "p block." These names indicate the quantum states being filled.

The term transition metals may not include the elements in group IIB of the d-block, for reason related to the fact that their s and d shells have been completely filled. The f-block elements are sometimes referred to as the inner transition metals.

Further, according to the 2005 IUPAC Red Book the lanthanides and actinides should be more properly called the lanthanoids and actinoids, since "ide" usually means negative ion. Since "oid" means "-like," according to IUPAC the lanthanoids should not really include lanthanum, and the actinoids should not include actinium. However, the IUPAC does include them because of common usage. A rare triumph of scientific common sense over lousy terminology. If lanthanum and actinium are to be included, the lanthanides and actinides should of course simply have been renamed the lanthanum and actinium groups, or equivalent, not lanthanoids and actinoids.

More significantly, unlike figure 5.8 suggests, lutetium is included in the lanthanoids and lawrencium in the actinoids. The term rare-earth metals include the lanthanoids, as well as scandium and yttrium as found immediately above lutetium.

Also, both lutetium and lawrencium are according to IUPAC included in the f-block. That makes the f-block 15 columns wide instead of the 14 column block shown at the bottom of figure 5.8. Of course, that does not make any sense at all. The name f-block supposedly indicates that an f-shell is being filled. An f-shell holds 14 electrons, not 15. For lutetium, the f-shell is full and other shells have begun to fill. The same is, at the time of writing, believed to be true for lawrencium. And while the first f-shell electrons for lanthanum and actinium get *temporarily* bumped to the d-shell, that is obviously a minor error in the overall logic of filling the f-shell. (Apparently, there is a long-standing controversy whether lanthanum and actinium or lutetium and lawrencium should be included in the f-block. By compromising and putting both in the f-block of their 2007 periodic table, the IUPAC got the worst of both worlds.)

A nice recent example of a more conventional periodic table by an authoritative source is from NIST[1]. This also includes the latest updates on various data, unlike the periodic table in this book. An earlier version can be found at the web location[2] of this document. The hieroglyphs found in the NIST table are explained in chapter 10.7.1.

Periodic table figure 5.8 was based on data from various sources. Shell fillings and ionization energies agree with the NIST listing and table. The uncertain

[1] `https://www.nist.gov/pml/periodic-table-elements`

[2] `http://www.dommelen.net/quansup/periodic-table.pdf`

shell fillings at atomic numbers 103 and 104 were left out. The classification whether the elements must be artificially prepared was taken from the NIST periodic table. The electronegativities are based on the Pauling scale. They were taken from Wikipedia "use" values, that were in turn taken from WebElements, and are mostly the same as those in the 2003 CRC Handbook of Chemistry and Physics, and the 1999 Lange's Handbook of Chemistry. Discrepancies between these sources of more than 10% occur for atomic numbers 71, 74, 82, and 92.

5.10 Pauli Repulsion

Before proceeding to a description of chemical bonds, one important point must first be made. While the earlier descriptions of the hydrogen molecular ion and hydrogen molecule produced many important observations about chemical bonds, they are highly misleading in one aspect.

In the hydrogen molecule cases, the repulsive force that eventually stops the atoms from getting together any closer than they do is the electrostatic repulsion between the nuclei. It is important to recognize that this is the exception, rather than the norm. Normally, the main repulsion between atoms is not due to repulsion between the nuclei, but due to the Pauli exclusion principle for their electrons. Such repulsion is called "exclusion-principle repulsion" or "Pauli repulsion."

To understand why the repulsion arises, consider two helium ions, and assume that you put them right on top of each other. Of course, with the nuclei right on top of each other, the nuclear repulsion will be infinite, but ignore that for now. There is another effect, and that is the interesting one here. *There are now 4 electrons in the 1s shell.*

Without the Pauli exclusion principle, that would not be a big deal. The repulsion between the electrons would go up, but so would the combined nuclear strength double. However, Pauli says that only two electrons may go into the 1s shell. The other two 1s electrons will have to divert to the 2s shell, and that requires a lot of energy.

Next consider what happens when two helium atoms are not on top of each other, but are merely starting to intrude on each other's 1s shell space. Recall that the Pauli principle is just the antisymmetrization requirement of the electron wave function applied to a description in terms of given energy states. When the atoms get closer together, the energy states get confused, but the antisymmetrization requirement stays in full force. When the filled shells start to intrude on each other's space, the electrons start to divert to increasingly higher energy to continue to satisfy the antisymmetrization requirement. This process ramps up much more quickly than the nuclear repulsions and dominates the net repulsion in almost all circumstances.

In everyday terms, the standard example of repulsion forces that ramp up very quickly is billiard balls. If billiard balls are a millimeter away from touching, there is no repulsion between them, but move them closer a millimeter, and suddenly there is this big repulsive force. The repulsion between filled atom shells does not ramp up that quickly in relative terms, of course, but it does ramp up quickly. So describing atoms with closed shells as billiard balls is quite reasonable if you are just looking for a general idea.

Key Points

- If electron wave functions intrude on each others space, it can cause repulsion due to the antisymmetrization requirement.
- This is called Pauli repulsion or exclusion principle repulsion.
- It is the dominant repulsion in almost all cases.

5.11 Chemical Bonds

The electron states, or "atomic orbitals", of the elements discussed in section 5.9 form the basis for the "valence bond" description of chemical bonds. This section summarizes some of the basic ideas involved.

5.11.1 Covalent sigma bonds

As pointed out in section 5.9, helium is chemically inert: its outermost, and only, shell can hold two electrons, and it is full. But hydrogen has only one electron, leaving a vacant position for another 1s electron. As discussed earlier in chapter 5.2, two hydrogen atoms are willing to *share* their electrons. This gives each atom in some sense two electrons in its shell, filling it up. The shared state has lower energy than the two separate atoms, so the H_2 molecule stays together. A sketch of the shared 1s electrons was given in figure 5.2.

Fluorine has one vacant spot for an electron in its outer shell just like hydrogen; its outer shell can contain 8 electrons and fluorine has only seven. One of its 2p states, assume it is the horizontal axial state $2p_z$, has only one electron in it instead of two. Two fluorine atoms can share their unpaired electrons much like hydrogen atoms do and form an F_2 molecule. This gives each of the two atoms a filled shell. The fluorine molecular bond is sketched in figure 5.9 (all other electrons have been omitted.) This bond between p electrons looks quite different from the H_2 bond between s electrons in figure 5.2, but it is again a covalent one, in which the electrons are shared. In addition, both bonds are called "sigma" bonds: if you look at either bond *from the side,* it looks rotationally symmetric, just like an s state. (Sigma is the Greek equivalent of the letter s; it is written as σ.)

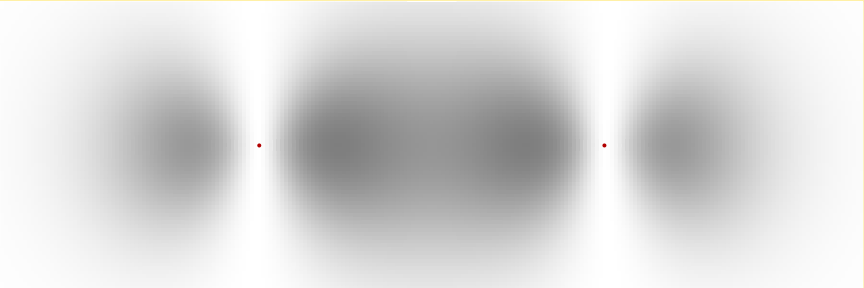


Figure 5.9: Covalent sigma bond consisting of two $2p_z$ states.

Key Points

- Two fluorine or similar atoms can share their unpaired 2p electrons in much the same way that two hydrogen atoms can share their unpaired 2s electrons.
- Since such bonds look like s states when seen from the side, they are called sigma or σ bonds.

5.11.2 Covalent pi bonds

The N_2 nitrogen molecule is another case of covalent bonding. Nitrogen atoms have a total of three unpaired electrons, which can be thought of as one each in the $2p_x$, $2p_y$, and $2p_z$ states. Two nitrogen atoms can share their unpaired $2p_z$ electrons in a sigma bond the same way that fluorine does, longitudinally.

However, the $2p_x$ and $2p_y$ states are normal to the line through the nuclei; these states must be matched up sideways. Figure 5.10 illustrates this for the bond between the two vertical $2p_x$ states. This covalent bond, and the corresponding one between the $2p_y$ states, looks like a p state when seen from the side, and it is called a "pi" or π bond.

So, the N_2 nitrogen molecule is held together by two pi bonds in addition to a sigma bond, making a triple bond. It is a relatively inert molecule.

Key Points

- Unpaired p states can match up sideways in what are called pi or π bonds.

5.11.3 Polar covalent bonds and hydrogen bonds

Oxygen, located in between fluorine and nitrogen in the periodic table, has two unpaired electrons. It can share these electrons with another oxygen atom to

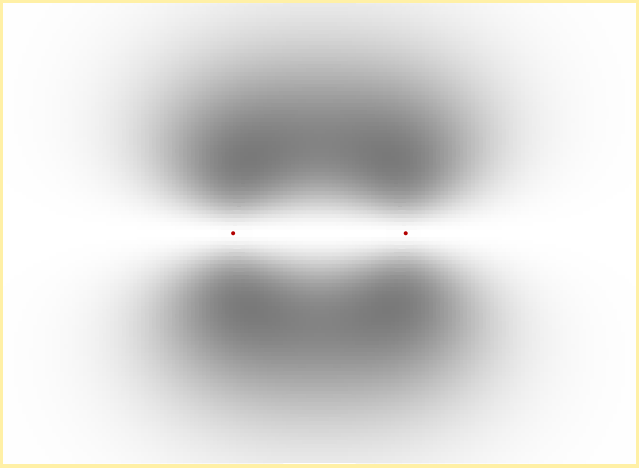


Figure 5.10: Covalent pi bond consisting of two $2p_x$ states.

form O_2, the molecular oxygen we breath. However, it can instead bind with two hydrogen atoms to form H_2O, the water we drink.

In the water molecule, the lone $2p_z$ electron of oxygen is paired with the 1s electron of one hydrogen atom, as shown in figure 5.11. Similarly, the lone $2p_y$

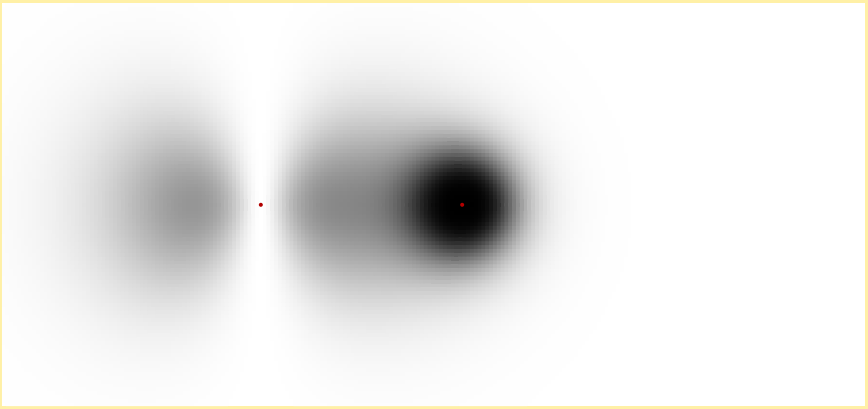


Figure 5.11: Covalent sigma bond consisting of a $2p_z$ and a 1s state.

electron is paired with the 1s electron of the other hydrogen atom. Both bonds are sigma bonds: they are located on the connecting line between the nuclei. But in this case each bond consists of a 1s and a 2p state, rather than two states of the same type.

Since the x and y axes are orthogonal, the two hydrogen atoms in water should be at a 90 degree angle from each other, relative to the oxygen nucleus. (Without valence bond theory, the most logical guess would surely have been that they would be at opposite sides of the oxygen atom.) The predicted 90 degree angle is in fair approximation to the experimental value of 105 degrees.

The reason that the actual angle is a bit more may be understood from the fact that the oxygen atom has a higher attraction for the shared electrons, or electronegativity, than the hydrogen atoms. It will pull the electrons partly away from the hydrogen atoms, giving itself some negative charge, and the hydrogen atoms a corresponding positive one. The positively charged hydrogen atoms repel each other, increasing their angle a bit. If you go down one place in the periodic table below oxygen, to the larger sulfur atom, H_2S has its hydrogen atoms under about 93 degrees, quite close to 90 degrees.

Bonds like the one in water, where the negative electron charge shifts towards the more electronegative atom, are called "polar" covalent bonds.

It has significant consequences for water, since the positively charged hydrogen atoms can electrostatically attract the negatively charged oxygen atoms on *other* molecules. This has the effect of creating bonds between different molecules called "hydrogen bonds." While much weaker than typical covalent bonds, they are strong enough to affect the physical properties of water. For example, they are the reason that water is normally a liquid instead of a gas, quite a good idea if you are thirsty, and that ice floats on water instead of sinking to the bottom of the oceans. Hydrogen is particularly efficient at creating such bonds because it does not have any other electrons to shield its nucleus.

Key Points

- The geometry of the quantum states reflects in the geometry of the formed molecules.
- When the sharing of electrons is unequal, a bond is called polar.
- A special case is hydrogen, which is particularly effective in also creating bonds between different molecules, hydrogen bonds, when polarized.
- Hydrogen bonds give water unusual properties that are critical for life on earth.

5.11.4 Promotion and hybridization

While valence bond theory managed to explain a number of chemical bonds so far, two important additional ingredients need to be added. Otherwise it will not at all be able to explain organic chemistry, the chemistry of carbon critical to life.

Carbon has two unpaired 2p electrons just like oxygen does; the difference between the atoms is that oxygen has in addition two paired 2p electrons. With two unpaired electrons, it might seem that carbon should form two bonds like oxygen.

But that is not what happens; normally carbon forms four bonds instead of two. In chemical bonds, one of carbon's paired 2s electrons moves to the empty 2p state, leaving carbon with four unpaired electrons. It is said that the 2s electron is "promoted" to the 2p state. This requires energy, but the energy gained by having four bonds more than makes up for it.

Promotion explains why a molecule such as CH_4 forms. Including the 4 shared hydrogen electrons, the carbon atom has 8 electrons in its outer shell, so its shell is full. It has made as many bonds as it can support.

However, promotion is still not enough to explain the molecule. If the CH_4 molecule was merely a matter of promoting one of the 2s electrons into the vacant $2p_y$ state, the molecule should have three hydrogen atoms under 90 degrees, sharing the $2p_x$, $2p_y$, and $2p_z$ electrons respectively, and one hydrogen atom elsewhere, sharing the remaining 2s electron. In reality, the CH_4 molecule is shaped like a regular tetrahedron, with angles of 109.5 degrees between all four hydrogens.

The explanation is that, rather than using the $2p_x$, $2p_y$, $2p_z$, and 2s states directly, the carbon atom forms new combinations of the four called "hybrid" states. (This is not unlike how the torus-shaped ψ_{211} and ψ_{21-1} states were recombined in chapter 4.3 to produce the equivalent $2p_x$ and $2p_y$ pointer states.)

In case of CH_4, the carbon converts the 2s, $2p_x$, $2p_y$, and $2p_z$ states into four new states. These are called sp^3 states, since they are formed from one s and three p states. They are given by:

$$|\mathrm{sp}^3_a\rangle = \tfrac{1}{2}(|2\mathrm{s}\rangle + |2\mathrm{p}_x\rangle + |2\mathrm{p}_y\rangle + |2\mathrm{p}_z\rangle)$$

$$|\mathrm{sp}^3_b\rangle = \tfrac{1}{2}(|2\mathrm{s}\rangle + |2\mathrm{p}_x\rangle - |2\mathrm{p}_y\rangle - |2\mathrm{p}_z\rangle)$$

$$|\mathrm{sp}^3_c\rangle = \tfrac{1}{2}(|2\mathrm{s}\rangle - |2\mathrm{p}_x\rangle + |2\mathrm{p}_y\rangle - |2\mathrm{p}_z\rangle)$$

$$|\mathrm{sp}^3_d\rangle = \tfrac{1}{2}(|2\mathrm{s}\rangle - |2\mathrm{p}_x\rangle - |2\mathrm{p}_y\rangle + |2\mathrm{p}_z\rangle)$$

where the kets denote the wave functions of the indicated states.

All four sp^3 hybrids have the same shape, shown in figure 5.12. The asymmetrical shape can increase the overlap between the wave functions in the bond. The four sp^3 hybrids are under equal 109.5 degrees angles from each other, producing the tetrahedral structure of the CH_4 molecule. And of diamond, for that matter. With the atoms bound together in all spatial directions, diamond is an extremely hard material.

But carbon is a very versatile atom. In graphite, and carbon nanotubes, carbon atoms arrange themselves in layers instead of three-dimensional structures. Carbon achieves this trick by leaving the 2p-state in the direction normal to the plane, call it p_x, out of the hybridization. The two 2p states in the plane

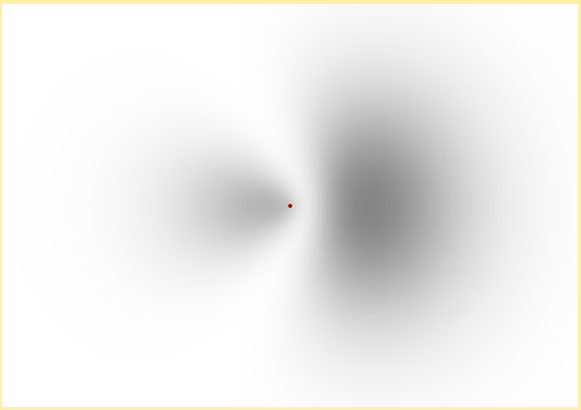


Figure 5.12: Shape of an sp^3 hybrid state.

plus the 2s state can then be combined into three sp^2 states:

$$|\mathrm{sp}_a^2\rangle = \frac{1}{\sqrt{3}}|2\mathrm{s}\rangle + \frac{2}{\sqrt{6}}|2\mathrm{p}_z\rangle$$
$$|\mathrm{sp}_b^2\rangle = \frac{1}{\sqrt{3}}|2\mathrm{s}\rangle - \frac{1}{\sqrt{6}}|2\mathrm{p}_z\rangle + \frac{1}{\sqrt{2}}|2\mathrm{p}_y\rangle$$
$$|\mathrm{sp}_c^2\rangle = \frac{1}{\sqrt{3}}|2\mathrm{s}\rangle - \frac{1}{\sqrt{6}}|2\mathrm{p}_z\rangle - \frac{1}{\sqrt{2}}|2\mathrm{p}_y\rangle$$

Each is shaped as shown in figure 5.13.

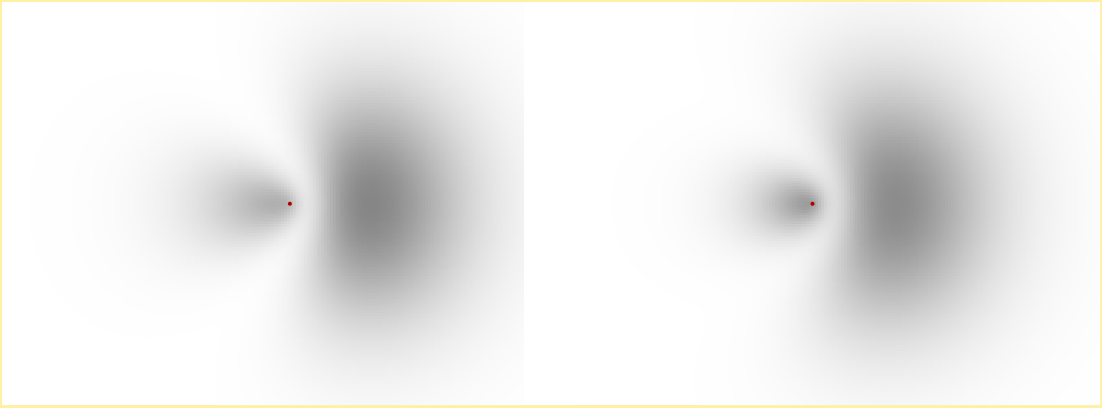


Figure 5.13: Shapes of the sp^2 (left) and sp (right) hybrids.

These planar hybrids are under 120 degree angles from each other, giving graphite its hexagonal structure. The left-out p electrons normal to the plane can form pi bonds with each other. A planar molecule formed using sp^2 hybridization is ethylene (C_2H_4); it has all six nuclei in the same plane. The pi

bond normal to the plane prevents out-of-plane rotation of the nuclei around the line connecting the carbons, keeping the plane rigid.

Finally, carbon can combine the 2s state with a single 2p state to form two sp hybrids under 180 degrees from each other:

$$|\mathrm{sp}_a\rangle = \frac{1}{\sqrt{2}}(|2\mathrm{s}\rangle + |2\mathrm{p}_z\rangle)$$

$$|\mathrm{sp}_b\rangle = \frac{1}{\sqrt{2}}(|2\mathrm{s}\rangle - |2\mathrm{p}_z\rangle)$$

An example sp hybridization is acetylene, (C_2H_2), which has all its four nuclei on a single line.

Key Points

- The chemistry of carbon is critical for life as we know it.
- It involves two additional ideas; one is promotion, where carbon kicks one of its 2s electrons into a 2p state. This gives carbon one 2s and three 2p electrons.
- The second idea is hybridization, where carbon combines these four states in creative new combinations called hybrids.
- In sp^3 hybridization, carbon creates four hybrids in a regular tetrahedron combination.
- In sp^2 hybridization, carbon creates three hybrids in a plane, spaced at 120 degree intervals. That leaves a conventional 2p state in the direction normal to the plane.
- In sp hybridization, carbon creates two hybrids along a line, pointing in opposite directions. That leaves two conventional 2p states normal to the line of the hybrids and to each other.

5.11.5 Ionic bonds

Ionic bonds are the extreme polar bonds; they occur if there is a big difference between the electronegativities of the atoms involved.

An example is kitchen salt, NaCl. The sodium atom has only one electron in its outer shell, a loosely bound 3s one. The chlorine has seven electrons in its outer shell and needs only one more to fill it. When the two react, the chlorine does not just share the lone electron of the sodium atom, it simply takes it away. It makes the chlorine a negatively charged ion. Similarly, it leaves the sodium as a positively charged ion.

The charged ions are bound together by electrostatic forces. Since these forces act in all directions, each ion does not just attract the opposite ion it

exchanged the electron with, but all surrounding opposite ions. And since in salt each sodium ion is surrounded by six chlorine ions and vice versa, the number of bonds that exists is large.

Since so many bonds must be broken to take a ionic substance apart, their properties are quite different from covalently bounded substances. For example, salt is a solid with a high melting point, while the covalently bounded Cl_2 chlorine molecule is normally a gas, since the bonds between different molecules are weak. Indeed, the covalently bound hydrogen molecule that has been discussed much in this chapter remains a gas until especially low cryogenic temperatures.

Chapter 10.2 will give a more quantitative discussion of ionic molecules and solids.

Key Points

- When a bond is so polar that practically speaking one atom takes the electron away from the other, the bond is called ionic.
- Ionic substances like salt tend to form strong solids, unlike typical purely covalently bound molecules like hydrogen that tend to form gases.

5.11.6 Limitations of valence bond theory

Valence bond theory does a terrific job of describing chemical bonds, producing a lot of essentially correct, and very nontrivial predictions, but it does have limitations.

One place it fails is for the O_2 oxygen molecule. In the molecule, the atoms share their unpaired $2p_x$ and $2p_z$ electrons. With all electrons symmetrically paired in the spatial states, the electrons should all be in singlet spin states having no net spin. However, it turns out that oxygen is strongly paramagnetic, indicating that there is in fact net spin. The problem in valence bond theory that causes this error is that it ignores the already paired-up electrons in the $2p_y$ states. In the molecule, the filled $2p_y$ states of the atoms are next to each other and they do interact. In particular, one of the total of four $2p_y$ electrons jumps over to the $2p_x$ states, where it only experiences repulsion by two other electrons instead of by three. The spatial state of the electron that jumps over is no longer equal to that of its twin, allowing them to have equal instead of opposite spin.

Valence bond theory also has problems with single-electron bonds such as the hydrogen molecular ion, or with benzene, in which the carbon atoms are held together with what is essentially 1.5 bonds, or rather, bonds shared as in a two state system. Excited states produce major difficulties. Various fixes and improved theories exist.

Key Points

- Valence bond theory is extremely useful. It is conceptually simple and explains much of the most important chemical bonds.
- However, it does have definite limitations: some types of bonds are not correctly or not at all described by it.
- Little in life is ideal, isn't it?

Chapter 6

Macroscopic Systems

Abstract

Macroscopic systems involve extremely large numbers of particles. Such systems are very hard to analyze exactly in quantum mechanics. An exception is a system of noninteracting particles stuck in a rectangular box. This chapter therefore starts with an examination of that model. For a model of this type, the system energy eigenfunctions are found to be products of single-particle states.

One thing that becomes quickly obvious is that macroscopic system normally involve a gigantic number of single-particle states. It is unrealistic to tabulate them each individually. Instead, average statistics about the states are derived. The primary of these is the so-called density of states. It is the number of single-particle states per unit energy range.

But knowing the number of states is not enough by itself. Information is also needed on how many particles are in these states. Fortunately, it turns out to be possible to derive the average number of particles per state. This number depends on whether it is a system of bosons, like photons, or a system of fermions, like electrons. For bosons, the number of particles is given by the so-called Bose-Einstein distribution, while for electrons it is given by the so-called Fermi-Dirac distribution. Either distribution can be simplified to the so-called Maxwell-Boltzmann distribution under conditions in which the average number of particles per state is much less than one.

Each distribution depends on both the temperature and on a so-called chemical potential. Physically, temperature differences promote the diffusion of thermal energy, heat, from hot to cold. Similarly, differences in chemical potential promote the diffusion of particles from high chemical potential to low.

At first, systems of identical bosons are studied. Bosons behave quite strangely at very low temperatures. Even for a nonzero temperature, a finite fraction of them may stay in the single-particle state of lowest

energy. That behavior is called Bose-Einstein condensation. Bosons also show a lack of low-energy global energy eigenfunctions.

A first discussion of electromagnetic radiation, including light, will be given. The discussed radiation is the one that occurs under conditions of thermal equilibrium, and is called blackbody radiation.

Next, systems of electrons are covered. It is found that electrons in typical macroscopic systems have vast amounts of kinetic energy even at absolute zero temperature. It is this kinetic energy that is responsible for the volume of solids and liquids and their resistance to compression. The electrons are normally confined to a solid despite all their kinetic energy. But at some point, they may escape in a process called thermionic emission.

The electrical conduction of metals can be explained using the simple model of noninteracting electrons. However, electrical insulators and semiconductors cannot. It turns out that these can be explained by including a simple model of the forces on the electrons.

Then semiconductors are discussed, including applications such as diodes, transistors, solar cells, light-emitting diodes, solid state refrigeration, thermocouples, and thermoelectric generators. A somewhat more general discussion of optical issues is also given.

6.1 Intro to Particles in a Box

Since most macroscopic systems are very hard to analyze in quantum-mechanics, simple systems are very important. They allow insight to be achieved that would be hard to obtain otherwise. One of the simplest and most important systems is that of multiple noninteracting particles in a box. For example, it is a starting point for quantum thermodynamics and the quantum description of solids.

It will be assumed that the particles do not interact with each other, nor with anything else in the box. That is a dubious assumption; interactions between particles are essential to achieve statistical equilibrium in thermodynamics. And in solids, interaction with the atomic structure is needed to explain the differences between electrical conductors, semiconductors, and insulators. However, in the box model such effects can be treated as a perturbation. That perturbation is ignored to leading order.

In the absence of interactions between the particles, the possible quantum states, or energy eigenfunctions, of the complete system take a relatively simple form. They turn out to be products of *single particle* energy eigenfunctions. A generic energy eigenfunction for a system of I particles is:

$$\psi^{\rm S}_{\vec{n}_1,\vec{n}_2,\ldots,\vec{n}_i,\ldots,\vec{n}_I}(\vec{r}_1,S_{z1},\vec{r}_2,S_{z2},\ldots,\vec{r}_i,S_{zi},\ldots,\vec{r}_I,S_{zI}) =$$

$$\psi^{\rm p}_{\vec{n}_1}(\vec{r}_1, S_{z1}) \times \psi^{\rm p}_{\vec{n}_2}(\vec{r}_2, S_{z2}) \times \ldots \times \psi^{\rm p}_{\vec{n}_i}(\vec{r}_i, S_{zi}) \times \ldots \times \psi^{\rm p}_{\vec{n}_I}(\vec{r}_I, S_{zI}) \quad (6.1)$$

In such a system eigenfunction, particle number i out of I is in a single-particle energy eigenfunction $\psi^{\rm p}_{\vec{n}_i}(\vec{r}_i, S_{zi})$. Here $\vec{r}_i$ is the position vector of the particle, and S_{zi} its spin in a chosen z-direction. The subscript $\vec{n}_i$ stands for whatever quantum numbers characterize the single-particle eigenfunction. A system wave function of the form above, a simple product of single-particles ones, is called a "Hartree product."

For noninteracting particles confined inside a box, the single-particle energy eigenfunctions, or single-particle states, are essentially the same ones as those derived in chapter 3.5 for a particle in a pipe with a rectangular cross section. However, to account for nonzero particle spin, a spin-dependent factor must be added. In any case, this chapter will not really be concerned that much with the detailed form of the single-particle energy states. The main quantities of interest are their quantum numbers and their energies. Each possible set of quantum numbers will be graphically represented as a point in a so-called "wave number space." The single-particle energy is found to be related to how far that point is away from the origin in that wave number space.

For the complete system of I particles, the most interesting physics has to do with the (anti) symmetrization requirements. In particular, for a system of identical fermions, the Pauli exclusion principle says that there can be at most one fermion in a given single-particle state. That implies that in the above Hartree product each set of quantum numbers $\vec{n}$ must be different from all the others. In other words, any system wave function for a system of I fermions must involve at least I different single-particle states. For a macroscopic number of fermions, that puts a tremendous restriction on the wave function. The most important example of a system of identical fermions is a system of electrons, but systems of protons and of neutrons appear in the description of atomic nuclei.

The antisymmetrization requirement is really more subtle than the Pauli principle implies. And the symmetrization requirements for bosons like photons or helium-4 atoms are nontrivial too. This was discussed earlier in chapter 5.7. Simple Hartree product energy eigenfunctions of the form (6.1) above are not acceptable by themselves; they must be combined with others with the same single-particle states, but with the particles shuffled around between the states. Or rather, because shuffled around sounds too much like Las Vegas, with the particles exchanged between the states.

Key Points

- Systems of noninteracting particles in a box will be studied.
- Interactions between the particles may have to be included at some later stage.
- System energy eigenfunctions are obtained from products of single-particle energy eigenfunctions.

0⊸ (anti) symmetrization requirements further restrict the system energy eigenfunctions.

6.2 The Single-Particle States

As the previous section noted, the objective is to understand systems of non-interacting particles stuck in a closed, impenetrable, box. To do so, the key question is what are the single-particle quantum states, or energy eigenfunctions, for the particles. They will be discussed in this section.

The box will be taken to be rectangular, with its sides aligned with the coordinate axes. The lengths of the sides of the box will be indicated by ℓ_x, ℓ_y, and ℓ_z respectively.

The single-particle energy eigenfunctions for such a box were derived in chapter 3.5 under the guise of a pipe with a rectangular cross section. The single-particle energy eigenfunctions are:

$$\boxed{\psi^{\rm p}_{n_x n_y n_z}(\vec{r}) = \sqrt{\frac{8}{\mathcal{V}}}\,\sin(k_x x)\sin(k_y y)\sin(k_z z)} \tag{6.2}$$

Here $\mathcal{V} = \ell_x \ell_y \ell_z$ is the volume of the box. The "wave numbers" k_x, k_y, and k_z take the values:

$$\boxed{k_x = n_x \frac{\pi}{\ell_x} \quad k_y = n_y \frac{\pi}{\ell_y} \quad k_z = n_z \frac{\pi}{\ell_z}} \tag{6.3}$$

where n_x, n_y, and n_z are natural numbers. Each set of three natural numbers n_x, n_y, n_z gives one single-particle eigenfunction. In particular, the single-particle eigenfunction of lowest energy is $\psi^{\rm p}_{111}$, having $n_x = n_y = n_z = 1$.

However, the precise form of the eigenfunctions is not really that important here. What is important is how many there are and what energy they have. That information can be summarized by plotting the allowed wave numbers in a k_x, k_y, k_z axis system. Such a plot is shown in the left half of figure 6.1.

Each point in this "wave number space" corresponds to one spatial single-particle state. The coordinates k_x, k_y, and k_z give the wave numbers in the three spatial directions. In addition, the distance k from the origin indicates the single-particle energy. More precisely, the single particle energy is

$$\boxed{E^{\rm p} = \frac{\hbar^2}{2m}k^2 \qquad k \equiv \sqrt{k_x^2 + k_y^2 + k_z^2}} \tag{6.4}$$

The energy is therefore just a constant times the square of this distance. (The above expression for the energy can be verified by applying the kinetic energy operator on the given single-particle wave function.)

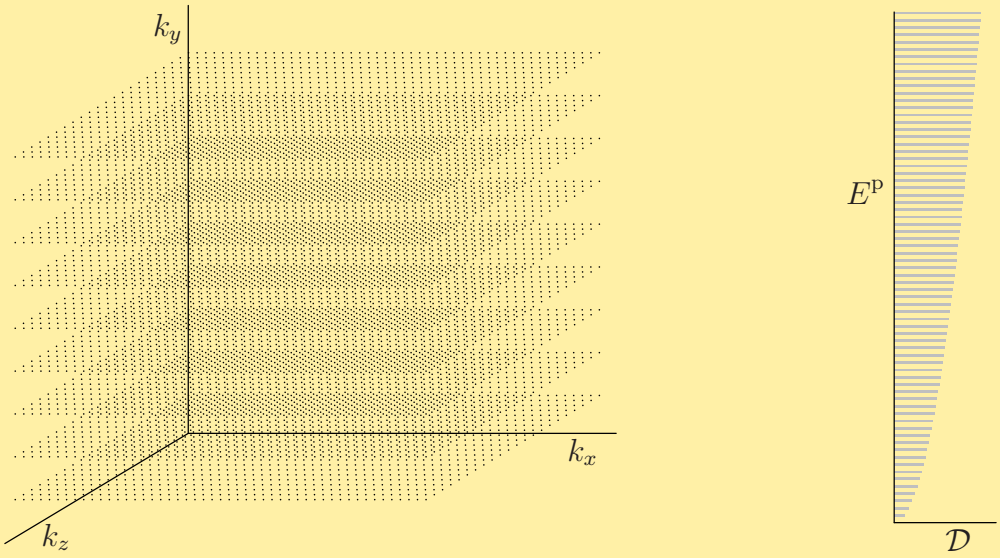


Figure 6.1: Allowed wave number vectors, left, and energy spectrum, right.

One more point must be made. The single-particle energy eigenfunctions described above are *spatial* states. Particles with nonzero spin, which includes all fermions, can additionally have different spin in whatever is chosen to be the z-direction. In particular, for fermions with spin $\frac{1}{2}$, including electrons, there is a "spin-up" and a "spin-down" version of each spatial energy eigenfunction:

$$\psi^{\rm p}_{n_x n_y n_z, \frac{1}{2}}(\vec{r}, S_z) = \sqrt{\frac{8}{\mathcal{V}}} \sin(k_x x) \sin(k_y y) \sin(k_z z) \uparrow(S_z)$$

$$\psi^{\rm p}_{n_x n_y n_z, -\frac{1}{2}}(\vec{r}, S_z) = \sqrt{\frac{8}{\mathcal{V}}} \sin(k_x x) \sin(k_y y) \sin(k_z z) \downarrow(S_z)$$

That means that each point in the wave number space figure 6.1 stands for *two* single-particle states, not just one.

In general, if the particles have spin s, each point in wave number space corresponds to $2s + 1$ different single-particle states. However, photons are an exception to this rule. Photons have spin $s = 1$ but each spatial state corresponds to only 2 single-particle states, not 3. (That is related to the fact that the spin angular momentum of a photon in the direction of motion can only be $\hbar$ or $-\hbar$, not 0. And that is in turn related to the fact that the electromagnetic field cannot have a component in the direction of motion. If you are curious, see addendum {A.21.6} for more.)

Key Points

- Each single particle state is characterized by a set of three "wave numbers" k_x, k_y, and k_z.

- Each point in the "wave number space" figure 6.1 corresponds to one specific spatial single-particle state.
- The distance of the point from the origin is a measure of the energy of the single-particle state.
- In the presence of nonzero particle spin s, each point in wave number space corresponds to $2s+1$ separate single-particle states that differ in the spin in the chosen z-direction. For photons, make that $2s$ instead of $2s+1$.

6.3 Density of States

Up to this point, this book has presented energy levels in the form of an energy spectrum. In these spectra, each single-particle energy was shown as a tick mark along the energy axis. The single-particle states with that energy were usually listed next to the tick marks. One example was the energy spectrum of the electron in a hydrogen atom as shown in figure 4.8.

However, the number of states involved in a typical macroscopic system can easily be of the order of 10^{20} or more. There is no way to show anywhere near that many energy levels in a graph. Even if printing technology was up to it, and it can only dream about it, your eyes would have only about $7\ 10^6$ cones and $1.3\ 10^8$ rods to see them.

For almost all practical purposes, the energy levels of a macroscopic system of noninteracting particles in a box form a continuum. That is schematically indicated by the hatching in the energy spectrum to the right in figure 6.1. The spacing between energy levels is however very many orders of magnitude tighter than the hatching can indicate.

	helium atom	electron	photon
$E^{\rm p}_{111}$, eV:	$1.5\ 10^{-18}$	$1.1\ 10^{-14}$	$1.1\ 10^{-4}$
$T_{\rm equiv}$, K:	$1.2\ 10^{-14}$	$8.7\ 10^{-11}$	0.83

Table 6.1: Energy of the lowest single-particle state in a cube with 1 cm sides.

It can also normally be assumed that the lowest energy is zero for noninteracting particles in a box. While the lowest single particle energy is strictly speaking somewhat greater than zero, it is extremely small. That is numerically illustrated by the values for a 1 cm^3 cubic box in table 6.1. The table gives the lowest energy as computed using the formulae given in the previous section. The lowest energy occurs for the state $\psi^{\rm p}_{111}$ with $n_x = n_y = n_z = 1$. As is common

for single-particle energies, the energy has been expressed in terms of electron volts, one eV being about 1.6 10^{-19} J. The table also shows the same energy in terms of an equivalent temperature, found by dividing it by 1.5 times the Boltzmann constant. These temperatures show that at room temperature, for all practical purposes the lowest energy is zero. However, at very low cryogenic temperatures, photons in the lowest energy state, or "ground state," may have a relatively more significant energy.

The spacing between the lowest and second lowest energy is comparable to the lowest energy, and similarly negligible. It should be noted, however, that in Bose-Einstein condensation, which is discussed later, there is a macroscopic effect of the finite spacing between the lowest and second-lowest energy states, miniscule as it might be.

The next question is why quantum mechanics is needed here at all. Classical nonquantum physics too would predict a continuum of energies for the particles. And it too would predict the energy to start from zero. The energy of a noninteracting particle is all kinetic energy; classical physics has that zero if the particle is at rest and positive otherwise.

Still, the (anti) symmetrization requirements cannot be accommodated using classical physics. And there is at least one other important quantum effect. Quantum mechanics predicts that there are more single-particle states in a given energy range at high energy than at low energy.

To express that more precisely, physicists define the "density of states" as the number of single-particle states per unit energy range. For particles in a box, the density of states is not that hard to find. First, the number $\mathrm{d}N$ of single-particle states in a small wave number range from k to $k + \mathrm{d}k$ is given by, {D.26},

$$\mathrm{d}N = \mathcal{V}\mathcal{D}_k \,\mathrm{d}k \qquad \mathcal{D}_k = \frac{2s+1}{2\pi^2}k^2 \tag{6.5}$$

Here $\mathcal{V}$ is the volume of the box that holds the particles. As you would expect, the bigger the box, the more particles it can hold, all else being the same. Similarly, the larger the wave number range $\mathrm{d}k$, the larger the number of states in it. The factor $\mathcal{D}_k$ is the density of states on a wave number basis. It depends on the spin s of the particles; that reflects that there are $2s + 1$ possible values of the spin for every given spatial state.

(It should be noted that for the above expression for $\mathcal{D}_k$ to be valid, the wave number range $\mathrm{d}k$ should be small. However, $\mathrm{d}k$ should still be large enough that there are a lot of states in the range $\mathrm{d}k$; otherwise $\mathcal{D}_k$ cannot be approximated by a simple continuous function. If the spacing $\mathrm{d}k$ truly becomes zero, $\mathcal{D}_k$ turns into a distribution of infinite spikes.)

To get the density of states on an energy basis, eliminate k in favor of the single-particle energy E^{p} using $E^{\mathrm{p}} = \hbar^2k^2/2m$, where m is the particle mass.

That gives:

$$\boxed{\mathrm{d}N = \mathcal{V}\mathcal{D}\,\mathrm{d}E^{\mathrm{p}} \qquad \mathcal{D} = \frac{2s+1}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E^{\mathrm{p}}}} \tag{6.6}$$

The requirements on the energy range $\mathrm{d}E^{\mathrm{p}}$ are like those on $\mathrm{d}k$.

The factor $\mathcal{D}$ is what is conventionally defined as the density of states; it is on a unit energy range and unit volume basis. In the spectrum to the right in figure 6.1, the density of states is indicated by means of the width of the spectrum.

Note that the density of states grows like $\sqrt{E^{\mathrm{p}}}$: quickly at first, more slowly later, but it continues to grow. There are more states per unit energy range at higher energy than at lower energy. And that means that at nonzero energies, the energy states are spaced many times tighter together still than the ground state spacing of table 6.1 indicates. Assuming that the energies form a continuum is an extremely accurate approximation in most cases.

The given expression for the density of states is not valid if the particle speed becomes comparable to the speed of light. In particular for photons the Planck-Einstein expression for the energy must be used, $E^{\mathrm{p}} = \hbar\omega$, where the electromagnetic frequency is $\omega = ck$ with c the speed of light. In addition, as mentioned in section 6.2, photons have only two independent spin states, even though their spin is 1.

It is conventional to express the density of states for photons on a frequency basis instead of an energy basis. Replacing k with ω/c in (6.5) and $2s+1$ by 2 gives

$$\boxed{\mathrm{d}N = \mathcal{V}\mathcal{D}_\omega\,\mathrm{d}\omega \qquad \mathcal{D}_\omega = \frac{1}{\pi^2 c^3}\omega^2} \tag{6.7}$$

The factor $\mathcal{D}_\omega$ is commonly called the "density of modes" instead of density of states on a frequency basis.

Key Points

- The spectrum of a macroscopic number of noninteracting particles in a box is practically speaking continuous.
- The lowest single-particle energy can almost always be taken to be zero.
- The density of states $\mathcal{D}$ is the number of single-particle states per unit energy range and unit volume.
- More precisely, the number of states in an energy range $\mathrm{d}E^{\mathrm{p}}$ is $\mathcal{V}\mathcal{D}\,\mathrm{d}E^{\mathrm{p}}$.
- To use this expression, the energy range $\mathrm{d}E^{\mathrm{p}}$ should be small. However, $\mathrm{d}E^{\mathrm{p}}$ should still be large enough that there are a lot of states in the range.

For photons, use the density of modes.

6.4 Ground State of a System of Bosons

The ground state for a system of noninteracting spinless bosons is simple. The ground state is defined as the state of lowest energy, so every boson has to be in the single-particle state $\psi^{\text{p}}_{111}(\vec{r})$ of lowest energy. That makes the system energy eigenfunction for spinless bosons equal to:

$$\psi_{\text{gs, bosons}} = \psi^{\text{p}}_{111}(\vec{r}_1) \times \psi^{\text{p}}_{111}(\vec{r}_2) \times \ldots \times \psi^{\text{p}}_{111}(\vec{r}_I) \tag{6.8}$$

If the bosons have spin, this is additionally multiplied by an arbitrary combination of spin states. That does not change the system energy. The system energy either way is IE^{p}_{111}, the number of bosons times the single-particle ground state energy.

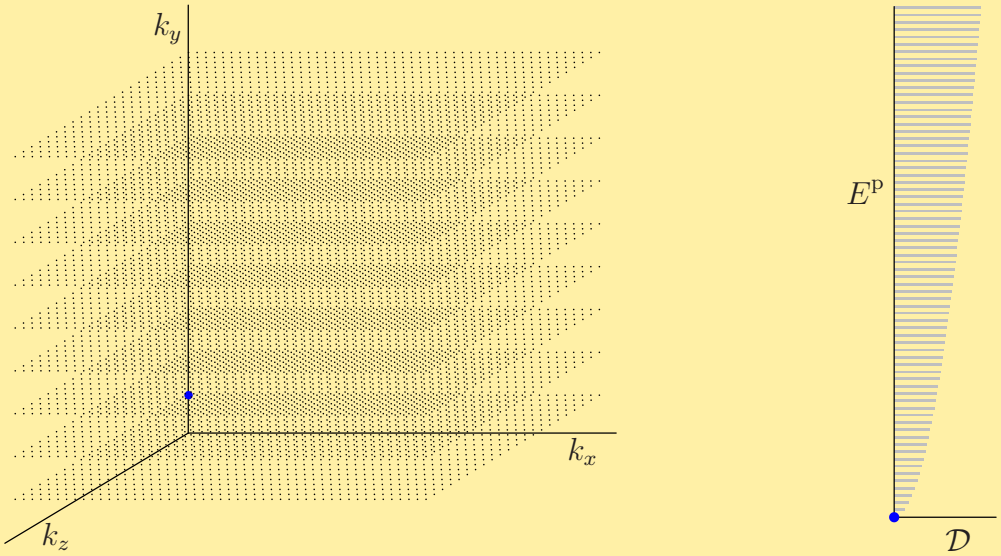


Figure 6.2: Ground state of a system of noninteracting bosons in a box.

Graphically, the single-particle ground state ψ^{p}_{111} is the point closest to the origin in wave number space. It is shown as a fat blue dot in figure 6.2 to indicate that all I bosons are bunched together in that state.

Physicists like to talk about "occupation numbers." The occupation number of a single-particle state is simply the number of particles in that state. In particular, for the ground state of the system of noninteracting spinless bosons above, the single-particle state ψ^{p}_{111} has occupation number I, while all other single-particle states have zero.

Note that for a macroscopic system, I will be a humongous number. Even a millimol of particles means well over 10^{20} particles. Bosons in their ground state are very unfair to the single-particle states: $\psi^{\rm p}_{111}$ gets all of them, the rest gets nothing.

Key Points

- For a system of bosons in the ground state, every boson is in the single particle state of lowest energy.

6.5 About Temperature

The previous section discussed the wave function for a macroscopic system of bosons in its ground state. However, that is really a very theoretical exercise.

A macroscopic system of particles is only in its ground state at what is called absolute zero temperature. Absolute zero temperature is −273.15 °C in degrees Celsius (Centigrade) or −459.67 °F in degrees Fahrenheit. It is the coldest that a stable system could ever be.

Of course, you would hardly think something special was going on from the fact that it is −273.15 °C or −459.67 °F. That is why physicists have defined a more meaningful temperature scale than Centigrade or Fahrenheit; the Kelvin scale. The Kelvin scale takes absolute zero temperature to be 0 K, zero degrees Kelvin. A one degree temperature *difference* in Kelvin is still the same as in Centigrade. So 1 K is the same as −272.15 °C; both are one degree above absolute zero. Normal ambient temperatures are near 300 K. More precisely, 300 K is equal to 27.15 °C or 80.6 °F.

A temperature measured from absolute zero, like a temperature expressed in Kelvin, is called an "absolute temperature." Any theoretical computation that you do requires the use of absolute temperatures. (However, there are some empirical relations and tables that are mistakenly phrased in terms of Celsius or Fahrenheit instead of in Kelvin.)

Absolute zero temperature is impossible to achieve experimentally. Even getting close to it is very difficult. Therefore, real macroscopic systems, even very cold ones, have an energy noticeably higher than their ground state. So they have a temperature above absolute zero.

But what exactly is that temperature? Consider the classical picture of a substance, in which the molecules that it consists of are in constant chaotic thermal motion. Temperature is often described as a measure of the translational kinetic energy of this chaotic motion. The higher the temperature, the larger the thermal motion. In particular, classical statistical physics would say that the average thermal kinetic energy per particle is equal to $\frac{3}{2}k_{\rm B}T$, with $k_{\rm B}$

$= 1.38\ 10^{-23}$ J/K the Boltzmann constant and T the absolute temperature in degrees Kelvin.

Unfortunately, this story is only true for the translational kinetic energy of the molecules in an ideal gas. For any other kind of substance, or any other kind of kinetic energy, the quantum effects are much too large to be ignored. Consider, for example, that the electron in a hydrogen atom has 13.6 eV worth of kinetic energy even at absolute zero temperature. (The binding energy also happens to be 13.6 eV, {A.17}, even though physically it is not the same thing.) Classically that kinetic energy would correspond to a gigantic temperature of about 100 000 K. Not to 0 K. More generally, the Heisenberg uncertainty principle says that particles that are in any way confined must have kinetic energy even in the ground state. Only for an ideal gas is the containing box big enough that it does not make a difference. Even then that is only true for the translational degrees of freedom of the ideal gas molecules. Don't look at their electrons or rotational or vibrational motion.

The truth is that temperature is not a measure of kinetic energy. Instead the temperature of a system is a measure of its capability to transfer thermal energy to other systems. By definition, if two systems have the same temperature, neither is able to transfer net thermal energy to the other. It is said that the two systems are in thermal equilibrium with each other. If however one system is hotter than the other, then if they are put in thermal contact, energy will flow from the hotter system to the colder one. That will continue until the temperatures become equal. Transferred thermal energy is referred to as "heat," so it is said that heat flows from the hotter system to the colder.

The simplest example is for systems in their ground state. If two systems in their ground state are brought together, no heat will transfer between them. By definition the ground state is the state of lowest possible energy. Therefore neither system has any spare energy available to transfer to the other system. It follows that all systems in their ground state have the same temperature. This temperature is simply defined to be absolute zero temperature, 0 K. Systems at absolute zero have zero capability of transferring heat to other systems.

Systems not in their ground state are not at zero temperature. Besides that, basically all that can be said is that they still have the same temperature as any other system that they are in thermal equilibrium with. But of course, this only defines *equality* of temperatures. It does not say what the *value* of that temperature is.

For identification and computational purposes, you would like to have a specific numerical value for the temperature of a given system. To get it, look at an ideal gas that the system is in thermal equilibrium with. A numerical value of the temperature can simply be *defined* by demanding that the average translational kinetic energy of the ideal gas molecules is equal to $\frac{3}{2}k_{\rm B}T$, where $k_{\rm B}$ is the Boltzmann constant, $1.380\,65\ 10^{-23}$ J/K. That kinetic energy can be deduced from such easily measurable quantities as the pressure, volume, and

mass of the ideal gas.

Key Points

- A macroscopic system is in its ground state if the absolute temperature is zero.
- Absolute zero temperature means 0 K (Kelvin), which is equal to -273.15 °C (Centigrade) or -459.67 °F (Fahrenheit).
- Absolute zero temperature can never be fully achieved.
- If the temperature is greater than absolute zero, the system will have an energy greater than that of the ground state.
- Temperature is not a measure of the thermal kinetic energy of a system, except under very limited conditions in which there are no quantum effects.
- Instead the defining property of temperature is that it is the same for systems that are in thermal equilibrium with each other.
- For systems that are not in their ground state, a numerical value for their temperature can be defined using an ideal gas at the same temperature.

6.6 Bose-Einstein Condensation

This section examines what happens to a system of noninteracting bosons in a box if the temperature is somewhat greater than absolute zero.

As noted in the second last section, in the ground state all bosons are in the single-particle state of lowest energy. This was indicated by the fat blue dot next to the origin in the wave number space figure 6.2. Nonzero temperature implies that the bosons obtain an additional amount of energy above the ground state. Therefore they will spread out a bit towards states of higher energy. The single fat blue point will become a colored cloud as shown in figures 6.3 and 6.4. So far, that all seems plausible enough.

But something weird occurs for identical bosons:

> *Below a certain critical temperature a finite fraction of the bosons remains bunched together in the single-particle state of lowest energy.*

That is indicated by the fat blue dot in figure 6.3. The lowest energy state, the one closest to the origin, holds less bosons than at absolute zero, but below a certain critical temperature, it remains a finite fraction of the total.

That is weird because the average thermal energy available to each boson dwarfs the difference in energy between the lowest energy single-particle state

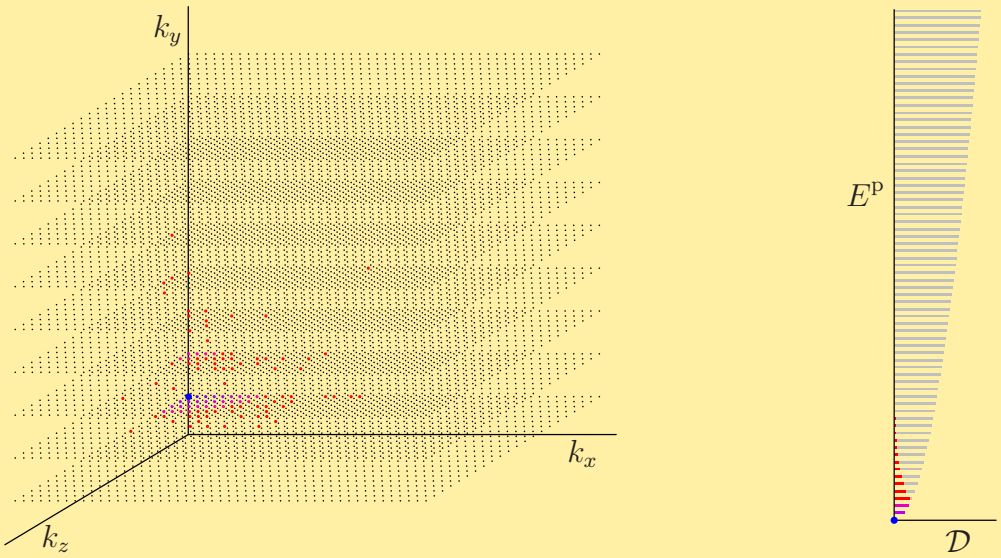


Figure 6.3: The system of bosons at a very low temperature.

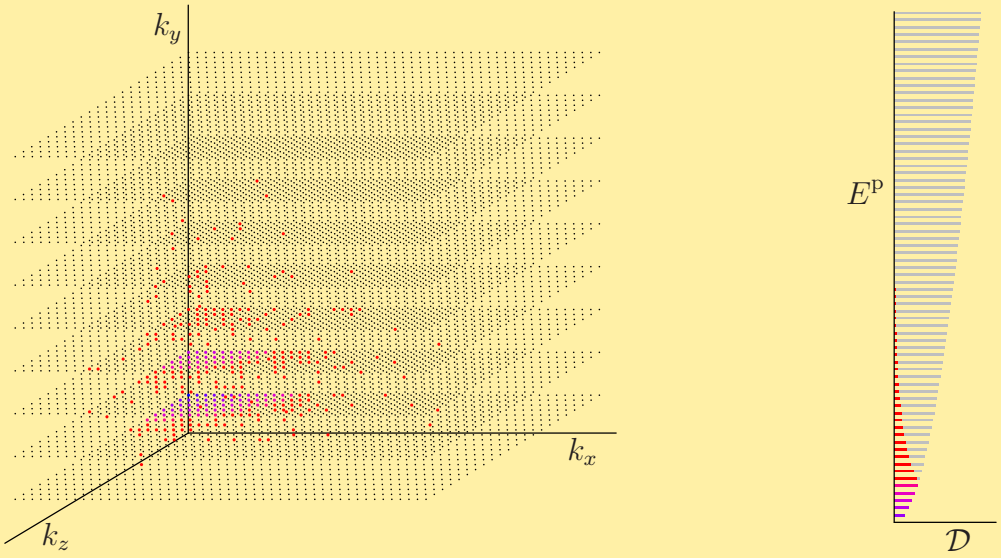


Figure 6.4: The system of bosons at a relatively low temperature.

and its immediate neighbors. If the energy difference between the lowest energy state and its neighbors is negligibly small, you would reasonably expect that they will hold similar numbers of bosons. And if a lot of states near the origin each hold about the same number of bosons, then that number must be a small fraction of the total, not a finite one. But reasonable or not, it is untrue. For a system of noninteracting bosons below the critical temperature, the lowest energy state holds a finite fraction of the bosons, much more than its immediate neighbors.

If you raise the temperature of the system, you "boil" away the bosons in the lowest energy state into the surrounding cloud. Above the critical temperature, the excess bosons are gone and the lowest energy state now only holds a similar number of bosons as its immediate neighbors. That is illustrated in figure 6.4. Conversely, if you lower the temperature of the system from above to below the critical temperature, the bosons start "condensing" into the lowest energy state. This process is called "Bose-Einstein condensation" after Bose and Einstein who first predicted it.

Bose-Einstein condensation is a pure quantum effect; it is due to the symmetrization requirement for the wave function. It does not occur for fermions, or if each particle in the box is distinguishable from every other particle. "Distinguishable" should here be taken to mean that there are no antisymmetrization requirements, as there are not if each particle in the system is a different type of particle from every other particle.

It should be noted that the given description is simplistic. In particular, it is certainly possible to cool a *microscopic* system of distinguishable particles down until say about half the particles are in the single-particle state of lowest energy. Based on the above discussion, you would then conclude that Bose-Einstein condensation has occurred. That is not true. The problem is that this supposed "condensation" disappears when you scale up the system to macroscopic dimensions and a corresponding macroscopic number of particles.

Given a microscopic system of distinguishable particles with half in the single-particle ground state, if you hold the temperature constant while increasing the system size, the size of the cloud of occupied states in wave number space remains about the same. However, the bigger macroscopic system has much more energy states, spaced much closer together in wave number space. Distinguishable particles spread out over these additional states, leaving only a vanishingly small fraction in the lowest energy state. This does not happen if you scale up a Bose-Einstein condensate; here the fraction of bosons in the lowest energy state stays finite regardless of system size.

Bose-Einstein condensation was achieved in 1995 by Cornell, Wieman, *et al* by cooling a dilute gas of rubidium atoms to below about 170 nK (nano Kelvin). Based on the extremely low temperature and fragility of the condensate, practical applications are very likely to be well into the future, and even determination of the condensate's basic properties will be difficult.

A process similar to Bose-Einstein condensation is also believed to occur in liquid helium when it turns into a superfluid below 2.17 K. However, this case is more tricky, {N.21}. For one, the atoms in liquid helium can hardly be considered to be noninteracting. That makes the entire concept of "single-particle states" poorly defined. Still, it is quite widely believed that for helium below 2.17 K, a finite fraction of the atoms starts accumulating in what is taken to be a single-particle state of zero wave number. Unlike for normal Bose-Einstein condensation, for helium it is believed that the number of atoms in this state remains limited. At absolute zero only about 9% of the atoms end up in the state.

Currently there is a lot of interest in other systems of particles undergoing Bose-Einstein condensation. One example is liquid helium-3. Compared to normal helium, helium-3 misses a neutron in its nucleus. That makes its spin half-integer, so it is not a boson but a fermion. Therefore, it should not turn into a superfluid like normal liquid helium. And it does not. Helium 3 behaves in almost all aspects exactly the same as normal helium. It becomes liquid at a similar temperature, 3.2 K instead of 4.2 K. But it does not become a superfluid like normal helium at any temperature comparable to 2.17 K. That is very strong evidence that the superfluid behavior of normal helium is due to the fact that it is a boson.

Still it turns out that at temperatures three orders of magnitude smaller, helium-3 does turn into a superfluid. That is believed to be due to the fact that the atoms pair up. A composite of two fermions has integer spin, so it is a boson. Similarly, superconductivity of simple solids is due to the fact that the electrons pair up into "Cooper pairs." They get tied together due to their interaction with the surrounding atoms.

A variety of other particles can pair up to. At the time of writing, there is interest in polariton condensates. A polariton is a quantum mechanical superposition of a photon and an electronic excitation in a solid. It is hoped that these will allow Bose-Einstein condensation to be studied at room temperature. There is still much to be learned about it. For example, while the relationship between superfluidity and Bose-Einstein condensation is quite generally accepted, there are some issues. Snoke & Baym point out, (in the introduction to *Bose-Einstein Condensation*, Griffin, A., Snoke, D.W., & Stringari, S., Eds, 1995, Cambridge), that examples indicate that Bose-Einstein condensation is neither necessary nor sufficient for superfluidity. With only approximate theoretical models and approximate experimental data, it is often difficult to make solid specific statements.

Key Points

- In Bose-Einstein condensation, a finite fraction of the bosons is in the single-particle state of lowest energy.

- It happens when the temperature falls below a critical value.
- It applies to macroscopic systems.
- The effect is unique to bosons.

6.6.1 Rough explanation of the condensation

The reason why bosons show Bose-Einstein condensation while systems of distinguishable particles do not is complex. It is discussed in chapter 11. However, the idea can be explained qualitatively by examining a very simple system

$$\psi^{\rm S}_{\rm gs}(\vec{r}_1,\vec{r}_2,\vec{r}_3) = \psi^{\rm p}_1(\vec{r}_1)\psi^{\rm p}_1(\vec{r}_2)\psi^{\rm p}_1(\vec{r}_3)$$
$$E^{\rm S}_{\rm gs} = E^{\rm p}_1 + E^{\rm p}_1 + E^{\rm p}_1 = 3E^{\rm p}_1$$

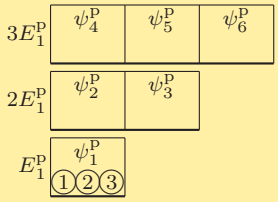


Figure 6.5: Ground state system energy eigenfunction for a simple model system. The system has only 6 single-particle states; each of these has one of 3 energy levels. In the specific case shown here, the system contains 3 distinguishable spinless particles. All three are in the single-particle ground state. Left: mathematical form. Right: graphical representation.

Assume that there are just three different single-particle energy levels, with values $E^{\rm p}_1$, $2E^{\rm p}_1$, and $3E^{\rm p}_1$. Also assume that there is just one single-particle state with energy $E^{\rm p}_1$, but two with energy $2E^{\rm p}_1$ and 3 with energy $3E^{\rm p}_1$. That makes a total of 6 single particle-states; they are shown as "boxes" that can hold particles at the right hand side of figure 6.5. Assume also that there are just three particles and for now take them to be distinguishable. Figure 6.5 then shows the system ground state in which every particle is in the single-particle ground state with energy $E^{\rm p}_1$. That makes the total system energy $3E^{\rm p}_1$.

$$\psi^{\rm S}_{151}(\vec{r}_1,\vec{r}_2,\vec{r}_3) = \psi^{\rm p}_1(\vec{r}_1)\psi^{\rm p}_5(\vec{r}_2)\psi^{\rm p}_1(\vec{r}_3)$$
$$E^{\rm S}_{151} = E^{\rm p}_1 + 3E^{\rm p}_1 + E^{\rm p}_1 = 5E^{\rm p}_1$$

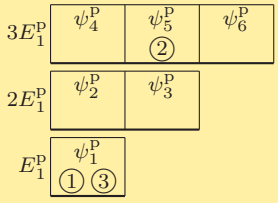


Figure 6.6: Example system energy eigenfunction with five times the single-particle ground state energy.

However, now assume that the system is at a nonzero temperature. In particular, assume that the total system energy is $5E^{\rm p}_1$. An example system energy eigenfunction with that energy is illustrated in figure 6.6.

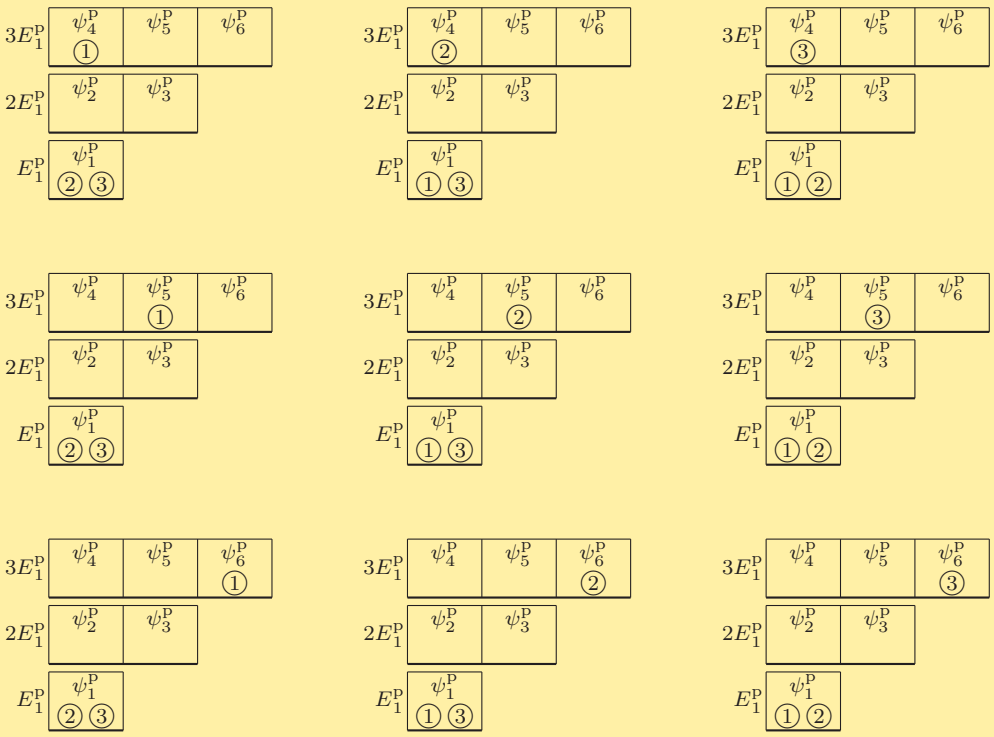


Figure 6.7: For distinguishable particles, there are 9 system energy eigenfunctions that have energy distribution A.

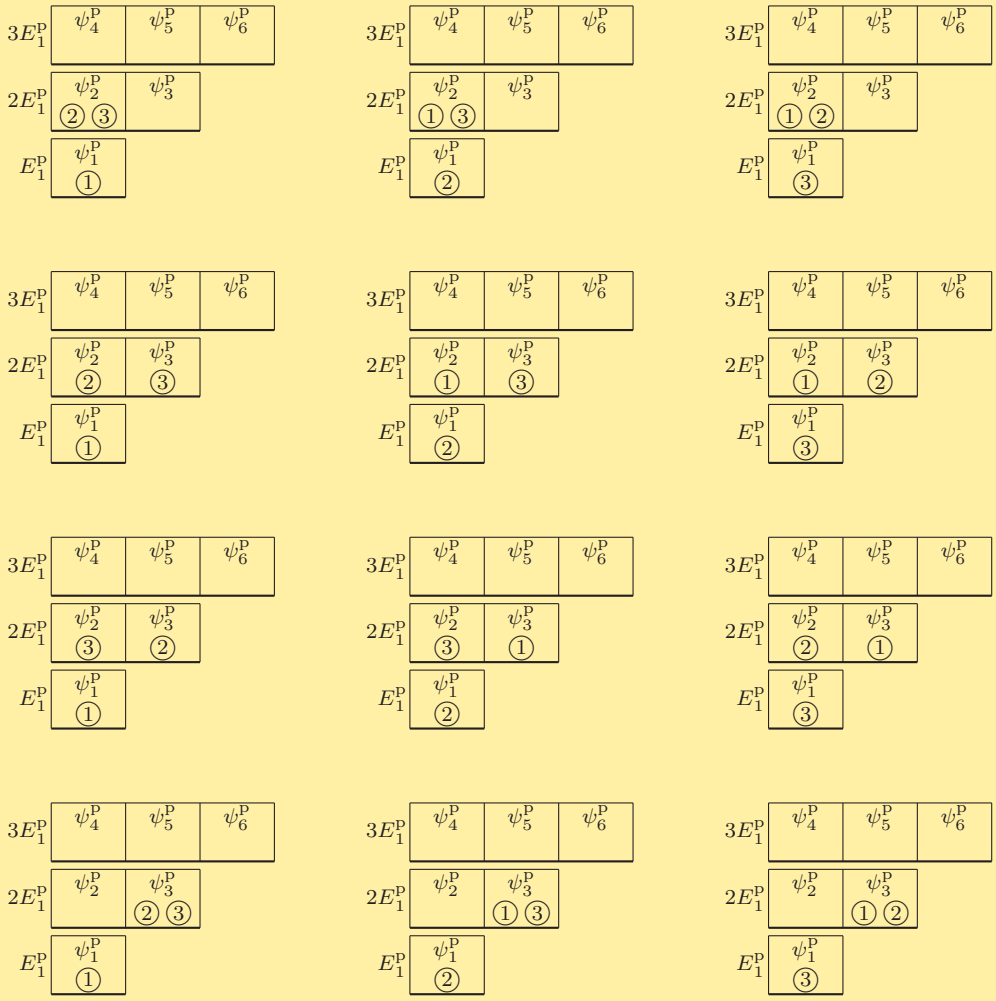


Figure 6.8: For distinguishable particles, there are 12 system energy eigenfunctions that have energy distribution B.

But there are a lot more system eigenfunctions with energy $5E_1^{\rm p}$. There are two general ways to achieve that energy:

Energy distribution A: Two particles in the ground state with energy $E_1^{\rm p}$ and one in a state with energy $3E_1^{\rm p}$.

Energy distribution B: One particle in the ground state with energy $E_1^{\rm p}$ and two in states with energy $2E_1^{\rm p}$.

As figures 6.7 and 6.8 show, there are 9 system energy eigenfunctions that have energy distribution A, but 12 that have energy distribution B.

Therefore, all else being the same, energy distribution B is more likely to be observed than A!

Of course, the difference between 9 system eigenfunctions and 12 is minor. Also, everything else is not the same; the eigenfunctions differ. But it turns out that if the system size is increased to macroscopic dimensions, the differences in numbers of energy eigenfunctions become gigantic. There will be one energy distribution for which there are *astronomically more* system eigenfunctions than for any other energy distribution. Common-sense statistics then says that this energy distribution is the only one that will ever be observed. If there are countless orders of magnitude more eigenfunctions for a distribution B than for a distribution A, what are the chances of A ever being found?

It is curious to think of it: only one energy distribution is observed for a given macroscopic system. And that is not because of any physics; other energy distributions are physically just as good. It is because of a mathematical count; there are just so many more energy eigenfunctions with that distribution.

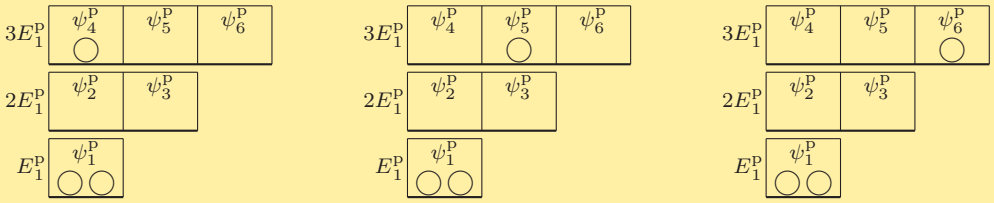


Figure 6.9: For identical bosons, there are only 3 system energy eigenfunctions that have energy distribution A.

Bose-Einstein condensation has to do with the fact that the count of eigenfunctions is different for identical bosons than for distinguishable particles. The details were worked out in chapter 5.7. The symmetrization requirement for bosons implies that system eigenfunctions that are the same except for exchanges of particles must be combined together into one. In particular for distribution A, in each of the rows of figure 6.7 the eigenfunctions are the same except for such exchanges. Simply put, they merely differ in what number is stamped on each particle. Therefore, for each row, the eigenfunctions must

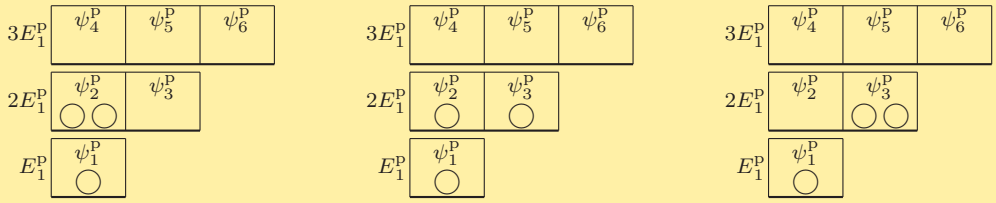


Figure 6.10: For identical bosons, there are also only 3 system energy eigenfunctions that have energy distribution B.

be combined together into a single eigenfunction. That leaves only the three system eigenfunctions shown in figure 6.9.

In the combination eigenfunction, every particle occupies every single-particle state involved equally. Therefore, numbers on the particles would not add any nontrivial information and may as well be left away. Sure, you could put all three numbers 1,2, and 3 in each of the particles in figure 6.9. But what good would that do?

Comparing figures 6.7 and 6.9, you can see why particles satisfying symmetrization requirements are commonly called "indistinguishable." Classical quantum mechanics may imagine to stamp numbers on the three identical bosons to keep them apart, but you sure do not see the difference between them in the system energy eigenfunctions.

For distribution B of figure 6.8, under the symmetrization requirement the three energy eigenfunctions in the first row must be combined into one, the six in the second and third rows must be combined into one, and the three in the fourth row must be combined into one. That gives a total of 3 system eigenfunctions for distribution B, as shown in figure 6.10.

It follows that the symmetrization requirement reduces the number of eigenfunctions for distribution A, with 2 particles in the ground state, from 9 to 3. However, it reduces the eigenfunctions for distribution B, with 1 particle in the ground state, from 12 to 3. Not only does the symmetrization requirement reduce the number of energy eigenfunctions, but it also tends to shift the balance towards eigenfunctions that have more particles in the ground state.

And so, if the system size is increased under conditions of Bose-Einstein condensation, it turns out that there are astronomically more system eigenfunctions for an energy distribution that keeps a finite number of bosons in the ground state than for anything else.

It may be noted from comparing figures 6.7 and 6.8 with 6.9 and 6.10 that any energy distribution that is physically possible for distinguishable particles is just as possible for identical bosons. Bose-Einstein condensation does not occur because the physics says it must, but because there are so gigantically more system eigenfunctions that have a finite fraction of bosons in the ground state than ones that do not.

It may also be noted that the reduction in the number of system energy eigenfunctions for bosons is believed to be an important factor in superfluidity. It eliminates low-energy eigenfunctions that cannot be interpreted as phonons, traveling particle wave solutions, [18, 39]. The lack of alternate eigenfunctions leaves no mechanism for the traveling particles to get scattered by small effects.

Key Points

- Energy distributions describe how many particles are found at each energy level.
- For macroscopic systems, one particular energy distribution has astronomically more energy eigenfunctions than any other one.
- That energy distribution is the only one that is ever observed.
- Under conditions of Bose-Einstein condensation, the observed distribution has a finite fraction of bosons in the ground state.
- This happens because the system eigenfunction count for bosons promotes it.

6.7 Bose-Einstein Distribution

As the previous section explained, the energy distribution of a macroscopic system of particles can be found by merely counting system energy eigenfunctions.

The details of doing so are messy but the results are simple. For a system of identical bosons, it gives the so-called:

$$\boxed{\text{Bose-Einstein distribution:} \quad \iota^{\mathrm{b}} = \frac{1}{e^{(E^{\mathrm{p}}-\mu)/k_{\mathrm{B}}T} - 1}} \tag{6.9}$$

Here ι^{b} is the average number of bosons in a single-particle state with single-particle energy E^{p}. Further T is the absolute temperature, and k_{B} is the Boltzmann constant, equal to $1.380\,65\ 10^{-23}$ J/K.

Finally, μ is known as the chemical potential and is a function of the temperature and particle density. The chemical potential is an important physical quantity, related to such diverse areas as particle diffusion, the work that a device can produce, and to chemical and phase equilibria. It equals the so-called Gibbs free energy on a molar basis. It is discussed in more detail in chapter 11.12.

The Bose-Einstein distribution is derived in chapter 11. In fact, for various reasons that chapter gives three different derivations of the distribution. Fortunately they all give the same answer. Keep in mind that whatever this book tells you thrice is absolutely true.

The Bose-Einstein distribution may be used to better understand Bose-Einstein condensation using a bit of simple algebra. First note that the chemical potential for bosons must always be less than the lowest single-particle energy $E^{\rm p}_{\rm gs}$. Just check it out using the formula above: if μ would be greater than $E^{\rm p}_{\rm gs}$, then the number of particles $\iota^{\rm b}$ in the lowest single-particle state would be negative. Negative numbers of particles do not exist. Similarly, if μ would equal $E^{\rm p}_{\rm gs}$ then the number of particles in the lowest single-particle state would be infinite.

The fact that μ must stay less than $E^{\rm p}_{\rm gs}$ means that the number of particles in anything but the lowest single-particle state has a limit. It cannot become greater than

$$\iota^{\rm b}_{\rm max} = \frac{1}{e^{(E^{\rm p}-E^{\rm p}_{\rm gs})/k_{\rm B}T}-1}$$

Now assume that you keep the box size and temperature both fixed and start putting more and more particles in the box. Then eventually, all the single-particle states except the ground state hit their limit. Any further particles have nowhere else to go than into the ground state. That is when Bose-Einstein condensation starts.

The above argument also illustrates that there are two main ways to produce Bose-Einstein condensation: you can keep the box and number of particles constant and lower the temperature, or you can keep the temperature and box constant and push more particles in the box. Or a suitable combination of these two, of course.

If you keep the box and number of particles constant and lower the temperature, the mathematics is more subtle. By itself, lowering the temperature lowers the number of particles $\iota^{\rm b}$ in all states. However, that would lower the total number of particles, which is kept constant. To compensate, μ inches closer to $E^{\rm p}_{\rm gs}$. This eventually causes all states except the ground state to hit their limit, and beyond that stage the left-over particles must then go into the ground state.

You may recall that Bose-Einstein condensation is only Bose-Einstein condensation if it does not disappear with increasing system size. That too can be verified from the Bose-Einstein distribution under fairly general conditions that include noninteracting particles in a box. However, the details are messy and will be left to chapter 11.14.1.

Key Points

- The Bose-Einstein distribution gives the number of bosons per single-particle state for a macroscopic system at a nonzero temperature.
- It also involves the Boltzmann constant and the chemical potential.
- It can be used to explain Bose-Einstein condensation.

6.8 Blackbody Radiation

The Bose-Einstein distribution of the previous section can also be used for understanding the emission of light and other electromagnetic radiation. If you turn on an electric stove, the stove plate heats up until it becomes red hot. The red glow that you see consists of photons with energies in the visible red range. When the stove plate was cold, it also emitted photons, but those were of too low energy to be seen by our unaided eyes.

The radiation system that is easiest to analyze is the inside of an empty box. Empty should here be read as devoid of matter. For if the temperature inside the box is above absolute zero, then the inside of the box will still be filled with the electromagnetic radiation that the atoms in the box surfaces emit. This radiation is representative of the radiation that truly black surfaces emit. Therefore, the radiation inside the box is called "blackbody radiation."

Before the advent of quantum mechanics, Rayleigh and Jeans had computed using classical physics that the energy of the radiation in the box would vary with electromagnetic frequency ω and temperature T as

$$\rho(\omega) = \frac{\omega^2}{\pi^2 c^3} k_{\rm B} T$$

where $k_{\rm B} = 1.38\ 10^{-23}$ J/K is the Boltzmann constant and c the speed of light. That was clearly all wrong except at low frequencies. For one thing, the radiation energy would become infinite at infinite frequencies!

It was this very problem that led to the beginning of quantum mechanics. To fix the problem, in 1900 Planck made the unprecedented assumption that energy would not come in arbitrary amounts, but only in discrete chunks of size $\hbar\omega$. The constant $\hbar$ was a completely new physical constant whose value could be found by fitting theoretical radiation spectra to experimental ones. Planck's assumption was however somewhat vague about exactly what these chunks of energy were physically. It was Einstein who proposed, in his 1905 explanation of the photoelectric effect, that $\hbar\omega$ gives the energy of photons, the particles of electromagnetic radiation.

Photons are bosons, relativistic ones, to be sure, but still bosons. Therefore the Bose-Einstein distribution should describe their statistics. More specifically, the average number of photons in each single-particle state should be

$$\iota_\gamma^{\rm b} = \frac{1}{e^{E^{\rm p}/k_{\rm B}T} - 1} \tag{6.10}$$

where γ is the standard symbol for a photon. Note the missing chemical potential. As discussed in chapter 11, the chemical potential is related to conservation of the number of particles. It does not apply to photons that are readily created out of nothing or absorbed by the atoms in the walls of the box. (One consequence is that Bose-Einstein condensation does not occur for photons, {N.21}.)

To get the energy of the photons in a small frequency range $\mathrm{d}\omega$, simply multiply the number of single particle states in that range, (6.7), by the number of photons per state $\iota_\gamma^{\mathrm{b}}$ above, and that by the single-photon energy $E^{\mathrm{p}} = \hbar\omega$.

That gives the radiation energy per unit volume of the box and per unit energy range as

$$\boxed{\rho(\omega) = \frac{\omega^2}{\pi^2 c^3} \frac{\hbar\omega}{e^{\hbar\omega/k_{\mathrm{B}}T} - 1}} \tag{6.11}$$

This expression is known as "Planck's blackbody spectrum."

For low frequencies, the final ratio is about $k_{\mathrm{B}}T$, giving the Rayleigh-Jeans result. That is readily verified from writing a Taylor series for the exponential in the denominator. For high frequencies the energy is much less because of the rapid growth of the exponential for large values of its argument. In particular, the energy no longer becomes infinite at high frequencies. It becomes zero instead.

To rewrite the blackbody spectrum in terms of the frequency $f = \omega/2\pi$ in cycles per second, make sure to convert the actual energy in a frequency range, $\mathrm{d}E = \rho(\omega)\,\mathrm{d}\omega$, to $\mathrm{d}E = \bar{\rho}(f)\,\mathrm{d}f$. Merely trying to convert ρ will get you into trouble. The same if you want to rewrite the blackbody spectrum in terms of the wave length $\lambda = c/f$.

For engineering purposes, what is often the most important is the amount of radiation emitted by a surface into its surroundings. Now it so happens that if you drill a little hole in the box, you get a perfect model for a truly black surface. An ideal black surface is *defined* as a surface that absorbs, rather than reflects, all radiation that hits it. If the hole in the box is small enough, any radiation that hits the hole enters the box and is never seen again. In that sense the hole is perfectly black.

And note that a black surface does not have to *look* black. If the black plate of your electric stove is hot enough, it will glow red. Similarly, if you would heat the inside of the box to the same temperature, the radiation inside the box would make the hole shine just as red. If you would heat the box to 6 000 K, about as hot as the surface of the sun, the hole would radiate sunlight.

The amount of radiation that is emitted by the hole can be found by simply multiplying Planck's spectrum by one quarter of the speed of light c, {D.27}. That gives for the radiation energy emitted per unit area, per unit frequency range, and per unit time:

$$\boxed{\mathcal{I}(\omega) = \frac{\omega^2}{4\pi^2 c^2} \frac{\hbar\omega}{e^{\hbar\omega/k_{\mathrm{B}}T} - 1}} \tag{6.12}$$

A perfectly black surface area would radiate the same amount as the hole.

If you see the hole under an angle, it will look just as bright per unit area as when you see it straight on, but it will seem smaller. So your eyes will receive less radiation. More generally, if A_e is a small black surface at temperature

T that emits radiation, then the amount of that radiation received by a small surface A_r is given by

$$\mathrm{d}E = \frac{\omega^2}{4\pi^3 c^2} \frac{A_e \cos\theta_e A_r \cos\theta_r}{r^2} \frac{\hbar\omega}{e^{\hbar\omega/k_{\mathrm{B}}T} - 1} \,\mathrm{d}\omega \,\mathrm{d}t \tag{6.13}$$

Here r is the distance between the small surfaces, while θ_e and θ_r are the angles that the connecting line between the surfaces makes with the normals to the emitting and receiving surfaces respectively.

Often the total amount of energy radiated away by a black surface is of interest. To get it, simply integrate the emitted radiation (6.12) over all values of the frequency. You will want to make a change of integration variable to $\hbar\omega/k_{\mathrm{B}}T$ while doing this and then use a table book like [41, 18.80, p. 132]. The result is called the "Stefan-Boltzmann law:

$$\boxed{\mathrm{d}E_{\text{total emitted}} = A\sigma_B T^4 \,\mathrm{d}t \qquad \sigma_B = \frac{\pi^2 k_{\mathrm{B}}^4}{60\hbar^3 c^2} \approx 5.67\ 10^{-8}\ \mathrm{W/m^2\ K^4}} \tag{6.14}$$

Since this is proportional to T^4, at 6 000 K 160 000 times as much radiation will be emitted as at room temperature. In addition, a much larger part of that radiation will be in the visible range. That is the reason you will see light coming from a hole in a box if it is at 6 000 K, but not when it is at room temperature.

A surface that is not perfectly black will absorb only a fraction of the radiation that hits it. The fraction is called the "absorptivity" a. Such a surface will also radiate less energy than a perfectly black one by a factor called the "emissivity" e. This assumes that the surface is in stable thermal equilibrium. More simply put, it assumes that no external source of energy is directed at the surface.

Helmholtz discovered that the absorptivity and emissivity of a surface are equal in thermal equilibrium, {D.28}. So poor absorbers are also poor emitters of radiation. That is why lightweight emergency blankets typically have reflective metallic coatings. You would think that they would want to absorb, rather than reflect, the heat of incoming radiation. But if they did, then according to Helmholtz they would also radiate precious body heat away to the surroundings.

Since a surface cannot absorb more radiation than hits it, the absorptivity cannot be greater than one, It follows that the emissivity cannot be greater than one either. No surface can absorb better or emit better than a perfectly black one. At least not when in thermodynamic equilibrium.

Note that absorptivity and emissivity typically depend on electromagnetic frequency. Substances that seem black to the eye may not be at invisible electromagnetic frequencies and vice-verse. It remains true for any given electromagnetic frequency that the absorptivity and emissivity at that frequency are equal. To soak up the heat of the sun in a solar energy application, you want

your material to be black in the visible frequency range emitted by the 6 000 K surface of the sun. However, you want it to be "white" in the infrared range emitted at the operating temperature of the material, in order that it does not radiate the heat away again.

Absorptivity and emissivity may also depend on the direction of the radiation, polarization, temperature, pressure, etcetera. In thermodynamic equilibrium, absorptivity and emissivity must still be equal, but only at the same frequency and same directions of radiation and polarization.

For surfaces that are not black, formula (6.13) will need to be modified for the relevant emissivity. A simplifying "grey body" assumption is often made that the absorptivity, and so the emissivity, is constant. Absorptivity and emissivity are usually defined as material properties, cited for infinitely thick samples. For objects, the terms absorptance and emittance are used.

Fluorescence/phosphorescence and stimulated emission (lasers) are important examples of radiative processes that are not in thermal equilibrium. The above discussion simply does not apply to them.

Key Points

- Blackbody radiation is the radiation emitted by a black surface that is in thermal equilibrium.
- Planck's blackbody spectrum determines how much is radiated at each frequency.
- Surfaces that are not black emit radiation that is less by a factor called the emissivity.
- Emissivity equals absorptivity for the same frequency and direction of radiation.
- If the material is not in thermal equilibrium, like energized materials, it is a completely different ball game.

6.9 Ground State of a System of Electrons

So far, only the physics of bosons has been discussed. However, by far the most important particles in physics are electrons, and electrons are fermions. The electronic structure of matter determines almost all engineering physics: the strength of materials, all chemistry, electrical conduction and much of heat conduction, power systems, electronics, etcetera. It might seem that nuclear engineering is an exception because it primarily deals with nuclei. However, nuclei consist of protons and neutrons, and these are spin $\frac{1}{2}$ fermions just like electrons. The analysis below applies to them too.

Noninteracting electrons in a box form what is called a "free-electron gas." The valence electrons in a block of metal are often modeled as such a free-electron gas. These electrons can move relatively freely through the block. As long as they do not try to get off the block, that is. Sure, a valence electron experiences repulsions from the surrounding electrons, and attractions from the nuclei. However, in the interior of the block these forces come from all directions and so they tend to average away.

Of course, the electrons of a "free" electron gas are confined. Since the term "noninteracting-electron gas" would be correct and understandable, there were few possible names left. So "free-electron gas" it was.

At absolute zero temperature, a system of fermions will be in the ground state, just like a system of bosons. However, the ground state of a macroscopic system of electrons, or any other type of fermions, is dramatically different from that of a system of bosons. For a system of bosons, in the ground state all bosons crowd together in the single-particle state of lowest energy. That was illustrated in figure 6.2. Not so for electrons. The Pauli exclusion principle allows only two electrons to go into the lowest energy state; one with spin up and the other with spin down. A system of I electrons needs at least $I/2$ spatial states to occupy. Since for a macroscopic system I is a some gigantic number like 10^{20}, that means that a gigantic number of states needs to be occupied.

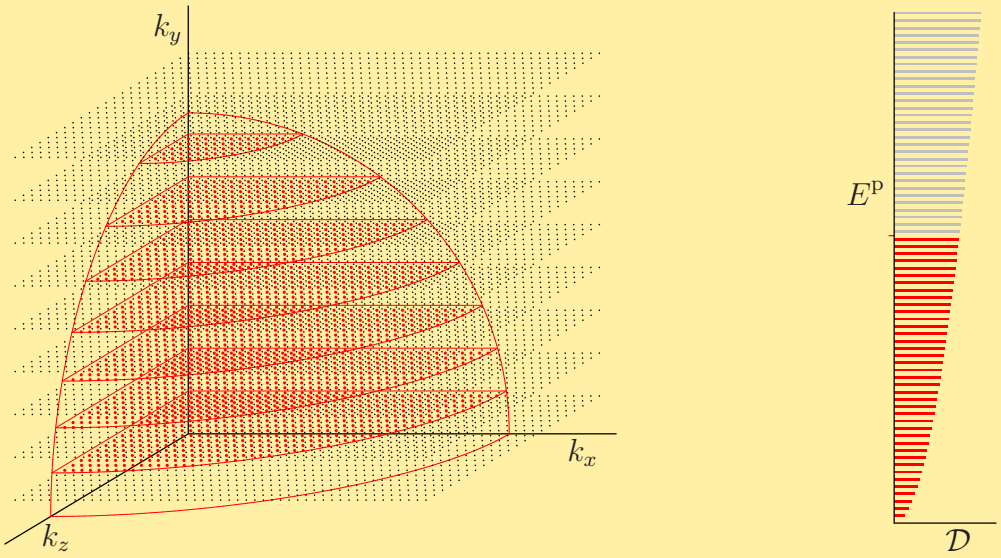


Figure 6.11: Ground state of a system of noninteracting electrons, or other fermions, in a box.

In the system ground state, the electrons crowd into the $I/2$ spatial states of lowest energy. Now the energy of the spatial states increases with the distance from the origin in wave number space. Therefore, the electrons occupy the $I/2$ states closest to the origin in this space. That is shown to the left in figure 6.11.

Every red spatial state is occupied by 2 electrons, while the black states are unoccupied. The occupied states form an octant of a sphere. Of course, in a real macroscopic system, there would be many more states than a figure could show.

The spectrum to the right in figure 6.11 shows the occupied energy levels in red. The width of the spectrum indicates the density of states, the number of single-particle states per unit energy range.

Key Points

- Noninteracting electrons in a box are called a free-electron gas.
- In the ground state, the $I/2$ spatial states of lowest energy are occupied by two electrons each. The remaining states are empty.
- The ground state applies at absolute zero temperature.

6.10 Fermi Energy of the Free-Electron Gas

As the previous section discussed, a system of noninteracting electrons, a free-electron gas, occupies a range of single-particle energies. Now the electrons with the highest single-particle energies are particularly important. The reason is that these electrons have empty single-particle states available at just very slightly higher energy. Therefore, these electrons are easily excited to do useful things, like conduct electricity for example. In contrast, electrons in energy states of lower energy do not have empty states within easy reach. Therefore lower energy electron are essentially stuck in their states; they do not usually contribute to nontrivial electronic effects.

Valence electrons in metals behave qualitatively much like a free-electron gas. For them too, the electrons in the highest energy single-particle states are the critical ones for the metallic properties. Therefore, the highest single-particle energy occupied by electrons in the system ground state has been given a special name; the "Fermi energy." In the energy spectrum of the free-electron gas to the right in figure 6.11, the Fermi energy is indicated by a red tick mark on the axis.

Also, the surface that the electrons of highest energy occupy in wave number space is called the "Fermi surface." For the free-electron gas the wave number space was illustrated to the left in figure 6.11. The Fermi surface is outlined in red in the figure; it is the spherical outside surface of the occupied region.

One issue that is important for understanding the properties of systems of electrons is the overall magnitude of the Fermi energy. Recall first that for a system of bosons, in the ground state all bosons are in the single-particle state of lowest energy. That state corresponds to the point closest to the origin in

wave number space. It has very little energy, even in terms of atomic units of electronic energy. That was illustrated numerically in table 6.1. The lowest single-particle energy is, assuming that the box is cubic

$$E^{\rm p}_{111} = 3\pi^2 \frac{\hbar^2}{2m_e} \frac{1}{\mathcal{V}^{2/3}} \tag{6.15}$$

where $m_{\rm e}$ is the electron mass and $\mathcal{V}$ the volume of the box.

Unlike for bosons, for electrons only two electrons can go into the lowest energy state. Or in any other spatial state for that matter. And since a macroscopic system has a gigantic number of electrons, it follows that a gigantic number of states must be occupied in wave number space. Therefore the states on the Fermi surface in figure 6.11 are many orders of magnitude further away from the origin than the state of lowest energy. And since the energy is proportional to the square distance from the origin, that means that the Fermi energy is many orders of magnitude larger than the lowest single-particle energy $E^{\rm p}_{111}$.

More precisely, the Fermi energy of a free-electron gas can be expressed in terms of the number of electrons per unit volume $I/\mathcal{V}$ as:

$$\boxed{E^{\rm p}_{\rm F} = \left(3\pi^2\right)^{2/3} \frac{\hbar^2}{2m_e} \left(\frac{I}{\mathcal{V}}\right)^{2/3}} \tag{6.16}$$

To check this relationship, integrate the density of states (6.6) given in section 6.3 from zero to the Fermi energy. That gives the total number of occupied states, which equals the number of electrons I. Inverting the expression to give the Fermi energy in terms of I produces the result above.

It follows that the Fermi energy is larger than the lowest single-particle energy by the gigantic factor

$$\frac{I^{2/3}}{\left(3\pi^2\right)^{1/3}}$$

It is instructive to put some ballpark number to the Fermi energy. In particular, take the valence electrons in a block of copper as a model. Assuming one valence electron per atom, the electron density $I/\mathcal{V}$ in the expression for the Fermi energy equals the atom density. That can be estimated to be 8.5 10^{28} atoms/m^3 by dividing the mass density, 9 000 kg/m^3, by the molar mass, 63.5 kg/kmol, and then multiplying that by Avogadro's number, 6.02 10^{26} particles/kmol. Plugging it in (6.16) then gives a Fermi energy of 7 eV (electron Volt). That is quite a lot of energy, about half the 13.6 eV ionization energy of hydrogen atoms.

The Fermi energy gives the maximum energy that an electron can have. The average energy that they have is comparable but somewhat smaller:

$$E^{\rm p}_{\rm average} = \tfrac{3}{5} E^{\rm p}_{\rm F} \tag{6.17}$$

To verify this expression, find the total energy $E = \int E^{\mathrm{p}} \mathcal{V}\mathcal{D}\,\mathrm{d}E^{\mathrm{p}}$ of the electrons using (6.6) and divide by the number of electrons $I = \int \mathcal{V}\mathcal{D}\,\mathrm{d}E^{\mathrm{p}}$. The integration is again over the occupied states, so from zero to the Fermi energy.

For copper, the ballpark average energy is 4.2 eV. To put that in context, consider the equivalent temperature at which classical particles would need to be to have the same average kinetic energy. Multiplying 4.2 eV by $e/\frac{3}{2}k_{\mathrm{B}}$ gives an equivalent temperature of 33 000 K. That is gigantic even compared to the melting point of copper, 1 356 K. It is all due to the exclusion principle that prevents the electrons from dropping down into the already filled states of lower energy.

Key Points

- The Fermi energy is the highest single-particle energy that a system of electrons at absolute zero temperature will occupy.
- It is normally a very high energy.
- The Fermi surface is the surface that the electrons with the Fermi energy occupy in wave number space.
- The average energy per electron for a free-electron gas is 60% of the Fermi energy.

6.11 Degeneracy Pressure

According to the previous sections, electrons, being fermions, behave in a way very differently from bosons. A system of bosons has very little energy in its ground state, as all bosons collect in the spatial state of lowest energy. Electrons cannot do so. At most two electrons can go into a single spatial state. A macroscopic system of electrons must occupy a gigantic number of states, ranging from the lowest energy state to states with many orders of magnitude more energy.

As a result, a "free-electron gas" of I noninteracting electrons ends up with an average energy per electron that is larger than of a corresponding system of bosons by a gigantic factor of order $I^{2/3}$. That is all kinetic energy; all forces on the electrons are ignored in the interior of a free-electron gas, so the potential energy can be taken to be zero.

Having so much kinetic energy, the electrons exert a tremendous pressure on the walls of the container that holds them. This pressure is called "degeneracy pressure." It explains qualitatively why the volume of a solid or liquid does not collapse under normally applied pressures.

Of course, degeneracy pressure is a poorly chosen name. It is really due to the fact that the energy distribution of electrons is *not* degenerate, unlike that

of bosons. Terms like "exclusion-principle pressure" or "Pauli pressure" would capture the essence of the idea. So they are not acceptable.

The magnitude of the degeneracy pressure for a free-electron gas is

$$P_{\mathrm{d}} = \tfrac{2}{5}\left(3\pi^2\right)^{2/3} \frac{\hbar^2}{2m_e} \left(\frac{I}{\mathcal{V}}\right)^{5/3} \tag{6.18}$$

This may be verified by equating the work $-P_{\mathrm{d}}\,\mathrm{d}\mathcal{V}$ done when compressing the volume a bit to the increase in the total kinetic energy E^{S} of the electrons:

$$-P_{\mathrm{d}}\,\mathrm{d}\mathcal{V} = \mathrm{d}E^{\mathrm{S}}$$

The energy E^{S} is I times the average energy per electron. According to section 6.10, that is $\frac{3}{5}I$ times the Fermi energy (6.16).

A ballpark number for the degeneracy pressure is very instructive. Consider once again the example of a block of copper, with its valence electrons modeled as a free-electron gas, Using the same numbers as in the previous section, the degeneracy pressure exerted by these valence electrons is found to be 40 10^9 Pa, or 40 GPa.

This tremendous outward pressure is balanced by the nuclei that pull on electrons that try to leave the block. The details are not that simple, but electrons that try to escape repel other, easily displaced, electrons that might aid in their escape, leaving the nuclei unopposed to pull them back. Obviously, electrons are not very smart.

It should be emphasized that it is *not* mutual repulsion of the electrons that causes the degeneracy pressure; all forces on the electrons are ignored in the interior of the block. It is the uncertainty relationship that requires spatially confined electrons to have momentum, and the exclusion principle that explodes the resulting amount of kinetic energy, creating fast electrons that are as hard to contain as students on the day before Thanksgiving.

Compared to a 10^{10} Pa degeneracy pressure, the normal atmospheric pressure of 10^5 Pa cannot add any noticeable further compression. Pauli's exclusion principle makes liquids and solids quite incompressible under normal pressures.

However, under extremely high pressures, the electron pressure can lose out. In particular, for neutron stars the spatial electron states collapse under the very weight of the massive star. This is related to the fact that the degeneracy pressure grows less quickly with compression when the velocity of the electrons becomes relativistic. (For very highly relativistic particles, the kinetic energy is not given in terms of the momentum p by the Newtonian value $E^{\mathrm{p}} = p^2/2m$, but by the Planck-Einstein relationship $E^{\mathrm{p}} = pc$ like for photons.) That makes a difference since gravity too increases with compression. If gravity increases more quickly, all is lost for the electrons. For neutron stars, the collapsed electrons combine with the protons in the star to form neutrons. It is the degeneracy pressure of the neutrons, also spin ½ fermions but 2 000 times heavier, that carries the weight of a neutron star.

Key Points

- Because typical confined electrons have so much kinetic energy, they exert a great degeneracy pressure on what is holding them.
- This pressure makes it very hard to compress liquids and solids significantly in volume.
- Differently put, liquids and solids are almost incompressible under typical conditions.

6.12 Confinement and the DOS

The motion of a single particle in a confining box was described in chapter 3.5.9. Nontrivial motion in a direction in which the box is sufficiently narrow can become impossible. This section looks at what happens to the density of states for such a box. The density of states gives the number of single-particle states per unit energy range. It is interesting for many reasons. For example, for systems of electrons the density of states at the Fermi energy determines how many electrons in the box pick up thermal energy if the temperature is raised above zero. It also determines how many electrons will be involved in electrical conduction if their energy is raised.

By definition, the density of states $\mathcal{D}$ gives the number of single-particle states $\mathrm{d}N$ in an energy range from E^{p} to $E^{\mathrm{p}} + \mathrm{d}E^{\mathrm{p}}$ as

$$\mathrm{d}N = \mathcal{V}\mathcal{D}\,\mathrm{d}E^{\mathrm{p}}$$

where $\mathcal{V}$ is the volume of the box containing the particles. To use this expression, the size of the energy range $\mathrm{d}E^{\mathrm{p}}$ should be small, but still big enough that the number of states $\mathrm{d}N$ in it remains large.

For a box that is not confining, the density of states is proportional to $\sqrt{E^{\mathrm{p}}}$. To understand why, consider first the total number of states N that have energy less than some given value E^{p}. For example, the wave number space to the left in figure 6.11 shows all states with energy less than the Fermi energy in red. Clearly, the number of such states is about proportional to the volume of the octant of the sphere that holds them. And that volume is in turn proportional to the cube of the sphere radius k, which is proportional to $\sqrt{E^{\mathrm{p}}}$, (6.4), so

$$N = (\text{some constant}) \; \left(E^{\mathrm{p}}\right)^{3/2}$$

This gives the number of states that have energies less than some value E^{p}. To get the number of states in an energy range from E^{p} to $E^{\mathrm{p}} + \mathrm{d}E^{\mathrm{p}}$, take a differential:

$$\mathrm{d}N = (\text{some other constant}) \; \sqrt{E^{\mathrm{p}}}\,\mathrm{d}E^{\mathrm{p}}$$

So the density of states is proportional to $\sqrt{E^{\rm p}}$. (The constant of proportionality is worked out in derivation {D.26}.) This density of states is shown as the width of the energy spectrum to the right in figure 6.11.

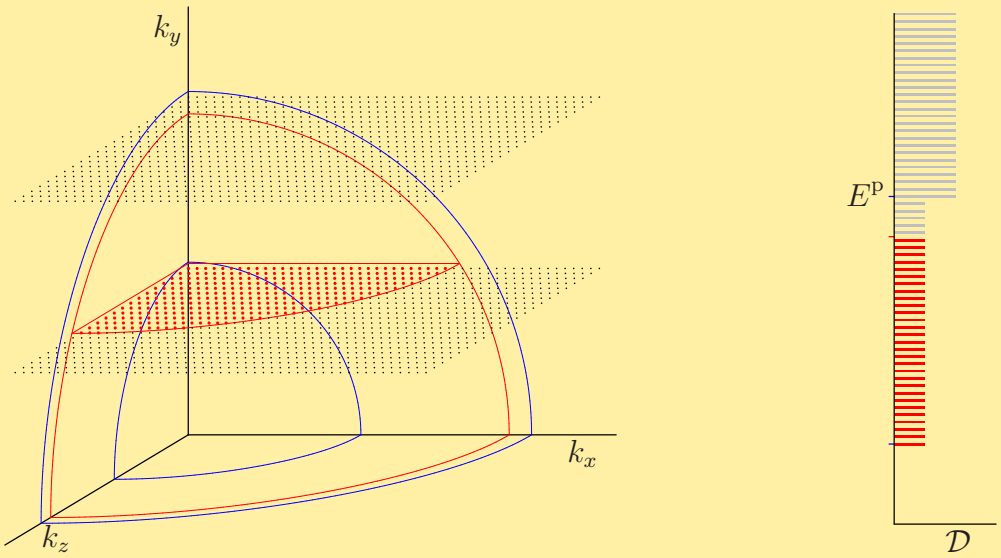


Figure 6.12: Severe confinement in the y-direction, as in a quantum well.

Confinement changes the spacing between the states. Consider first the case that the box containing the particles is very narrow in the y-direction only. That produces a quantum well, in which motion in the y-direction is inhibited. In wave number space the states become spaced very far apart in the k_y-direction. That is illustrated to the left in figure 6.12. The red states are again the ones with an energy below some given example value $E^{\rm p}$, say the Fermi energy. Clearly, now the number of states inside the red sphere is proportional not to its *volume*, but to the *area* of the quarter circle holding the red states. The density of states changes correspondingly, as shown to the right in figure 6.12.

Consider the variation in the density of states for energies starting from zero. As long as the energy is less than that of the smaller blue sphere in figure 6.12, there are no states at or below that energy, so there is no density of states either. However, when the energy becomes just a bit higher than that of the smaller blue sphere, the sphere gobbles up quite a lot of states compared to the small box volume. That causes the density of states to jump up. However, after that jump, the density of states does not continue grow like the unconfined case. The unconfined case keeps gobbling up more and more circles of states when the energy grows. The confined case remains limited to a single circle until the energy hits that of the larger blue sphere. At that point, the density of states jumps up again. Through jumps like that, the confined density of states eventually starts resembling the unconfined case when the energy levels get high enough.

As shown to the right in the figure, the density of states is piecewise constant for a quantum well. To understand why, note that the number of states on a circle is proportional to its square radius $k_x^2 + k_z^2$. That is the same as $k^2 - k_y^2$, and k^2 is directly proportional to the energy E^{p}. So the number of states varies linearly with energy, making its derivative, the density of states, constant. (The detailed mathematical expressions for the density of states for this case and the ones below can again be found in derivation {D.26}.)

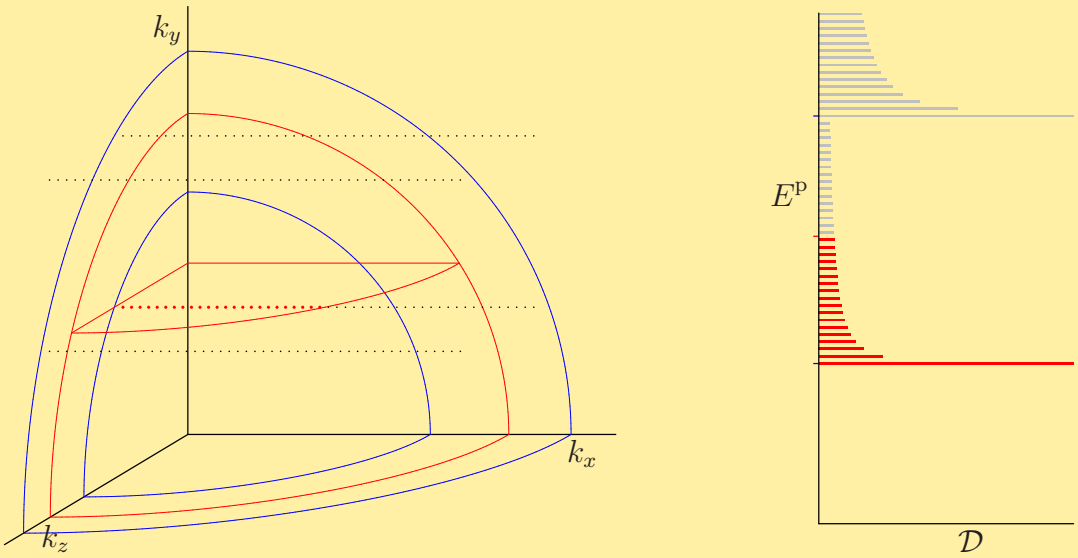


Figure 6.13: Severe confinement in both the y and z directions, as in a quantum wire.

The next case is that the box is very narrow in the z-direction as well as in the y-direction. This produces a quantum wire, where there is full freedom of motion only in the x-direction. This case is shown in figure 6.13. Now the states separate into individual lines of states. The smaller blue sphere just reaches the line of states closest to the origin. There are no energy states until the energy exceeds the level of this blue sphere. Just above that level, a lot of states are encountered relative to the very small box volume, and the density of states jumps way up. When the energy increases further, however, the density of states comes down again: compared to the less confined cases, no new lines of states are added until the energy hits the level of the larger blue sphere. When the latter happens, the density of states jumps way up once again. Mathematically, the density of states produced by each line is proportional to the reciprocal square root of the excess energy above the one needed to reach the line.

The final possibility is that the box holding the particles is very narrow in all three directions. This produces a quantum dot or artificial atom. Now each energy state is a separate point, figure 6.14. The density of states is now zero unless the energy sphere exactly hits one of the individual points, in which

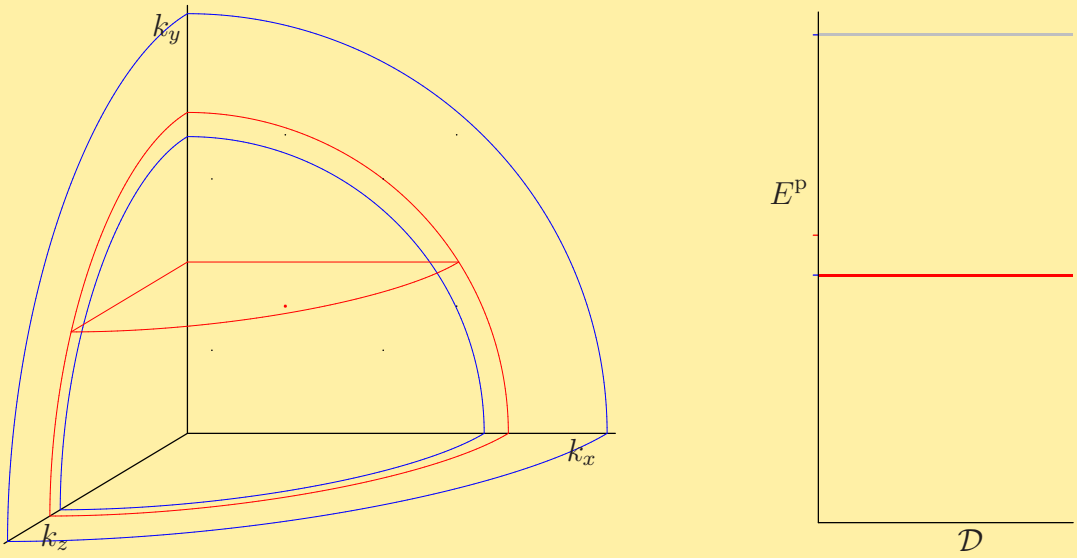


Figure 6.14: Severe confinement in all three directions, as in a quantum dot or artificial atom.

case the density of states is infinite. So, the density of states is a set of spikes. Mathematically, the contribution of each state to the density of states is a delta function located at that energy.

(It may be pointed out that very strictly speaking, every density of states is a set of delta functions. After all, the individual states always remain discrete points, however extremely densely spaced they might be. Only if you average the delta functions over a small energy range $\mathrm{d}E^{\mathrm{p}}$ do you get the smooth mathematical functions of the quantum wire, quantum well, and unconfined box. It is no big deal, as a perfect confining box does not exist anyway. In real life, energy spikes do broaden out bit; there is always some uncertainty in energy due to various effects.)

Key Points

- If one or more dimensions of a box holding a system of particles becomes very small, confinement effects show up.
- In particular, the density of states shows a staging behavior that is typical for each reduced dimensionality.

6.13 Fermi-Dirac Distribution

The previous sections discussed the ground state of a system of fermions like electrons. The ground state corresponds to absolute zero temperature. This

section has a look at what happens to the system when the temperature becomes greater than zero.

For nonzero temperature, the average number of fermions $\iota^{\rm f}$ per single-particle state can be found from the so-called

$$\boxed{\text{Fermi-Dirac distribution:} \quad \iota^{\rm f} = \frac{1}{e^{(E^{\rm p}-\mu)/k_{\rm B}T}+1}} \tag{6.19}$$

This distribution is derived in chapter 11. Like the Bose-Einstein distribution for bosons, it depends on the energy $E^{\rm p}$ of the single-particle state, the absolute temperature T, the Boltzmann constant $k_{\rm B} = 1.38\ 10^{-23}$ J/K, and a chemical potential μ. In fact, the mathematical difference between the two distributions is merely that the Fermi-Dirac distribution has a plus sign in the denominator where the Bose-Einstein one has a minus sign. Still, that small change makes for very different statistics.

The biggest difference is that $\iota^{\rm f}$ is always less than one: the Fermi-Dirac distribution can never have more than one fermion in a given single-particle state. That follows from the fact that the exponential in the denominator of the distribution is always greater than zero, making the denominator greater than one.

It reflects the exclusion principle: there cannot be more than one fermion in a given state, so the average per state cannot exceed one either. The Bose-Einstein distribution can have many bosons in a single state, especially in the presence of Bose-Einstein condensation.

Note incidentally that both the Fermi-Dirac and Bose-Einstein distributions count the different spin versions of a given spatial state as separate states. In particular for electrons, the spin-up and spin-down versions of a spatial state count as two separate states. Each can hold one electron.

Consider now the system ground state that is predicted by the Fermi-Dirac distribution. In the limit that the temperature becomes zero, single-particle states end up with either exactly one electron or exactly zero electrons. The states that end up with one electron are the ones with energies $E^{\rm p}$ below the chemical potential μ. Similarly the states that end up empty are the ones with $E^{\rm p}$ above μ.

To see why, note that for $E^{\rm p} - \mu < 0$, in the limit $T \to 0$ the argument of the exponential in the Fermi-Dirac distribution becomes minus infinity. That makes the exponential zero, and $\iota^{\rm f}$ is then equal to one. Conversely, for $E^{\rm p} - \mu > 0$, in the limit $T \to 0$ the argument of the exponential in the Fermi-Dirac distribution becomes positive infinity. That makes the exponential infinite, and $\iota^{\rm f}$ is then zero.

The correct ground state, as pictured earlier in figure 6.11, has one electron per state below the Fermi energy $E^{\rm p}_{\rm F}$ and zero electrons per state above the Fermi energy. The Fermi-Dirac ground state can only agree with this if the chemical

potential at absolute zero temperature is the same as the Fermi energy:

$$\boxed{\mu = E_{\mathrm{F}}^{\mathrm{p}} \quad \text{at} \quad T = 0} \tag{6.20}$$

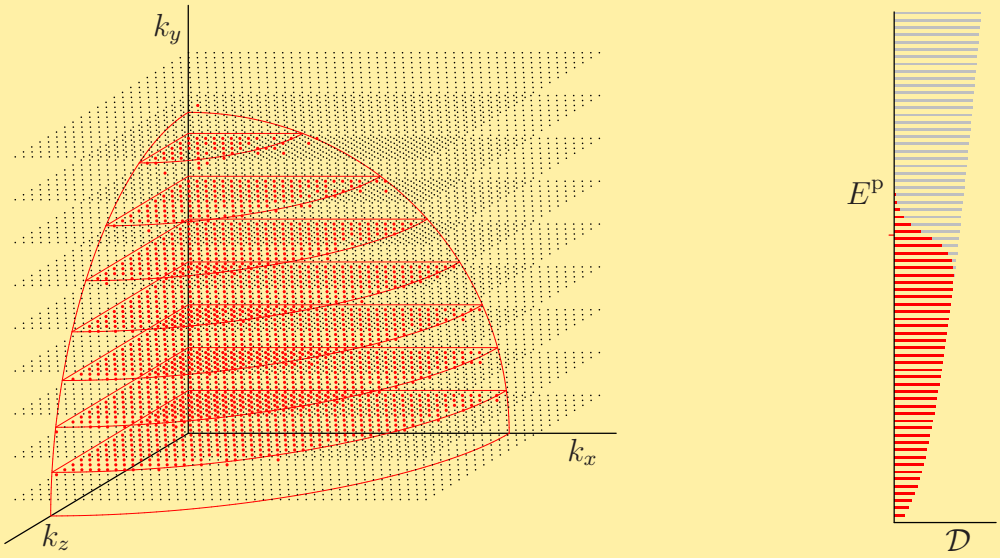


Figure 6.15: A system of fermions at a nonzero temperature.

Next consider what happens if the absolute temperature is not zero but a bit larger than that. The story given above for zero temperature does not change significantly unless the value of $E^{\mathrm{p}} - \mu$ is comparable to $k_{\mathrm{B}}T$. Only in a energy range of order $k_{\mathrm{B}}T$ around the Fermi energy does the average number of particles in a state change from its value at absolute zero temperature. Compare the spectrum at absolute zero temperature as sketched to the right in figure 6.11 to the one at a nonzero temperature shown in figure 6.15. The sharp transition from one particle per state, red, below the Fermi energy to zero particles per state, grey, above it smooths out a bit. As the wave number space to the left in figure 6.15 illustrates, at nonzero temperature a typical system energy eigenfunction has a few electrons slightly beyond the Fermi surface. Similarly it has a few "holes" (states that have lost their electron) immediately below the Fermi surface.

Put in physical terms, some electrons just below the Fermi energy pick up some thermal energy, which gives them an energy just above the Fermi energy. The affected energy range, and also the typical energy that the electrons in this range pick up, is comparable to $k_{\mathrm{B}}T$.

You may at first hardly notice the effect in the wave number space shown in figure 6.15. And that figure greatly exaggerates the effect to ensure that it is visible at all. Recall the ballpark Fermi energy given earlier for copper. It was equal to a $k_{\mathrm{B}}T$ value for an equivalent temperature of 33 000 K. Since the

melting point of copper is only 1 356 K, $k_B T$ is still negligibly small compared to the Fermi energy when copper melts. To good approximation, the electrons always remain like they were in their ground state at 0 K.

One of the mysteries of physics before quantum mechanics was why the valence electrons in metals do not contribute to the heat capacity. At room temperature, the atoms in typical metals were known to have picked up an amount of thermal energy comparable to $k_B T$ per atom. Classical physics predicted that the valence electrons, which could obviously move independently of the atoms, should pick up a similar amount of energy per electron. That should increase the heat capacity of metals. However, no such increase was observed.

The Fermi-Dirac distribution explains why: only the electrons within a distance comparable to $k_B T$ of the Fermi energy pick up the additional $k_B T$ of thermal energy. This is only a very small fraction of the total number of electrons, so the contribution to the heat capacity is usually negligible. While classically the electrons may seem to move freely, in quantum mechanics they are constrained by the exclusion principle. Electrons cannot move to higher energy states if there are already electrons in these states.

To discourage the absence of confusion, some or all of the following terms may or may not indicate the chemical potential μ, depending on the physicist: Fermi level, Fermi brim, Fermi energy, and electrochemical potential. It is more or less common to reserve "Fermi energy" to absolute zero temperature, but to not do the same for "Fermi level" or "Fermi brim." In any case, do not count on it. This book will occasionally use the term Fermi level for the chemical potential where it is common to do so. In particular, a Fermi-level electron has an energy equal to the chemical potential.

The term "electrochemical potential" needs some additional comment. The surfaces of solids are characterized by unavoidable layers of electric charge. These charge layers produce an electrostatic potential inside the solid that shifts all energy levels, including the chemical potential, by that amount. Since the charge layers vary, so does the electrostatic potential and with it the value of the chemical potential. It would therefore seem logical to define some "intrinsic" chemical potential, and add to it the electrostatic potential to get the total, or "electrochemical" potential.

For example, you might consider defining the "intrinsic" chemical potential μ_i of a solid as the value of the chemical potential μ when the solid is electrically neutral and isolated. Now, when you bring dissimilar solids at a given temperature into electrical contact, double layers of charge build up at the contact surfaces between them. These layers change the electrostatic potentials inside the solids and with it their total electrochemical potential μ.

In particular, the strengths of the double layers adjust so that in thermal equilibrium, the electrochemical potentials μ of all the solids (intrinsic plus additional electrostatic contribution due to the changed surface charge layers) are equal. They have to; solids in electrical contact become a single system of

electrons. A single system should have a single chemical potential.

Unfortunately, the assumed "intrinsic" chemical potential in the above description is a somewhat dubious concept. Even if a solid is uncharged and isolated, its chemical potential is not a material property. It still depends unavoidably on the surface properties: their contamination, roughness, and angular orientation relative to the atomic crystal structure. If you mentally take a solid attached to other solids out to isolate it, then what are you to make of the condition of the surfaces that were previously in contact with other solids?

Because of such concerns, nowadays many physicists disdain the concept of an intrinsic chemical potential and simply refer to μ as "the" chemical potential. Note that this means that the actual value of the chemical potential depends on the detailed conditions that the solid is in. But then, so do the electron energy levels. The location of the chemical potential relative to the spectrum is well defined regardless of the electrostatic potential.

And the chemical potentials of solids in contact and in thermal equilibrium still line up.

The Fermi-Dirac distribution is also known as the "Fermi factor." Note that in proper quantum terms, it gives the probability that a state is occupied by an electron.

Key Points

- The Fermi-Dirac distribution gives the number of electrons, or other fermions, per single-particle state for a macroscopic system at a nonzero temperature.
- Typically, the effects of nonzero temperature remain restricted to a, relatively speaking, small number of electrons near the Fermi energy.
- These electrons are within a distance comparable to $k_B T$ of the Fermi energy. They pick up a thermal energy that is also comparable to $k_B T$.
- Because of the small number of electrons involved, the effect on the heat capacity can usually be ignored.
- When solids are in electrical contact and in thermal equilibrium, their (electro)chemical potentials / Fermi levels / Fermi brims / whatever line up.

6.14 Maxwell-Boltzmann Distribution

The previous sections showed that the thermal statistics of a system of identical bosons is normally dramatically different from that of a system of identical fermions. However, if the temperature is high enough, and the box holding the particles big enough, the differences disappear. These are ideal gas conditions.

Under these conditions the average number of particles per single-particle state becomes much smaller than one. That average can then be approximated by the so-called

$$\boxed{\text{Maxwell-Boltzmann distribution:} \quad \iota^{\mathrm{d}} = \frac{1}{e^{(E^{\mathrm{p}}-\mu)/k_{\mathrm{B}}T}} \qquad \iota^{\mathrm{d}} \ll 1} \tag{6.21}$$

Here E^{p} is again the single-particle energy, μ the chemical potential, T the absolute temperature, and k_{B} the Boltzmann constant. Under the given conditions of a low particle number per state, the exponential is big enough that the ± 1 found in the Bose-Einstein and Fermi-Dirac distributions (6.9) and (6.19) can be ignored.

Figure 6.16 gives a picture of the distribution for noninteracting particles in a box. The energy spectrum to the right shows the average number of particles per state as the relative width of the red region. The wave number space to the left shows a typical system energy eigenfunction; states with a particle in them are in red.

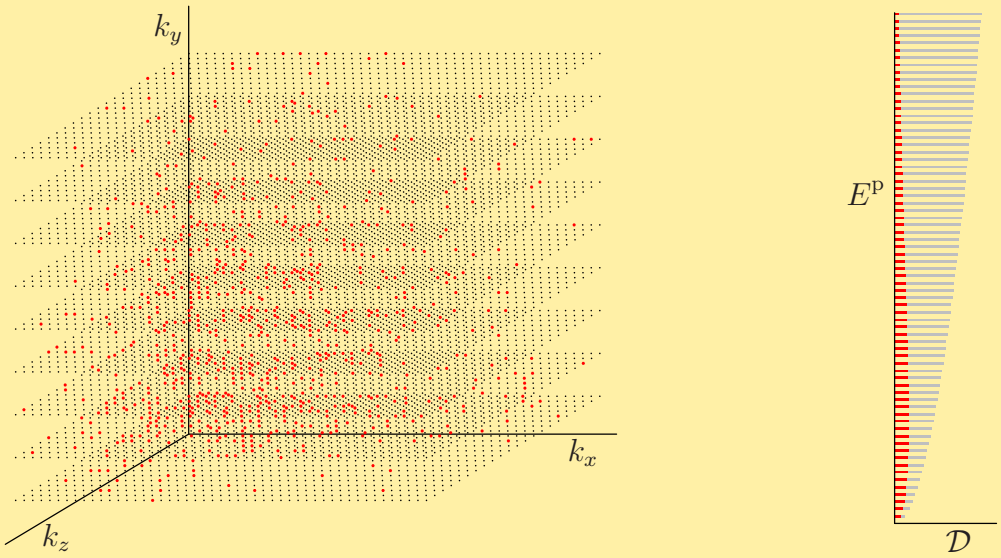


Figure 6.16: Particles at high-enough temperature and low-enough particle density.

Since the (anti) symmetrization requirements no longer make a difference, the Maxwell-Boltzmann distribution is often represented as applicable to "distinguishable" particles. But of course, where are you going to get a macroscopic number of, say, 10^{20} particles, each of a different type? The imagination boggles. Still, the "d" in ι^{d} refers to distinguishable.

The Maxwell-Boltzmann distribution was already known before quantum mechanics. The factor $e^{-E^{\mathrm{p}}/k_{\mathrm{B}}T}$ in it implies that the number of particles at a given energy decreases exponentially with the energy. A classical example

is the decrease of density with height in the atmosphere. In an equilibrium (i.e. isothermal) atmosphere, the number of molecules at a given height h is proportional to $e^{-mgh/k_{\rm B}T}$ where mgh is the gravitational potential energy of the molecules. (It should be noted that normally the atmosphere is not isothermal because of the heating of the earth surface by the sun and other effects.)

The example of the isothermal atmosphere can be used to illustrate the idea of intrinsic chemical potential. Think of the entire atmosphere as build up out of small boxes filled with particles. The walls of the boxes conduct some heat and they are very slightly porous, to allow an equilibrium to develop if you are very patient. Now write the energy of the particles as the sum of their gravitational potential energy plus an intrinsic energy (which is just their kinetic energy for the model of noninteracting particles). Similarly write the chemical potential as the sum of the gravitational potential energy plus an intrinsic chemical potential:

$$E^{\rm p} = mgh + E^{\rm p}_{\rm i} \qquad \mu = mgh + \mu_{\rm i}$$

Since $E^{\rm p} - \mu = E^{\rm p}_{\rm i} - \mu_{\rm i}$, the Maxwell-Boltzmann distribution is not affected by the switch to intrinsic quantities. But that implies that the relationship between kinetic energy, intrinsic chemical potential, and number of particles in each individual box is the same as if gravity was not there. In each box, the normal ideal gas law applies in terms of intrinsic quantities.

However, different boxes have different intrinsic chemical potentials. The entire system of boxes has one global temperature and one global chemical potential, since the porous walls make it a single system. But the global chemical potential that is the same in all boxes includes gravity. That makes the intrinsic chemical potential in boxes at different heights different, and with it the number of particles in the boxes.

In particular, boxes at higher altitudes have less molecules. Compare states with the same intrinsic, kinetic, energy for boxes at different heights. According to the Maxwell-Boltzmann distribution, the number of particles in a state with intrinsic energy $E^{\rm p}_{\rm i}$ is $1/e^{(E^{\rm p}_{\rm i}+mgh-\mu)/k_{\rm B}T}$. That decreases with height proportional to $e^{-mgh/k_{\rm B}T}$, just like classical analysis predicts.

Now suppose that you make the particles in one of the boxes hotter. There will then be a flow of heat out of that box to the neighboring boxes until a single temperature has been reestablished. On the other hand, assume that you keep the temperature unchanged, but increase the chemical potential in one of the boxes. That means that you must put more particles in the box, because the Maxwell-Boltzmann distribution has the number of particles per state equal to $e^{\mu/k_{\rm B}T}$. The excess particles will slowly leak out through the slightly porous walls until a single chemical potential has been reestablished. Apparently, then, too high a chemical potential promotes particle diffusion away from a site, just like too high a temperature promotes thermal energy diffusion away from a site.

While the Maxwell-Boltzmann distribution was already known classically, quantum mechanics adds the notion of discrete energy states. If there are more

energy states at a given energy, there are going to be more particles at that energy, because (6.21) is per state. For example, consider the number of thermally excited atoms in a thin gas of hydrogen atoms. The number I_2 of atoms that are thermally excited to energy E_2 is in terms of the number I_1 with the ground state energy E_1:

$$\frac{I_2}{I_1} = \frac{8}{2} e^{-(E_2-E_1)/k_{\mathrm{B}}T}$$

The final exponential is due to the Maxwell-Boltzmann distribution. The leading factor arises because there are eight electron states at energy E_2 and only two at energy E_1 in a hydrogen atom. At room temperature $k_{\mathrm{B}}T$ is about 0.025 eV, while $E_2 - E_1$ is 10.2 eV, so there are not going to be any thermally excited atoms at room temperature.

Key Points

- The Maxwell-Boltzmann distribution gives the number of particles per single-particle state for a macroscopic system at a nonzero temperature.
- It assumes that the particle density is low enough, and the temperature high enough, that (anti) symmetrization requirements can be ignored.
- In particular, the average number of particles per single-particle state should be much less than one.
- According to the distribution, the average number of particles in a state decreases exponentially with its energy.
- Systems for which the distribution applies can often be described well by classical physics.
- Differences in chemical potential promote particle diffusion.

6.15 Thermionic Emission

The valence electrons in a block of metal have tremendous kinetic energy, of the order of electron volts. These electrons would like to escape the confines of the block, but attractive forces exerted by the nuclei hold them back. However, if the temperature is high enough, typically 1 000 to 2 500 K, a few electrons can pick up enough thermal energy to get away. The metal then emits a current of electrons. This is called "thermionic emission." It is important for applications such as electron tubes and fluorescent lamps.

The amount of thermionic emission depends not just on temperature, but also on how much energy electrons inside the metal need to escape. Now the energies of the most energetic electrons inside the metal are best expressed in

terms of the Fermi energy level. Therefore, the energy required to escape is conventionally expressed relative to that level. In particular, the additional energy that a Fermi-level electron needs to escape is traditionally written in the form $e\varphi_w$ where e is the electron charge and φ_w is called the "work function." The magnitude of the work function is typically on the order of volts. That makes the energy needed for a Fermi-level electron to escape on the order of electron volts, comparable to atomic ionization energies.

The thermionic emission equation gives the current density of electrons as, {D.29},

$$j = AT^2 e^{-e\varphi_w/k_B T} \tag{6.22}$$

where T is the absolute temperature and k_B is the Boltzmann constant. The constant A is typically one quarter to one half of its theoretical value

$$A_{\text{theory}} = \frac{m_e e k_B}{2\pi^2 \hbar^3} \approx 1.2\ 10^6\ \text{amp/m}^2\text{K}^2 \tag{6.23}$$

Note that thermionic emission depends exponentially on the temperature; unless the temperature is high enough, extremely little emission will occur. You see the Maxwell-Boltzmann distribution at work here. This distribution is applicable since the number of electrons per state is very small for the energies at which the electrons can escape.

Despite the applicability of the Maxwell-Boltzmann distribution, classical physics cannot explain thermionic emission. That is seen from the fact that the constant A_{theory} depends nontrivially, and strongly, on $\hbar$. The dependence on quantum theory comes in through the density of states for the electrons that have enough energy to escape, {D.29}.

Thermionic emission can be helped along by applying an additional electric field $\mathcal{E}_{\text{ext}}$ that drives the electrons away from the surface of the solid. That is known as the "Schottky effect." The electric field has the approximate effect of lowering the work function value by an amount, {D.29},

$$\sqrt{\frac{e\mathcal{E}_{\text{ext}}}{4\pi\epsilon_0}} \tag{6.24}$$

For high-enough electric fields, significant numbers of electrons may also "tunnel" out due to their quantum uncertainty in position. That is called "field emission." It depends exponentially on the field strength, which must be very high as the quantum uncertainty in position is small.

It may be noted that the term "thermionic emission" may be used more generally to indicate the flow of charge carriers, either electrons or ions, over a potential barrier. Even for standard thermionic emission, it should be cautioned that the work function depends critically on surface conditions. For example, surface pollution can dramatically change it.

Key Points

- Some electrons can escape from solids if the temperature is sufficiently high. That is called thermionic emission.
- The work function is the minimum energy required to take a Fermi-level electron out of a solid, per unit charge.
- An additional electric field can help the process along, in more ways than one.

6.16 Chemical Potential and Diffusion

The chemical potential, or Fermi level, that appears in the Fermi-Dirac distribution is very important for solids in contact. If two solids are put in electrical contact, at first electrons will diffuse to the solid with the lower chemical potential. It is another illustration that differences in chemical potential cause particle diffusion.

Of course the diffusion cannot go on forever. The electrons that transfer to the solid with the lower chemical potential will give it a negative charge. They will also leave a net positive charge behind on the solid with the higher chemical potential. Therefore, eventually an electrostatic force builds up that terminates the further transfer of electrons. With the additional electrostatic contribution, the chemical potentials of the two solids have then become equal. As it should. If electrons can transfer from one solid to the other, the two solids have become a single system. In thermal equilibrium, a single system should have a single Fermi-Dirac distribution with a single chemical potential.

The transferred net charges will collect at the surfaces of the two solids, mostly where the two meet. Consider in particular the contact surface of two metals. The interiors of the metals have to remain completely free of net charge, or there would be a variation in electric potential and a current would flow to eliminate it. The metal that initially has the lower Fermi energy receives additional electrons, but these stay within an extremely thin layer at its surface. Similarly, the locations of missing electrons in the other metal stay within a thin layer at its surface. Where the two metals meet, a "double layer" exists; it consists of a very thin layer of highly concentrated negative net charges next to a similar layer of highly concentrated positive net charges. Across this double layer, the mean electrostatic potential changes almost discontinuously from its value in the first metal to that in the second. The step in electrostatic potential is called the "Galvani potential."

Galvani potentials are not directly measurable; attaching voltmeter leads to the two solids adds two new contact surfaces whose potentials will change the measured potential difference. More specifically, they will make the measured

potential difference exactly zero. To see why, assume for simplicity that the two leads of the voltmeter are made of the same material, say copper. All chemical potentials will level up, including those in the two copper leads of the meter. But then there is no way for the actual voltmeter to see any difference between its two leads.

Of course, it would have to be so. If there really was a net voltage in thermal equilibrium that could move a voltmeter needle, it would violate the second law of thermodynamics. You cannot get work for nothing.

Note however that if some contact surfaces are at different temperatures than others, then a voltage can in fact be measured. But the physical reason for that voltage is not the Galvani potentials at the contact surfaces. Instead diffusive processes in the bulk of the materials cause it. See section 6.28.2 for more details. Here it must suffice to note that the usable voltage is powered by temperature differences. That does not violate the second law; you are depleting temperature differences to get whatever work you extract from the voltage.

Similarly, chemical reactions can produce usable electric power. That is the principle of the battery. It too does not violate the second law; you are using up chemical fuel. The chemical reactions do physically occur at contact surfaces.

Somewhat related to Galvani potentials, there is an electric field in the gap between two different metals that are in electrical contact elsewhere. The corresponding change in electric potential across the gap is called the "contact potential" or "Volta potential."

As usual, the name is poorly chosen: the potential does not occur at the contact location of the metals. In fact, you could have a contact potential between different surfaces of the same metal, if the two surface properties are different. "Surface potential difference" or "gap potential" would have been a much more reasonable term. Only physicists would describe what really is a "gap potential" as a "contact potential."

The contact potential is equal to the difference in the work functions of the surfaces of the metals. As discussed in the previous section, the work function is the energy needed to take a Fermi-level electron out of the solid, per unit charge. To see why the contact potential equals the difference in work functions, imagine taking a Fermi-level electron out of the first metal, moving it through the gap, and putting it into the second metal. Since the electron is back at the same Fermi level that it started out at, the net work in this process should be zero. But if the work function of the second metal is different from the first, putting the electron back in the second metal does not recover the work needed to take it out of the first metal. Then electric work in the gap must make up the difference.

Key Points

- When two solids are brought in contact, their chemical potentials, or Fermi levels, must line up. A double layer of positive and negative

charges forms at the contact surface between the solids. This double layer produces a step in voltage between the interiors of the solids.

 There is a voltage difference in the gap between two metals that are electrically connected and have different work functions. It is called the contact potential.

6.17 Intro to the Periodic Box

This chapter so far has shown that lots can be learned from the simple model of noninteracting particles inside a closed box. The biggest limitation of the model is particle motion. Sustained particle motion is hindered by the fact that the particles cannot penetrate the walls of the box.

One way of dealing with that is to make the box infinitely large. That produces motion in infinite and empty space. It can be done, as shown in chapter 7.9 and following. However, the analysis is nasty, as the eigenfunctions cannot be properly normalized. In many cases, a much simpler approach is to assume that the particles are in a finite, but periodic box. A particle that exits such a box through one side reenters it at the same time through the opposing side.

To understand the idea, consider the one-dimensional case. Studying one-dimensional motion along an infinite straight line $-\infty < x < \infty$ is typically nasty. One-dimensional motion along a circle is likely to be easier. Unlike the straight line, the circumference of the circle, call it ℓ_x, is finite. So you can define a coordinate x along the circle with a finite range $0 < x < \ell_x$. Yet despite the finite circumference, a particle can keep moving along the circle without getting stuck. When the particle reaches the position $x = \ell_x$ along the circle, it is back at its starting point $x = 0$. It leaves the defined x-range through $x = \ell_x$, but it reenters it at the same time through $x = 0$. The position $x = \ell_x$ is physically exactly the same point as $x = 0$.

Similarly a periodic box of dimensions ℓ_x, ℓ_y, and ℓ_z assumes that $x = \ell_x$ is physically the same as $x = 0$, $y = \ell_y$ the same as $y = 0$, and $z = \ell_z$ the same as $z = 0$. That is of course hard to visualize. It is just a mathematical trick, but one that works well. Typically at the end of the analysis you take the limit that the box dimensions become infinite. That makes this artificial box disappear and you get the valid infinite-space solution.

The biggest difference between the closed box and the periodic box is linear momentum. For noninteracting particles in a periodic box, the energy eigenfunctions can be taken to be also eigenfunctions of linear momentum $\vec{\hat{p}}$. They then have definite linear momentum in addition to definite energy. In fact, the linear momentum is just a scaled wave number vector; $\vec{p} = \hbar\vec{k}$. That is discussed in more detail in the next section.

Key Points

- A periodic box is a mathematical concept that allows unimpeded motion of the particles in the box. A particle that exits the box through one side reenters it at the opposite side at the same time.
- For a periodic box, the energy eigenfunctions can be taken to be also eigenfunctions of linear momentum.

6.18 Periodic Single-Particle States

The single-particle quantum states, or energy eigenfunctions, for noninteracting particles in a closed box were given in section 6.2, (6.2). They were a product of a sine in each axial direction. Those for a periodic box can similarly be taken to be a product of a sine or cosine in each direction. However, it is usually much better to take the single-particle energy eigenfunctions to be exponentials:

$$\boxed{\psi^{\rm p}_{n_x n_y n_z}(\vec{r}) = \mathcal{V}^{-\frac{1}{2}} e^{{\rm i}(k_x x + k_y y + k_z z)} = \mathcal{V}^{-\frac{1}{2}} e^{{\rm i}\vec{k}\cdot\vec{r}}} \tag{6.25}$$

Here $\mathcal{V}$ is the volume of the periodic box, while $\vec{k} = (k_x, k_y, k_z)$ is the "wave number vector" that characterizes the state.

One major advantage of these eigenfunctions is that they are also eigenfunction of linear momentum. For example. the linear momentum in the x-direction equals $p_x = \hbar k_x$. That can be verified by applying the x-momentum operator $\hbar\partial/{\rm i}\partial x$ on the eigenfunction above. The same for the other two components of linear momentum, so:

$$\boxed{p_x = \hbar k_x \quad p_y = \hbar k_y \quad p_z = \hbar k_z \qquad \vec{p} = \hbar\vec{k}} \tag{6.26}$$

This relationship between wave number vector and linear momentum is known as the "de Broglie relation."

The reason that the momentum eigenfunctions are also energy eigenfunctions is that the energy is all kinetic energy. It makes the energy proportional to the square of linear momentum. (The same is true inside the closed box, but momentum eigenstates are not acceptable states for the closed box. You can think of the surfaces of the closed box as infinitely high potential energy barriers. They reflect the particles and the energy eigenfunctions then must be a 50/50 mix of forward and backward momentum.)

Like for the closed box, for the periodic box the single-particle energy is still given by

$$\boxed{E^{\rm p} = \frac{\hbar^2}{2m}k^2 \qquad k \equiv \sqrt{k_x^2 + k_y^2 + k_z^2}} \tag{6.27}$$

That may be verified by applying the kinetic energy operator on the eigenfunctions. It is simply the Newtonian result that the kinetic energy equals $\frac{1}{2}mv^2$ since the velocity is $v = p/m$ by the definition of linear momentum and $p = \hbar k$ in quantum terms.

Unlike for the closed box however, the wave numbers k_x, k_y, and k_z are now constrained by the requirement that the box is periodic. In particular, since $x = \ell_x$ is supposed to be the same physical plane as $x = 0$ for a periodic box, $e^{\mathrm{i}k_x\ell_x}$ must be the same as $e^{\mathrm{i}k_x 0}$. That restricts $k_x\ell_x$ to be an integer multiple of 2π, (2.5). The same for the other two components of the wave number vector, so:

$$\boxed{k_x = n_x \frac{2\pi}{\ell_x} \qquad k_y = n_y \frac{2\pi}{\ell_y} \qquad k_z = n_z \frac{2\pi}{\ell_z}} \tag{6.28}$$

where the quantum numbers n_x, n_y, and n_z are integers.

In addition, unlike for the sinusoidal eigenfunctions of the closed box, zero and negative values of the wave numbers must now be allowed. Otherwise the set of eigenfunctions will not be complete. The difference is that for the closed box, $\sin(-k_x x)$ is just the negative of $\sin(k_x x)$, while for the periodic box, $e^{-\mathrm{i}k_x x}$ is not just a multiple of $e^{\mathrm{i}k_x x}$ but a fundamentally different function.

Figure 6.17 shows the wave number space for a system of electrons in a periodic box. The wave number vectors are no longer restricted to the first quadrant like for the closed box in figure 6.11; they now fill the entire space. In the ground state, the states occupied by electrons, shown in red, now form a complete sphere. For the closed box they formed just an octant of one. The Fermi surface, the surface of the sphere, is now a complete spherical surface.

It may also be noted that in later parts of this book, often the wave number vector or momentum vector is used to label the eigenfunctions:

$$\psi^{\mathrm{p}}_{n_x n_y n_z}(\vec{r}) = \psi^{\mathrm{p}}_{k_x k_y k_z}(\vec{r}) = \psi^{\mathrm{p}}_{p_x p_y p_z}(\vec{r})$$

In general, whatever is the most relevant to the analysis is used as label. In any scheme, the single-particle state of lowest energy is $\psi^{\mathrm{p}}_{000}(\vec{r})$; it has zero energy, zero wave number vector, and zero momentum.

Key Points

- The energy eigenfunctions for a periodic box are usually best taken to be exponentials. Then the wave number values can be both positive and negative.
- The single-particle kinetic energy is still $\hbar^2 k^2/2m$.
- The momentum is $\hbar\vec{k}$.
- The eigenfunction labelling may vary.

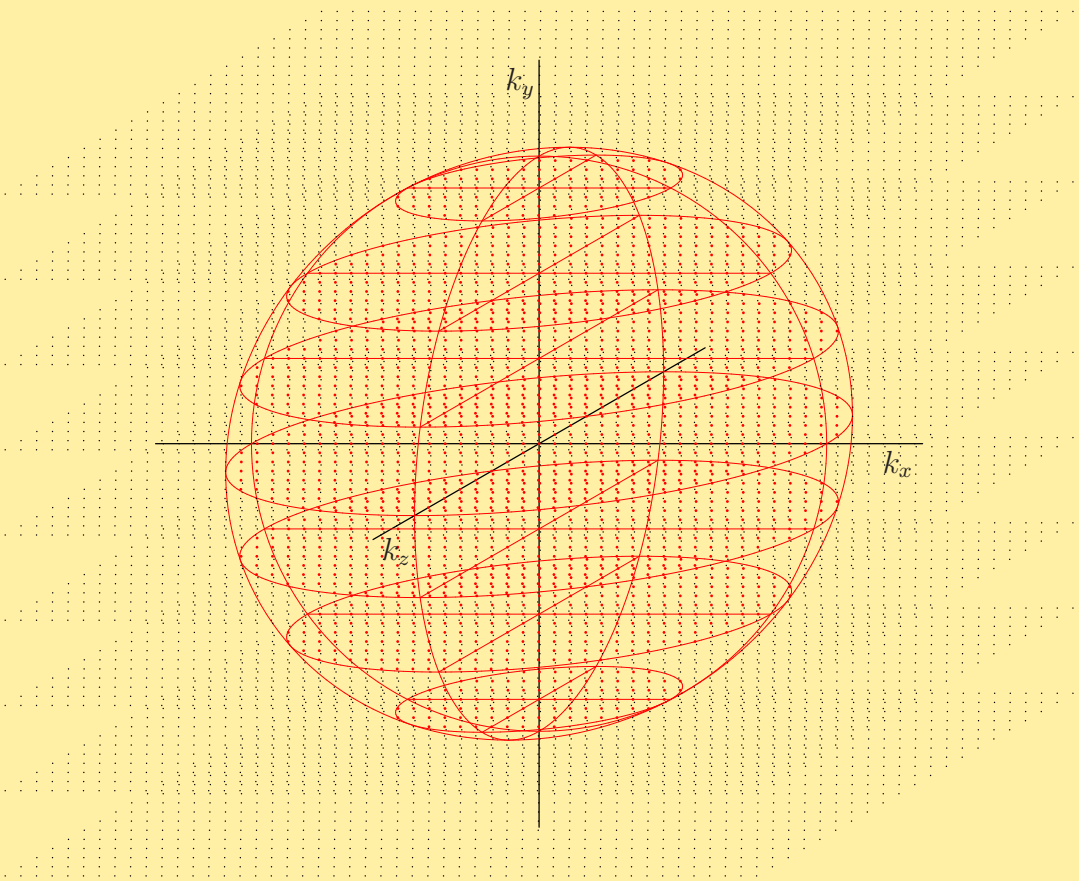


Figure 6.17: Ground state of a system of noninteracting electrons, or other fermions, in a periodic box.

6.19 DOS for a Periodic Box

The density of states is the number of single-particle states per unit energy range. It turns out that the formulae for the density of states given in section 6.3 may be used for the periodic box as well as for the closed box. A box can hold about the same number of particles per unit volume whether the boundary conditions are periodic or not.

It is not that hard to verify. For a periodic box, the wave numbers can be both positive and negative, not just positive like for a closed box. On the other hand, a comparison of (6.3) and (6.28) shows that the wave number spacing for a periodic box is twice as large as for a corresponding closed box. That cancels the effect of the additional negative wave numbers and the total number of wave number vectors in a given energy range remains the same. Therefore the density of states is the same.

For the periodic box it is often convenient to have the density of states on a linear momentum basis. It can be found by substituting $k = p/\hbar$ into (6.5).

That gives the number of single-particle states $\mathrm{d}N$ in a momentum range of size $\mathrm{d}p$ as:

$$\boxed{\mathrm{d}N = \mathcal{V}\mathcal{D}_p \,\mathrm{d}p \qquad \mathcal{D}_p = \frac{2s+1}{2\pi^2\hbar^3}p^2} \tag{6.29}$$

Here $\mathcal{D}_p$ is the density of states per unit momentum range and unit volume. Also, s is again the particle spin. Recall that $2s+1$ becomes $2s$ for photons.

The staging behavior due to confinement gets somewhat modified compared to section 6.12, since zero wave numbers are now included. The analysis is however essentially unchanged.

Key Points

- The density of states is essentially the same for a periodic box as for a closed one.

6.20 Intro to Electrical Conduction

Some of the basic physics of electrical conduction in metals can be understood using a very simple model. That model is a free-electron gas, i.e. noninteracting electrons, in a periodic box.

The classical definition of electric current is moving charges. That can readily be converted to quantum terms for noninteracting electrons in a periodic box. The single-particle energy states for these electrons have definite velocity. That velocity is given by the linear momentum divided by the mass.

Consider the possibility of an electric current in a chosen x-direction. Figure 6.18 shows a plot of the single-particle energy E^{p} against the single-particle velocity v_x^{p} in the x-direction. The states that are occupied by electrons are shown in red. The parabolic outer boundary reflects the classical expression $E^{\mathrm{p}} = \frac{1}{2}m_{\mathrm{e}}v^{\mathrm{p}2}$ for the kinetic energy: for the single-particle states on the outer boundary, the velocity is purely in the x-direction.

In the system ground state, shown to the left in figure 6.18, no current will flow, because there are just as many electrons that move toward negative x as there are that move towards positive x. To get net electron motion in the x-direction, electrons must be moved from states that have negative velocity in the x-direction to states that have positive velocity. That is indicated to the right in figure 6.18. The asymmetric occupation of states now produces net electron motion in the positive x-direction. That produces a current in the negative x-direction because of the fact that the charge $-e$ of electrons is negative.

Note that the electrons must pick up a bit of additional energy when they are moved from states with negative velocity to states with positive velocity. That is because the Pauli exclusion principle forbids the electrons from entering the lower energy states of positive velocity that are already filled with electrons.

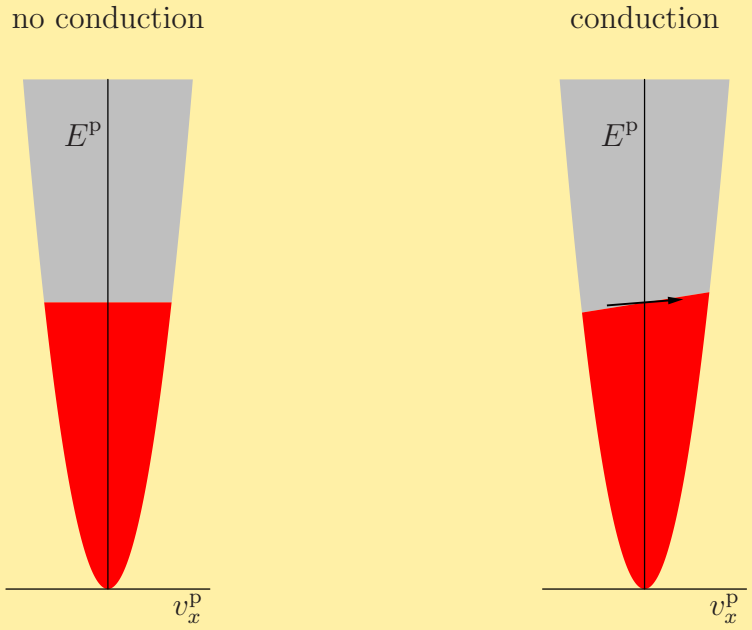


Figure 6.18: Conduction in the free-electron gas model.

However, the required energy is small. You might just briefly turn on an external voltage source to produce an electric field that gets the electrons moving. Then you can turn off the voltage source again, because once set into motion, the noninteracting electrons will keep moving forever.

In physical terms, it is not really that just a few electrons make a big velocity change from negative to positive due to the applied voltage. In quantum mechanics electrons are completely indistinguishable, and all the electrons are involved equally in the changes of state. It is better to say that all electrons acquire a small additional drift velocity Δv_x^{p} in the positive x-direction. In terms of the wave number space figure 6.17, this shifts the entire sphere of occupied states a bit towards the right, because velocity is proportional to wave number for a free-electron gas.

The net result is still the energy versus velocity distribution shown to the right in figure 6.18. Electrons at the highest energy levels with positive velocities go up a bit in energy. Electrons at the highest energy levels with negative velocities go down a bit in energy. The electrons at lower energy levels move along to ensure that there is no more than one electron in each quantum state. The fact remains that the system of electrons picks up a bit of additional energy. (The last subsection of derivation {D.45} discusses the effect of the applied voltage in more detail.)

Conduction electrons in an actual metal wire behave similar to free electrons. However, they must move around the metal atoms, which are normally arranged in some periodic pattern called the crystal structure. The conduction electrons

will periodically get scattered by thermal vibrations of the crystal structure, (in quantum terms, by phonons), and by crystal structure imperfections and impurities. That kills off their organized drift velocity $\Delta v_x^{\rm p}$, and a small permanent electric field is required to replenish it. In other words, there is resistance. But it is not a large effect. For one, in macroscopic terms the conduction electrons in a metal carry quite a lot of charge per unit volume. So they do not have to go fast. Furthermore, conduction electrons in copper or similar good metal conductors may move for thousands of Ångstroms before getting scattered, slipping past thousands of atoms. Electrons in extremely pure copper at liquid helium temperatures may even move millimeters or more before getting scattered. The average distance between scattering events, or "collisions," is called the "free path" length ℓ. It is very large on an atomic scale.

Of course, that does not make much sense from a classical point of view. Common sense says that a point-size classical electron in a solid should pretty much bounce off every atom it encounters. Therefore the free path of the electrons should be of the order of a single atomic spacing, not thousands of atoms or much more still. However, in quantum mechanics electrons are not particles with a definite position. Electrons are described by a wave function. It turns out that electron waves can propagate through perfect crystals without scattering, much like electromagnetic waves can. The free-electron gas wave functions adapt to the crystal structure, allowing the electrons to flow past the atoms without reflection.

It is of some interest to compare the quantum picture of conduction to that of a classical, nonquantum, description. In the classical picture, all conduction electrons would have a random thermal motion. The average velocity v of that motion would be proportional to $\sqrt{k_{\rm B}T/m_{\rm e}}$, with $k_{\rm B}$ the Boltzmann constant, T the absolute temperature, and $m_{\rm e}$ the electron mass. In addition to this random thermal motion in all directions, the electrons would also have a small organized drift velocity $\Delta v_x^{\rm p}$ in the positive x-direction that produces the net current. This organized motion would be created by the applied electric field in between collisions. Whenever the electrons collide with atoms, they lose much of their organized motion, and the electric field has to start over again from scratch.

Based on this picture, a ballpark expression for the classical conductivity can be written down. First, by definition the current density j_x equals the number of conduction electrons per unit volume $i_{\rm e}$, times the electric charge $-e$ that each carries, times the small organized drift velocity $\Delta v_x^{\rm p}$ in the x-direction that each has:

$$j_x = -i_{\rm e} e \Delta v_x^{\rm p} \tag{6.30}$$

The drift velocity $\Delta v_x^{\rm p}$ produced by the electric field between collisions can be found from Newton's second law as the force on an electron times the time interval between collisions during which this force acts and divided by the electron

mass. The average drift velocity would be half that, assuming for simplicity that the drift is totally lost in collisions, but the half can be ignored in the ballpark anyway. The force on an electron equals $-e\mathcal{E}_x$ where $\mathcal{E}_x$ is the electric field due to the applied voltage. The time between collisions can be computed as the distance between collisions, which is the free path length ℓ, divided by the velocity of motion v. Since the drift velocity is small compared to the random thermal motion, v can be taken to be the thermal velocity. The "conductivity" σ is the current density per unit electric field, so putting it all together,

$$\sigma \sim \frac{i_e e^2 \ell}{m_e v} \tag{6.31}$$

Neither the thermal velocity v nor the free path ℓ will be the same for all electrons, so suitable averages have to be used in more detailed expressions. The "resistivity" is defined as the reciprocal of the conductivity, so as $1/\sigma$. It is the resistance of a unit cube of material.

For metals, things are a bit different because of quantum effects. In metals random collisions are restricted to a small fraction of electrons at the highest energy levels. These energy levels are characterized by the Fermi energy, the highest occupied energy level in the spectrum to the left in figure 6.18. Electrons of lower energies do not have empty states nearby to be randomly scattered into. The velocity of electrons near the Fermi energy is much larger than the thermal value $\sqrt{k_B T/m_e}$, because there are much too few states with thermal-level energies to hold all conduction electrons, section 6.10. The bottom line is that for metals, in the ballpark for the conductivity the free path length ℓ and velocity v of the Fermi-level electrons must be used. In addition, the electron mass m_e may need to be changed into an effective one to account for the forces exerted by the crystal structure on the electrons. That will be discussed in more detail in section 6.22.3.

The classical picture works much better for semiconductors, since these have much less conduction electrons than would be needed to fill all the quantum states available at thermal energies. The mass correction remains required.

Key Points

- The free-electron gas can be used to understand conduction in metals in simple terms.
- In the absence of a net current the electrons are in states with velocities in all directions. The net electron motion therefore averages out to zero.
- A net current is achieved by giving the electrons an additional small organized motion.
- The energy needed to do this is small.

- In real metals, the electrons lose their organized motion due to collisions with phonons and crystal imperfections. Therefore a small permanent voltage must be applied to maintain the net motion. That means that there is electrical resistance. However, it is very small for typical metals.

6.21 Intro to Band Structure

Quantum mechanics is essential to describe the properties of solid materials, just as it is for lone atoms and molecules. One well-known example is superconductivity, in which current flows without any resistance. The complete absence of any resistance cannot be explained by classical physics, just like superfluidity cannot for fluids.

But even *normal* electrical conduction simply cannot be explained without quantum theory. Consider the fact that at ordinary temperatures, typical metals have electrical resistivities of a few times 10^{-8} ohm-m (and up to a hundred thousand times less still at very low temperatures), while Wikipedia lists a resistance for teflon of up to 10^{24} ohm-m. (Teflon's "one-minute" resistivity can be up to 10^{19} ohm-m.) That is a difference in resistance between the best conductors and the best insulators by over thirty orders of magnitude!

There is simply no way that classical physics could even begin to explain it. As far as classical physics is concerned, all of these materials are quite similar combinations of positive nuclei and negative electrons.

Consider an ordinary sewing needle. You would have as little trouble supporting its tiny 60 mg weight as a metal has conducting electricity. But multiply it by 10^{30}. Well, don't worry about supporting its weight. Worry about the entire earth coming up over your ears and engulfing you, because the needle now has ten times the mass of the earth. That is how widely different the electrical conductivities of solids are.

Only quantum mechanics can explain why it is possible, by making the electron energy levels discrete, and more importantly, by grouping them together in "bands."

Key Points

- Even excluding superconductivity, the electrical conductivities of solids vary enormously.

6.21.1 Metals and insulators

To understand electrical conduction in solids requires consideration of their electron energy levels.

Typical energy spectra are sketched in figure 6.19. The spectrum of a free-electron gas, noninteracting electrons in a box, is shown to the left. The energy $E^{\rm p}$ of the single-particle states is shown along the vertical axis. The energy levels allowed by quantum mechanics start from zero and reach to infinity. The energy levels are spaced many orders of magnitude more tightly together than the hatching in the figure can indicate. For almost all practical purposes, the energy levels form a continuum. In the ground state, the electrons fill the lowest of these energy levels, one electron per state. In the figure, the occupied states are shown in red. For a macroscopic system, the number of electrons is practically speaking infinite, and so is the number of occupied states.

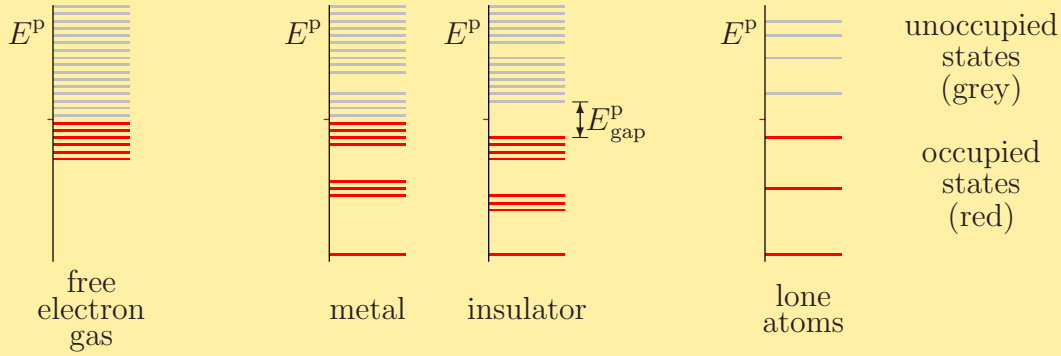


Figure 6.19: Sketch of electron energy spectra in solids at absolute zero temperature. (No attempt has been made to picture a density of states). Far left: the free-electron gas has a continuous band of extremely densely spaced energy levels. Far right: lone atoms have only a few discrete electron energy levels. Middle: actual metals and insulators have energy levels grouped into densely spaced bands separated by gaps. Insulators completely fill up the highest occupied band.

However, the free-electron gas assumes that there are no forces on the electrons. Inside a solid, this would only be true if the electric charges of the nuclei and fellow electrons would be homogeneously distributed throughout the entire solid. In that case the forces come equally from all directions and cancel each other out perfectly. In a true solid, forces from different directions do tend to cancel each other out, but this is far from perfect. For example, an electron very close to one particular nucleus experiences a strong attraction from that nucleus, much too strong for the rest of the solid to cancel.

The diametrical opposite of the free-electron gas picture is the case that the atoms of the solid are spaced so far apart that they are essentially lone atoms. In that case, of course, the "solid" would not physically be a solid at all, but a thin gas. Lone atoms do not have a continuum of electron energy levels, but discrete ones, as sketched to the far right in figure 6.19. One basic example is the hydrogen spectrum shown in figure 4.8. Every lone atom in the system

has the exact same discrete energy levels. Widely spaced atoms do not conduct electricity, assuming that not enough energy is provided to ionize them. While for the free-electron gas conduction can be achieved by moving a few electrons to slightly higher energy levels, for lone atoms there *are no* slightly higher energy levels.

When the lone atoms are brought closer together to form a true solid, however, the discrete atomic energy levels broaden out into bands. In particular, the outer electrons start to interact strongly with surrounding atoms. The different forms that these interactions can take produce varying energies, causing initially equal electron energies to broaden into bands. The result is sketched in the middle of figure 6.19. The higher occupied energy levels spread out significantly. (The inner atomic electrons, having the most negative net energies, do not interact significantly with different atoms, and their energy levels do not broaden much. This is not just because these electrons are farther from the surrounding atoms, but also because the inner electrons have much greater kinetic and much more negative potential energy levels to start with.)

For metals, conduction now becomes possible. Electrons at the highest occupied energy level, the Fermi energy, can be moved to slightly higher energy levels to provide net motion in a particular direction. That is just like they can for a free-electron gas as discussed in the previous section. The net motion produces a current.

Insulators are different. As sketched in figure 6.19, they completely fill up the highest occupied energy band. That filled band is called the "valence band." The next higher and empty band is called the "conduction band."

Now it is no longer possible to prod electrons to slightly higher energy levels to create net motion. There are no slightly higher energy levels available; all levels in the valence band are already filled with electrons.

To create a state with net motion, some electrons would have to be moved to the conduction band. But that would require large amounts of energy. The minimum energy required is the difference between the top of the valence band and the bottom of the conduction band. This energy is appropriately called the "band gap" energy $E_{\text{gap}}^{\text{p}}$. It is typically of the order of electron volts, comparable to atomic potentials for outer electrons. That is in turn comparable to ionization energies, a great amount of energy on an atomic scale.

Resistance is determined for voltages low enough that Ohm's law applies. Such voltages do not provide anywhere near the energy required to move electrons to the conduction band. So the electrons in an insulator are stuck. They cannot achieve net motion at all. And without net motion, there is no current. That makes the resistance infinite. In this way the band gaps are responsible for the enormous difference in resistance between metals and insulators.

Note that a normal applied voltage will not have a significant effect on the band structure. Atomic potential energies are in terms of eV or more. For the applied voltage to compete with that would require a voltage drop comparable

to volts *per atom*. On a microscopic scale, the applied potential does not change the states.

Key Points

- Quantum mechanics allows only discrete energy levels for the electrons in a solid, and these levels group together in bands with gaps in between them.
- If the electrons fill the spectrum right up to a gap between bands, the electrons are stuck. It will require a large amount of energy to activate them to conduct electricity or heat. Such a solid is an insulator at absolute zero temperature.
- The filled band is called the valence band, and the empty band above it the conduction band.

6.21.2 Typical metals and insulators

If a material completely fills up its valence band with electrons, it is an insulator. But what materials would do that? This subsection gives a few rules of thumb.

One important rule is that the elements towards the left in the periodic table figure 5.8 are metals. A relatively small group of elements towards the right are nonmetals.

Consider first the alkali metals found in group I to the far left in the table. The lone atoms have only one valence electron per atom. It is in an atomic "s" state that can hold two electrons, chapter 5.9.4. Every spatial state, including the s state, can hold two electrons that differ in spin.

Now if the lone atoms are brought closer together to form a solid, the spatial states change. Their energy levels broaden out into a band. However, the total number of states does not change. One spatial state per atom stays one spatial state per atom. Since each spatial state can hold two electrons, and there is only one, the band formed from the s states is only half filled. Therefore, like the name says, the alkali metals are metals.

In helium the spatial 1s states are completely filled with the two electrons per atom. That makes solid helium an insulator. It should be noted that helium is only a solid at very low temperatures and very high pressures. The atoms are barely held together by very weak Van der Waals forces.

The alkaline metals found in group II of the periodic table also have two valence electrons per atom. So you would expect them to be insulators too. However, like the name says, the alkaline metals are metals. What happens is that the filled band originating from the atomic s states merges with an empty band originating from the atomic p states. That produces a partially filled combined band.

This does not apply to helium because there are no 1p states. The lowest empty energy states for helium are the 2s ones. Still, computations predict that helium will turn metallic at extremely high pressures. Compressing a solid has the primary effect of increasing the kinetic energy of the electrons. Roughly speaking, the kinetic energy is inversely proportional to the square of the electron spacing, compare the Fermi energy (6.16). And increasing the kinetic energy of the electrons brings them closer to a free-electron gas.

A case resembling that of helium is ionic materials in which the ions have a noble-gas electron structure. A basic example is salt, sodium chloride. These materials are insulators, as it takes significant energy to take apart the noble-gas electron configurations. See however the discussion of ionic conductivity later in this section.

Another case that requires explanation is hydrogen. Like the alkali metals, hydrogen has only one valence electron per atom. That is not enough to fill up the energy band resulting from the atomic 1s states. So you would expect solid hydrogen to be a metal. But actually, hydrogen is an insulator. What happens is that the energy band produced by the 1s states splits into two. And the lower half is completely filled with electrons.

The reason for the splitting is that in the solid, the hydrogen atoms combine pairwise into molecules. In an hydrogen molecule, there are not two separate spatial 1s states of equal energy, chapter 5.2. Instead, there is a lowered-energy *two-electron* spatial state in which the two electrons are symmetrically shared. There is also a raised-energy *two-electron* spatial state in which the two electrons are antisymmetrically shared. So there are now two energy levels with a gap in between them. The two electrons occupy the lower-energy symmetric state with opposite spins. In the solid, the hydrogen molecules are barely held together by weak Van der Waals forces. The interactions between the molecules are small, so the two molecular energy levels broaden only slightly into two thin bands. The gap between the filled symmetric states and the empty antisymmetric ones remains.

Note that sharing electrons in pairs involves a nontrivial interaction between the two electrons in each pair. The truth must be stretched a bit to fit it within the band theory idea of noninteracting electrons. Truly noninteracting electrons would have the spatial states of the hydrogen molecular *ion* available to them, chapter 4.6. Here the lower energy state is one in which a single electron is symmetrically shared between the atoms. And the higher energy state is one in which a single electron is antisymmetrically shared. In the model of noninteracting electrons, both electrons occupy the lower-energy single-electron spatial state, again with opposite spins. One problem with this picture is that the single-electron states do not take into account where the other electron is. There is then a significant chance that both electrons can be found around the same atom. In the correct two-electron state, the electrons largely avoid that. Being around the same atom would increase their energy, since the electrons

repel each other.

Note also that using the actual hydrogen molecular ion states may not be the best approach. It might be better to account for the presence of the other electron approximately using some nuclear shielding approach like the one used for atoms in chapter 5.9. An improved, but still approximate way of accounting for the second electron would be to use a so-called "Hartree-Fock" method. More generally, the most straightforward band theory approach tends to work better for metals than for insulators. Alternative numerical methods exist that work better for insulators. At the time of writing there is no simple magic bullet that works well for every material.

Group IV elements like diamond, silicon, and germanium pull a similar trick as hydrogen. They are insulators at absolute zero temperature. However, their 4 valence electrons per atom are not enough to fill the merged band arising from the s and p states. That band can hold 8 electrons per atom. Like hydrogen, a gap forms within the band. First the s and p states are converted into hybrids, chapter 5.11.4. Then states are created in which electrons are shared symmetrically between atoms and states in which they are shared antisymmetrically. There is an energy gap between these states. The lower energy states are filled with electrons and the higher energy states are empty, producing again an insulator. But unlike in hydrogen, each atom is now bonded to four others. That turns the entire solid into essentially one big molecule. These materials are much stronger and more stable than solid hydrogen. Like helium, hydrogen is only a solid at very low temperatures.

It may be noted that under extremely high pressures, hydrogen might become metallic. Not only that, as the smallest atom of them all, and in the absence of 1p atomic states, metallic hydrogen is likely to have some very unusual properties. It makes metallic hydrogen the holy grail of high pressure physics.

It is instructive to examine how the band theory of noninteracting electrons accounts for the fact that hydrogen is an insulator. Unlike the discussion above, band theory does not actually look at the number of valence electrons *per atom*. For one, a solid may consist of atoms of more than one kind. In general, crystalline solids consist of elementary building blocks called "primitive cells" that can involve several atoms. Band theory predicts the solid to be a metal if the number of electrons *per primitive cell* is odd. If the number of electrons per primitive cell is even, the material may be an insulator. In solid hydrogen each primitive cell holds a complete molecule, so there are two atoms per primitive cell. Each atom contributes an electron, so the number of electrons per primitive cell is even. According to band theory, that allows hydrogen to be an insulator. In a similar way group V elements can fill up their valence bands with an odd number of valence electrons per atom. And like hydrogen, diamond, silicon, and germanium have two atoms per primitive cell, reflecting the gap that forms in the merged s and p bands.

Of course, that cannot be the complete story. It does not explain why atoms towards the right in the periodic table would group together into primitive cells that allow them to be insulators. Why don't the atoms to the left in the periodic table do the same? Why don't the alkali metals group together in two-atom molecules like hydrogen does? Qualitatively speaking, metals are characterized by valence electrons that are relatively loosely bound. Suppose you compare the size of the 2s state of a lithium atom with the spacing of the atoms in solid lithium. If you do, you find that on average the 2s valence electron is no closer to the atom to which it supposedly "belongs" than to the neighboring atoms. Therefore, the electrons are what is called "delocalized." They are not bound to one specific location in the atomic crystal structure. So they are not really interested in helping bond "their" particular atom to its immediate neighbors. On the other hand, to the right in the periodic table, including hydrogen and helium, the valence electrons are much more tightly held. To delocalize them would require that the atoms would be squeezed much more tightly together. That does not happen under normal pressures because it produces very high kinetic energy of the electrons.

Where hydrogen refuses to be a metal with one valence electron per atom, boron refuses to do so with three. However, boron is very ambivalent about it. It does not really feel comfortable with either metallic or covalent behavior. A bit of impurity can readily turn it metallic. That great sensitivity to impurity makes the element very hard to study. At the time of writing, it is believed that boron has a covalent ground state under normal pressures. The convoluted crystal structure is believed to have a unit cell with either 12 or 106 atoms, depending on precise conditions.

In group IV, tin is metallic above 13 °C, as white tin, but covalent below this temperature, as grey tin. It is often difficult to predict whether an element is a metal or covalent near the middle of the periodic table. Lead, of course, is a metal.

It should further be noted that band theory can be in error because it ignores the interactions between the electrons. "Mott insulators" and "charge transfer insulators" are, as the name says, insulators even though conventional band theory would predict that they are metals.

Key Points

- In the periodic table, the group I, II, and III elements are normally metals.
- Hydrogen and helium are nonmetals. Don't ask about boron.
- The group IV elements diamond, silicon, and germanium are insulators at absolute zero temperature.

6.21.3 Semiconductors

Temperature can have significant effects on electrical conduction. As the previous section noted, higher temperature decreases the conduction in metals, as there are more crystal vibrations that the moving electrons can get scattered by. But a higher temperature also changes which energy states the electrons occupy. And that can produce semiconductors.

Figure 6.19 showed which energy states the electrons occupy at absolute zero temperature. There are no electrons with energies above the Fermi level indicated by the red tick mark. Figure 6.20 shows how that changes for a nonzero temperature. Now random thermal motion allows electrons to reach energy levels up to roughly $k_B T$ above the Fermi level. Here k_B is the Boltzmann constant and T the absolute temperature. This change in electron energies is described mathematically by the Fermi-Dirac distribution discussed earlier.

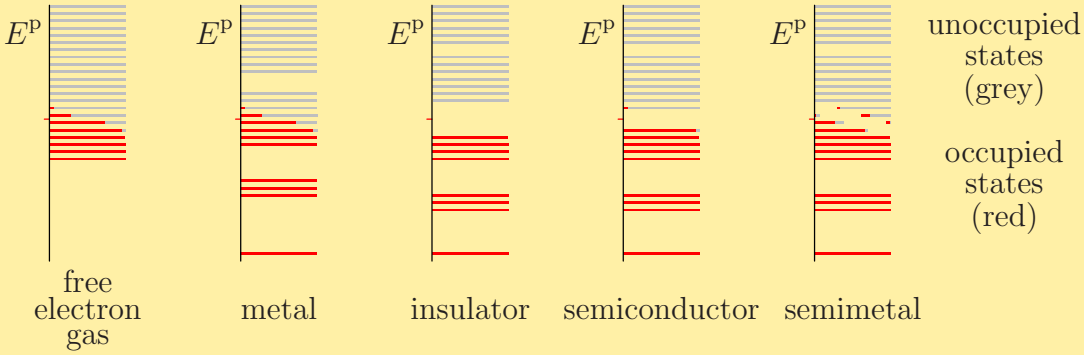


Figure 6.20: Sketch of electron energy spectra in solids at a nonzero temperature.

It does not make much difference for a free-electron gas or a metal. However, for an insulator it may make a dramatic difference. If the band gap is not too large compared to $k_B T$, random thermal motion will put a few very lucky electrons in the previously empty conduction band. These electrons can then be prodded to slightly higher energies to allow some electric current to flow. Also, the created "holes" in the valence band, the states that have lost their electrons, allow some electric current. Valence band electrons can be moved into holes that have a preferred direction of motion from states that do not. These electrons will then leave behind holes that have the opposite direction of motion.

It is often more convenient to think of the moving holes instead of the electrons as the electric current carriers in the valence band. Since a hole means that a negatively charged electron is *missing*, a hole acts much like a positively charged particle would.

Because both the electrons in the conduction band and the holes in the valence band allow some electrical conduction, the original insulator has turned

into what is called a "semiconductor."

The previous section mentioned that a classical picture of moving electrons simply does not work for metals. Their motion is much too much restrained by a lack of available empty energy states. However, the conduction band of semiconductors is largely empty. Therefore a classical picture works much better for the motion of the electrons in the conduction band of a semiconductor.

Key Points

- For semiconductors, conduction can occur because some electrons from the valence band are thermally excited to the conduction band.
- Both the electrons that get into the conduction band and the holes they leave behind in the valence band can conduct electricity.

6.21.4 Semimetals

One additional type of electron energy spectrum for solids should be mentioned. For a "semimetal," two distinct energy bands overlap slightly at the Fermi level. In terms of the simplistic spectra of figure 6.19, that would mean that semimetals are metals. Indeed they do allow conduction at absolute zero temperature. However, their further behavior is noticeably different from true metals because the overlap of the two bands is only small. One difference is that the electrical conduction of semimetals increases with temperature, unlike that of metals. Like for semiconductors, for semimetals a higher temperature means that there are more electrons in the upper band and more holes in the lower band. That effect is sketched to the far right in figure 6.20.

The classical semimetals are arsenic, antimony, and bismuth. Arsenic and antimony are not just semimetals, but also "metalloids," a group of elements whose chemical properties are considered to be intermediate between metals and nonmetals. But semimetal and metalloid are not the same thing. Semimetals do not have to consist of a single element. Conversely, metalloids include the semiconductors silicon and germanium.

A semimetal that is receiving considerable attention at the time of writing is graphite. Graphite consists of sheets of carbon atoms. A single sheet of carbon, called graphene, is right on the boundary between semimetal and semiconductor. A carbon nanotube can be thought of as a strip cut from a graphene sheet that then has its long edges attached together to produce a cylinder. Carbon nanotubes have electrical properties that are fundamentally different depending on the direction in which the strip is cut from the sheet. They can either be metallic or nonmetallic.

Key Points

- Semimetals have properties intermediate between metals and semiconductors.

6.21.5 Electronic heat conduction

The valence electrons in metals are not just very good conductors of electricity, but also of heat. In insulators electrons do not assist in heat conduction; it takes too much energy to excite them. However, atomic vibrations in solids can conduct heat too. For example, diamond, an excellent electrical insulator, is also an excellent conductor of heat. Therefore the differences in heat conduction between solids are not by far as large as those in electrical conduction. Because atoms can conduct significant heat, no solid material will be a truly superb thermal insulator. Practical thermal insulators are highly porous materials whose volume consists largely of voids.

Key Points

- Electrons conduct heat very well, but atoms can do it too.
- Practical thermal insulators use voids to reduce atomic heat conduction.

6.21.6 Ionic conductivity

It should be mentioned that electrons do not have an absolute monopoly on electrical conduction in solids. A different type of electrical conduction is possible in ionic solids. These solids consist of a mixture of positively and negatively charged ions. Positive ions, or "cations," are atoms that have lost one or more electrons. Negative ions, or "anions," are atoms that have absorbed one or more additional electrons. A simple example of a ionic solid is salt, which consists of Na^+ sodium cations and Cl^- chlorine anions. For ionic solids a small amount of electrical conduction may be possible due to motion of the ions. This requires defects in the atomic crystal structure in order to give the atoms some room to move.

Typical defects include "vacancies," in which an atom is missing from the crystal structure, and "interstitials," in which an additional atom has been forced into one of the small gaps between the atoms in the crystal. Now if a ion gets removed from its normal position in the crystal to create a vacancy, it must go somewhere. One possibility is that it gets squeezed in between the other atoms in the crystal. In that case both a vacancy and an interstitial have been produced at the same time. Such a combination of a vacancy and an interstitial is called a "Frenkel defect." Another possibility occurs in, for example, salt;

along with the original vacancy, a vacancy for a ion of the opposite kind is created. Such a combination of two opposite vacancies is called a "Schottky defect." In this case there is no need to squeeze an atom in the gaps in the crystal structure; there are now equal numbers of ions of each kind to fill the surrounding normal crystal sites. Creating defects in Frenkel or Schottky pairs ensures that the complete crystal remains electrically neutral as it should.

Impurities are another important defect. For example, in salt a Ca^{2+} calcium ion might be substituted for a Na^{+} sodium ion. The calcium ion has the charge of two sodium ions, so a sodium vacancy ensures electric neutrality of the crystal. In yttria-stabilized zirconia, (YSZ), oxygen vacancies are created in zirconia, ZrO_2, by replacing some Zr^{4+} zirconium ions with Y^{3+} yttrium ones. Calcium ions can also be used. The oxygen vacancies allow mobility for the oxygen ions. That is important for applications such as oxygen sensors and solid oxide fuel cells.

For salt, the main conduction mechanism is by natrium vacancies. But the ionic conductivity of salt is almost immeasurably small at room temperature. That is due to the high energy needed to create Schottky defects and for natrium ions to migrate into the natrium vacancies. Indeed, whatever little conduction there is at room temperature is due to impurities. Heating will help, as it increases the thermal energy available for both defect creation and ion mobility. As seen from the Maxwell-Boltzmann distribution discussed earlier, thermal effects increase exponentially with temperature. Still, even at the melting point of salt its conductivity is eight orders of magnitude less than that of metals.

There are however ionic materials that have much higher conductivities. They cannot compete with metals, but some ionic solids can compete with liquid electrolytes. These solids may be referred to as "solid electrolytes, "fast ion conductors," or "superionic conductors." They are important for such applications as batteries, fuel cells, and gas sensors. Yttria-stabilized zirconia is an example, although unfortunately only at temperatures around 1 000 °C. In the best ionic conductors, the crystal structure for one kind of ion becomes so irregular that these ions are effectively in a molten state. For example, this happens for the silver ions in the classical example of hot silver iodide. Throw in 25% of rubidium chloride and $RbAg_4Cl_5$ stays superionic to room temperature.

Crystal surfaces are also crystal defects, in a sense. They can enhance ionic conductivity. For example, nanoionics can greatly improve the ionic conductivity of poor ionic conductors by combining them in nanoscale layers.

Key Points

- In ionic solids, some electrical conduction may occur through the motion of the ions instead of individual electrons.
- It is important for applications such as batteries, fuel cells, and gas sensors.

6.22 Electrons in Crystals

A meaningful discussion of semiconductors requires some background on how electrons move through solids. The free-electron gas model simply assumes that the electrons move through an empty periodic box. But of course, to describe a real solid the box should really be filled with the countless atoms around which the conduction electrons move.

This subsection will explain how the motion of electrons gets modified by the atoms. To keep things simple, it will still be assumed that there is no direct interaction between the electrons. It will also be assumed that the solid is crystalline, which means that the atoms are arranged in a periodic pattern. The atomic period should be assumed to be many orders of magnitude shorter than the size of the periodic box. There must be many atoms in each direction in the box.

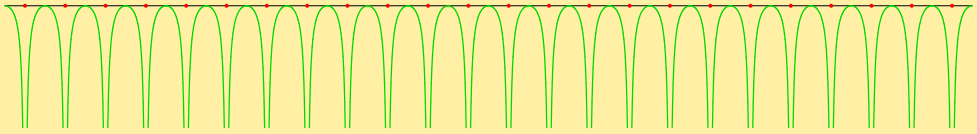


Figure 6.21: Potential energy seen by an electron along a line of nuclei. The potential energy is in green, the nuclei are in red.

The effect of the crystal is to introduce a periodic potential energy for the electrons. For example, figure 6.21 gives a sketch of the potential energy seen by an electron along a line of nuclei. Whenever the electron is right on top of a nucleus, its potential energy plunges. Close enough to a nucleus, a very strong attractive Coulomb potential is seen. Of course, on a line that does not pass exactly through nuclei, the potential will not plunge that low.

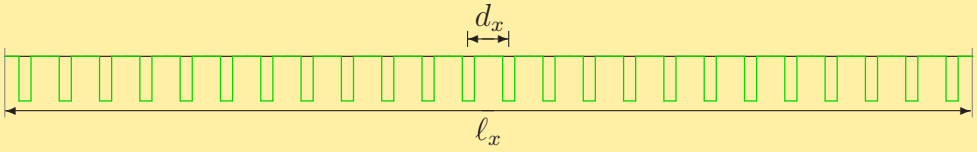


Figure 6.22: Potential energy seen by an electron in the one-dimensional simplified model of Kronig & Penney.

Kronig & Penney developed a very simple one-dimensional model that explains much of the motion of electrons through crystals. It assumes that the potential energy seen by the electrons is periodic on some atomic-scale period d_x. It also assumes that this potential consists of square dips, like in figure 6.22. You might think of the regions of lowered potential energy as the immediate vicinity of the nuclei. This is the model that will be examined. The atomic period d_x is assumed to be much smaller than the periodic box size ℓ_x, i.e. the

size of the complete "crystal." In particular, the box should contain a large and whole number of atomic periods.

Three-dimensional Kronig & Penney quantum states can be formed as products of one-dimensional ones, compare chapter 3.5.8. However, such states are limited to potentials that are sums of one-dimensional ones. In any case, this section will restrict itself mostly to the one-dimensional case.

6.22.1 Bloch waves

This subsection examines the single-particle quantum states, or energy eigenfunctions, of electrons in one-dimensional solids.

For free electrons, the energy eigenfunctions were given in section 6.18. In one dimension they are:

$$\psi^{\rm p}_{n_x}(x) = Ce^{{\rm i}k_x x}$$

where integer n_x merely numbers the eigenfunctions and C is a normalization constant that is not really important. What is important is that these eigenfunctions do not just have definite energy $E^{\rm p}_x = \hbar^2 k_x^2/2m_{\rm e}$, they also have definite linear momentum $p_x = \hbar k_x$. Here m_e is the electron mass and $\hbar$ the reduced Planck constant. In classical terms, the electron velocity is given by the linear momentum as $v^{\rm p}_x = p_x/m_{\rm e}$.

To find the equivalent one-dimensional energy eigenfunctions $\psi^{\rm p}_{n_x}(x)$ in the presence of a crystal potential $V_x(x)$ is messy. It requires solution of the one-dimensional Hamiltonian eigenvalue problem

$$-\frac{\hbar^2}{2m_{\rm e}}\frac{\partial^2\psi^{\rm p}}{\partial x^2} + V_x\psi^{\rm p} = E^{\rm p}_x\psi^{\rm p}$$

where $E^{\rm p}_x$ is the energy of the state. The solution is best done on a computer, even for a potential as simple as the Kronig & Penney one, {N.9}.

However, it can be shown that the eigenfunctions can always be written in the form:

$$\boxed{\psi^{\rm p}_{n_x}(x) = \psi^{\rm p}_{{\rm p},n_x}(x)e^{{\rm i}k_x x}} \tag{6.32}$$

in which $\psi^{\rm p}_{{\rm p},n_x}(x)$ is an *periodic* function on the atomic period. Note that as long as $\psi^{\rm p}_{{\rm p},n_x}(x)$ is a simple constant, this is exactly the same as the eigenfunctions of the free-electron gas in one dimension; mere exponentials. But if the periodic potential $V_x(x)$ is not a constant, then neither is $\psi^{\rm p}_{{\rm p},n_x}(x)$. In that case, all that can be said a priori is that it is periodic on the atomic period.

Energy eigenfunctions of the form (6.32) are called "Bloch waves." It may be pointed out that this form of the energy eigenfunctions was discovered by Floquet, not Bloch. However, Floquet was a mathematician. In naming the solutions after Bloch instead of Floquet, physicists celebrate the physicist who could do it too, just half a century later.

The reason why the energy eigenfunctions take this form, and what it means for the electron motion are discussed further in chapter 7.10.5. There are only two key points of interest for now. First, the possible values of the wave number k_x are exactly the same as for the free-electron gas, given in (6.28). Otherwise the eigenfunction would not be periodic on the period of the box. Second, the electron velocity can be found by differentiating the single particle energy $E_x^{\rm p}$ with respect to the "crystal momentum" $p_{{\rm cm},x} = \hbar k_x$. That is the same as for the free-electron gas. If you differentiate the one-dimensional free-electron gas kinetic energy $E_x^{\rm p} = (\hbar k_x)^2/2m_{\rm e}$ with respect to $p_x = \hbar k_x$, you get the velocity.

Key Points

- In the presence of a periodic crystal potential, the energy eigenfunctions pick up an additional factor that has the atomic period.
- The wave number values do not change.
- The velocity is found by differentiating the energy with respect to the crystal momentum.

6.22.2 Example spectra

As the previous section discussed, the difference between metals and insulators is due to differences in their energy spectra. The one-dimensional Kronig & Penney model can provide some insight into it.

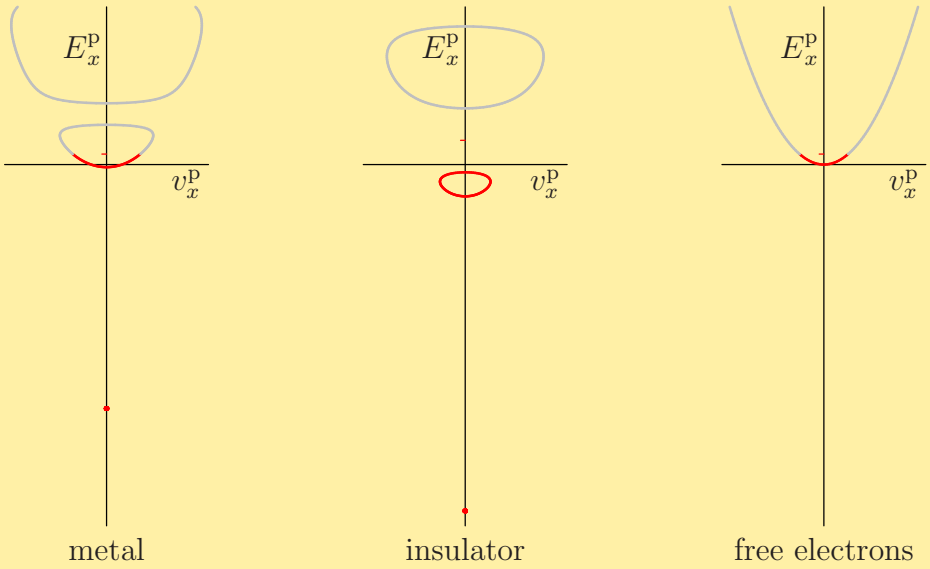


Figure 6.23: Example Kronig & Penney spectra.

Finding the energy eigenvalues is not difficult on a computer, {N.9}. A couple of example spectra are shown in figure 6.23. The vertical coordinate is

the single-electron energy, as usual. The horizontal coordinate is the electron velocity. (So the free electron example is the one-dimensional version of the spectrum in figure 6.18, but the axes are much more compressed here.) Quantum states occupied by electrons are again in red.

The example to the left in figure 6.23 tries to roughly model a metal like lithium. The depth of the potential drops in figure 6.22 was chosen so that for lone "atoms," (i.e. for widely spaced potential drops), there is one bound spatial state and a second marginally bound state. You might think of the bound state as holding lithium's two inner "1s" electrons, and the marginally bound state as holding its loosely bound single "2s" valence electron.

Note that the 1s state is just a red dot in the lower part of the left spectrum in figure 6.23. The energy of the inner electrons is not visibly affected by the neighboring "atoms." Also, the velocity does not budge from zero; electrons in the inner states would hardly move even *if* there were unfilled states. These two observations are related, because as mentioned earlier, the velocity is the derivative of the energy with respect to the crystal momentum. If the energy does not vary, the velocity is zero.

The second energy level has broadened into a half-filled "conduction band." Like for the free-electron gas in figure 6.18, it requires little energy to move some Fermi-level electrons in this band from negative to positive velocities to achieve net electrical conduction.

The spectrum in the middle of figure 6.23 tries to roughly model an insulator like diamond. (The one-dimensional model is too simple to model an alkaline metal with two valence electrons like beryllium. The spectra of these metals involve different energy bands that merge together, and merging bands do not occur in the one-dimensional model.) The voltage drops have been increased a bit to make the second energy level for lone "atoms" more solidly bound. And it has been assumed that there are now four electrons per "atom," so that the second band is completely filled.

Now the only way to achieve net electrical conduction is to move some electrons from the filled "valence band" to the empty "conduction band" above it. That requires much more energy than a normal applied voltage could provide. So the crystal is an insulator.

The reasons why the spectra look as shown in figure 6.23 are not obvious. Note {N.9} explains by example what happens to the free-electron gas energy eigenfunctions when there is a crystal potential. A much shorter explanation that hits the nail squarely on the head is "That is just the way the Schrödinger equation is."

Key Points

- A periodic crystal potential produces energy bands.

6.22.3 Effective mass

The spectrum to the right in figure 6.23 shows the one-dimensional free-electron gas. The relationship between velocity and energy is given by the classical expression for the kinetic energy in the x-direction:

$$E_x^{\rm p} = \tfrac{1}{2} m_{\rm e} v_x^{\rm p2}$$

This leads to the parabolic spectrum shown.

It is interesting to compare this spectrum to that of the "metal" to the left in figure 6.23. The occupied part of the conduction band of the metal is approximately parabolic just like the free-electron gas spectrum. To a fair approximation, in the occupied part of the conduction band

$$E_x^{\rm p} - E_{{\rm c},x}^{\rm p} = \tfrac{1}{2} m_{{\rm eff},x} v_x^{\rm p2}$$

where $E_{{\rm c},x}^{\rm p}$ is the energy at the bottom of the conduction band and $m_{{\rm eff},x}$ is a constant called the "effective mass."

This illustrates that conduction band electrons in metals behave much like free electrons. And the similarity to free electrons becomes even stronger if you define the zero level of energy to be at the bottom of the conduction band and replace the true electron mass by an effective mass. For the metal shown in figure 6.23, the effective mass is 61% of the true electron mass. That makes the parabola somewhat flatter than for the free-electron gas. For electrons that reach the conduction band of the insulator in figure 6.23, the effective mass is only 18% of the true mass.

In previous sections, the valence electrons in metals were repeatedly approximated as free electrons to derive such properties as degeneracy pressure and thermionic emission. The justification was given that the forces on the valence electrons tend to come from all directions and average out. But as the example above now shows, that approximation can be improved upon by replacing the true electron mass by an effective mass. For the valence electrons in copper, the appropriate effective mass is about one and a half times the true electron mass, [42, p. 257]. So the use of the true electron mass in the examples was not dramatically wrong.

And the agreement between conduction band electrons and free electrons is even deeper than the similarity of the spectra indicates. You can also use the density of states for the free-electron gas, as given in section 6.3, if you substitute in the effective mass.

To see why, assume that the relationship between the energy $E_x^{\rm p}$ and the velocity $v_x^{\rm p}$ is the same as that for a free-electron gas whose electrons have the appropriate effective mass. Then so is the relationship between the energy $E_x^{\rm p}$ and the wave number k_x the same as for that electron gas. That is because the velocity is merely the derivative of $E_x^{\rm p}$ with respect to $\hbar k_x$. You need the

same $E_x^{\rm p}$ versus k_x relation to get the same velocity. (This assumes that you measure both the energy and the wave number from the location of minimum conduction band energy.) And if the $E_x^{\rm p}$ versus k_x relation is the same as for the free-electron gas, then so is the density of states. That is because the quantum states have the same wave number spacing regardless of the crystal potential.

It should however be pointed out that in three dimensions, things get messier. Often the effective masses are different in different crystal directions. In that case you need to define some suitable average to use the free-electron gas density of states. In addition, for typical semiconductors the energy structure of the holes at the top of the valence band is highly complex.

Key Points

- The electrons in a conduction band and the holes in a valence band are often modeled as free particles.
- The errors can be reduced by giving them an effective mass that is different from the true electron mass.
- The density of states of the free-electron gas can also be used.

6.22.4 Crystal momentum

The crystal momentum of electrons in a solid is not the same as the linear momentum of free electrons. However, it is similarly important. It is related to optical properties such as the difference between direct and indirect gap semiconductors. Because of this importance, spectra are usually plotted against the crystal momentum, rather than against the electron velocity. The Kronig & Penney model provides a simple example to explain some of the ideas.

Figure 6.24 shows the single-electron energy plotted against the crystal momentum. Note that this is equivalent to a plot against the wave number k_x; the crystal momentum is just a simple multiple of the wave number, $p_{{\rm cm},x} = \hbar k_x$. The figure has nondimensionalized the wave number by multiplying it by the atomic period d_x. Both the example insulator and the free-electron gas are shown in the figure.

There is however an ambiguity in the figure:

> *The crystal wave number, and so the crystal momentum, is not unique.*

Consider once more the general form of a Bloch wave,

$$\psi_{n_x}^{\rm p}(x) = \psi_{{\rm p},n_x}^{\rm p}(x) e^{{\rm i}k_x x}$$

If you change the value of k_x by a whole multiple of $2\pi/d_x$, it remains a Bloch wave in terms of the new k_x. The change in the exponential can be absorbed in

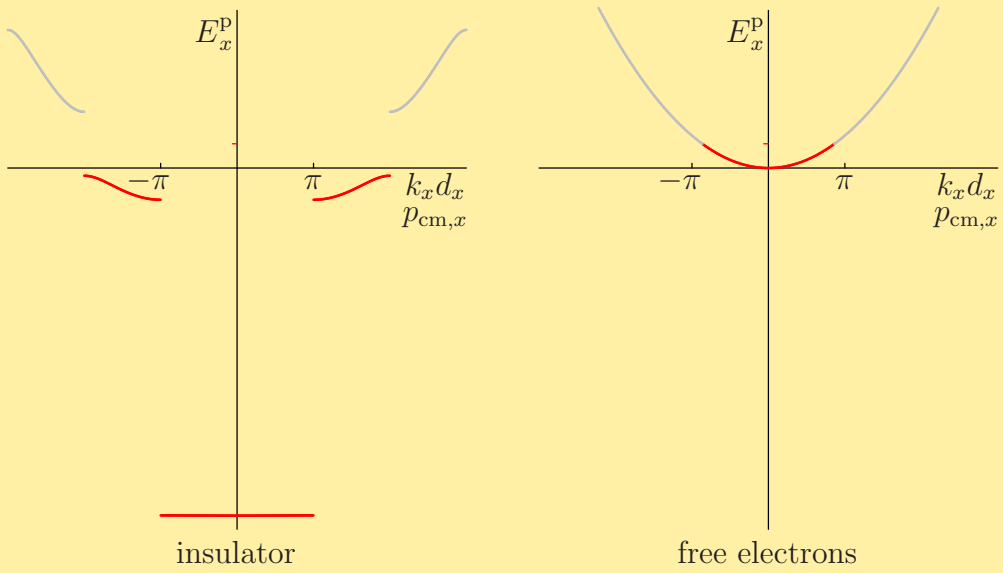


Figure 6.24: Spectrum against wave number in the extended zone scheme.

the periodic part $\psi^{\rm p}_{{\rm p},n_x}$. The periodic part changes, but it remains periodic on the atomic scale d_x.

Therefore there is a problem with how to define a unique value of k_x. There are different solutions to this problem. Figure 6.24 follows the so-called "extended zone scheme." It takes the wave number to be zero at the minimum energy and then keeps increasing the magnitude with energy. This is a good scheme for the free-electron gas. It also works nicely if the potential is so weak that the energy states are almost the free-electron gas ones.

A second approach is much more common, though. It uses the indeterminacy in k_x to shift it into the range $-\pi \leqslant k_x d_x \leqslant \pi$. That range is called the "first Brillouin zone." Restricting the wave numbers to the first Brillouin zone produces figure 6.25. This is called the "reduced zone scheme." Esthetically, it is clearly an improvement in case of a nontrivial crystal potential.

But it is much more than that. For one, the different energy curves in the reduced zone scheme can be thought of as modified atomic energy levels of lone atoms. The corresponding Bloch waves can be thought of as modified atomic states, modulated by a relatively slowly varying exponential $e^{{\rm i}k_x x}$.

Second, the reduced zone scheme is important for optical applications of semiconductors. In particular,

> *A lone photon can only produce an electron transition along the same vertical line in the reduced zone spectrum.*

The reason is that crystal momentum must be conserved. That is much like linear momentum must be preserved for electrons in free space. Since a photon has negligible crystal momentum, the crystal momentum of the electron cannot

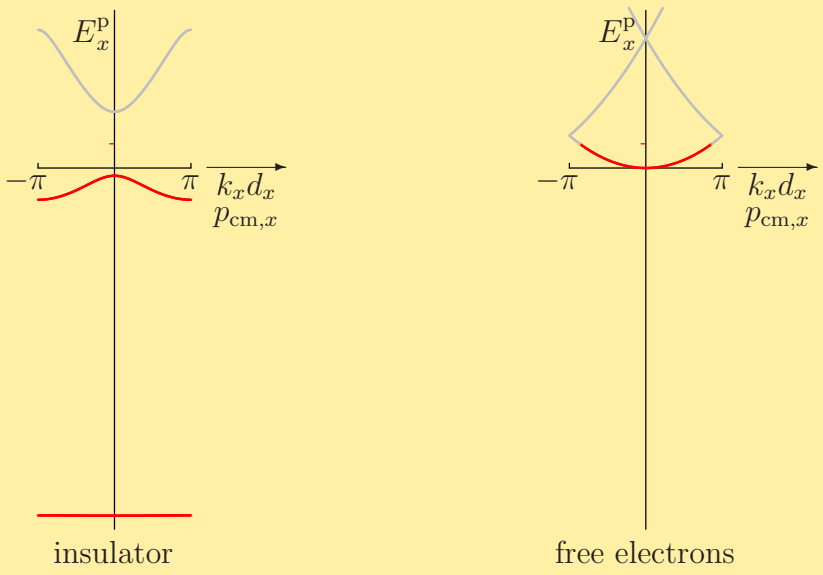


Figure 6.25: Spectrum against wave number in the reduced zone scheme.

change. That means it must stay on the same vertical line in the reduced zone scheme.

To see why that is important, suppose that you want to use a semiconductor to create light. To achieve that, you need to somehow excite electrons from the valence band to the conduction band. How to do that will be discussed in section 6.27.7. The question here is what happens next. The excited electrons will eventually drop back into the valence band. If all is well, they will emit the energy they lose in doing so as a photon. Then the semiconductor will emit light.

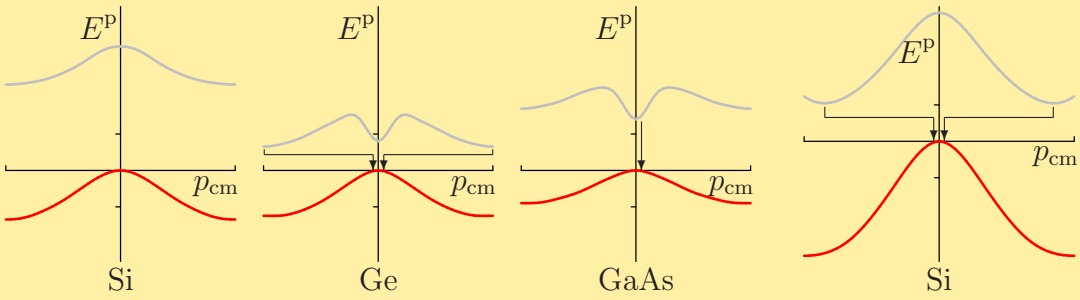


Figure 6.26: Some one-dimensional energy bands for a few basic semiconductors.

It turns out that the excited electrons are mostly in the lowest energy states in the conduction band. For various reasons. that tends to be true despite the absence of thermal equilibrium. They are created there or evolve to it. Also, the holes that the excited electrons leave behind in the valence band are mostly at the highest energy levels in that band.

Now consider the energy bands of some actual semiconductors shown to the left in figure 6.26. In particular, consider the spectrum of gallium arsenide. The excited electrons are at the lowest point of the conduction band. That is at zero crystal momentum. The holes are at the highest point in the valence band, which is also at zero crystal momentum. Therefore, the excited electrons can drop vertically down into the holes. The crystal momentum does not change, it stays zero. There is no problem. In fact, the first patent for a light emitting diode was for a gallium arsenide one, in 1961. The energy of the emitted photons is given by the band gap of gallium arsenide, somewhat less than 1.5 eV. That is slightly below the visible range, in the near infrared. It is suitable for remote controls and other nonvisible applications.

But now consider germanium in figure 6.26. The highest point of the valence band is still at zero crystal momentum. But the lowest point of the conduction band is now at maximum crystal momentum in the reduced zone scheme. When the excited electrons drop back into the holes, their crystal momentum changes. Since crystal momentum is conserved, something else must account for the difference. And the photon does not have any crystal momentum to speak of. It is a phonon of crystal vibration that must carry off the difference in crystal momentum. Or supply the difference, if there are enough pre-existing thermal phonons. The required involvement of a phonon in addition to the photon makes the entire process much more cumbersome. Therefore the energy of the electron is much more likely to be released through some alternate mechanism that produces heat instead of light.

The situation for silicon is like that for germanium. However, the lowest energy in the conduction band occurs for a different direction of the crystal momentum. The spectrum for that direction of the crystal momentum is shown to the right in figure 6.26. It still requires a change in crystal momentum.

At the time of writing, there is a lot of interest in improving the light emission of silicon. The reason is its prevalence in semiconductor applications. If silicon itself can be made to emit light efficiently, there is no need for the complications of involving different materials to do it. One trick is to minimize processes that allow electrons to drop back into the valence band without emitting photons. Another is to use surface modification techniques that promote absorption of photons in solar cell applications. The underlying idea is that at least in thermal equilibrium, the best absorbers of electromagnetic radiation are also the best emitters, section 6.8.

Gallium arsenide is called a "direct-gap semiconductor" because the electrons can fall straight down into the holes. Silicon and germanium are called "indirect-gap semiconductors" because the electrons must change crystal momentum. Note that these terms are accurate and understandable, a rarity in physics.

Conservation of crystal momentum does not just affect the emission of light. It also affects its absorption. Indirect-gap semiconductors do not absorb photons

very well if the photons have little more energy than the band gap. They absorb photons with enough energy to induce vertical electron transitions a lot better.

It may be noted that conservation of crystal momentum is often called "conservation of wave vector." It is the same thing of course, since the crystal momentum is simply $\hbar$ times the wave vector. However, those pesky new students often have a fairly good understanding of momentum conservation, and the term momentum would leave them insufficiently impressed with the brilliance of the physicist using it.

(If you wonder why crystal momentum is preserved, and how it even can be if the crystal momentum is not unique, the answer is in the discussion of conservation laws in chapter 7.3 and its note. It is not really momentum that is conserved, but the product of the single-particle eigenvalues $e^{\mathrm{i}k_x d_x}$ of the operator that translates the system involved over a distance d_x. These eigenvalues do not change if the wave numbers change by a whole multiple of $2\pi/d_x$, so there is no violation of the conservation law if they do. For a system of particles in free space, the potential is trivial; then you can take d_x equal to zero to eliminate the ambiguity in k_x and so in the momentum. But for a nontrivial crystal potential, d_x is fixed. Also, since a photon moves so fast, its wave number is almost zero on the atomic scale, giving it negligible crystal momentum. At least it does for the photons in the eV range that are relevant here.)

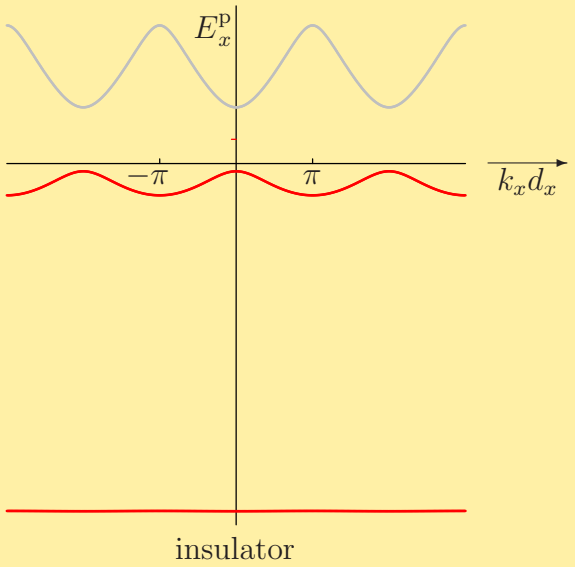


Figure 6.27: Spectrum against wave number in the periodic zone scheme.

Returning to the possible ways to plot spectra, the so-called "periodic zone scheme" takes the reduced zone scheme and extends it periodically, as in figure 6.27. That makes for very esthetic pictures, especially in three dimensions.

Of course, in three dimensions there is no reason for the spectra in the y and z directions to be the same as the one in the x-direction. Each can in principle

be completely different from the other two. Regardless of the differences, valid three-dimensional Kronig & Penney energy eigenfunctions are obtained as the product of the x, y and z eigenfunctions, and their energy is the sum of the eigenvalues.

Similarly, typical spectra for real solids have to show the spectrum versus wave number for more than one crystal direction to be comprehensive. One example was for silicon in figure 6.26. A more complete description of the one-dimensional spectra of real semiconductors is given in the next subsection.

Key Points

- The wave number and crystal momentum values are not unique.
- The extended, reduced, and periodic zone schemes make different choices for which values to use.
- The reduced zone scheme limits the wave numbers to the first Brillouin zone.
- For a photon to change the crystal momentum of an electron in the reduced zone scheme requires the involvement of a phonon.
- That makes indirect gap semiconductors like silicon and germanium undesirable for some optical applications.

6.22.5 Three-dimensional crystals

A complete description of the theory of three-dimensional crystals is beyond the scope of the current discussion. Chapter 10 provides a first introduction. However, because of the importance of semiconductors such as silicon, germanium, and gallium arsenide, it may be a good idea to explain a few ideas already.

Consider first a gallium arsenide crystal. Gallium arsenide has the same crystal structure as zinc sulfide, in the form known as zinc blende or sphalerite. The crystal is sketched in figure 6.28. The larger spheres represent the nuclei and inner electrons of the gallium atoms. The smaller spheres represent the nuclei and inner electrons of the arsenic atoms. Because arsenic has a more positively charged nucleus, it holds its electrons more tightly. The figure exaggerates the effect to keep the atoms visually apart.

The grey gas between these atom cores represents the valence electrons. Each gallium atom contributes 3 valence electrons and each arsenic atom contributes 5. That makes an average of 4 valence electrons per atom.

As the figure shows, each gallium atom core is surrounded by 4 arsenic ones and vice-versa. The grey sticks indicate the directions of the covalent bonds between these atom cores. You can think of these bonds as somewhat polar sp^3 hybrids. They are polar since the arsenic atom is more electronegative than the gallium one.

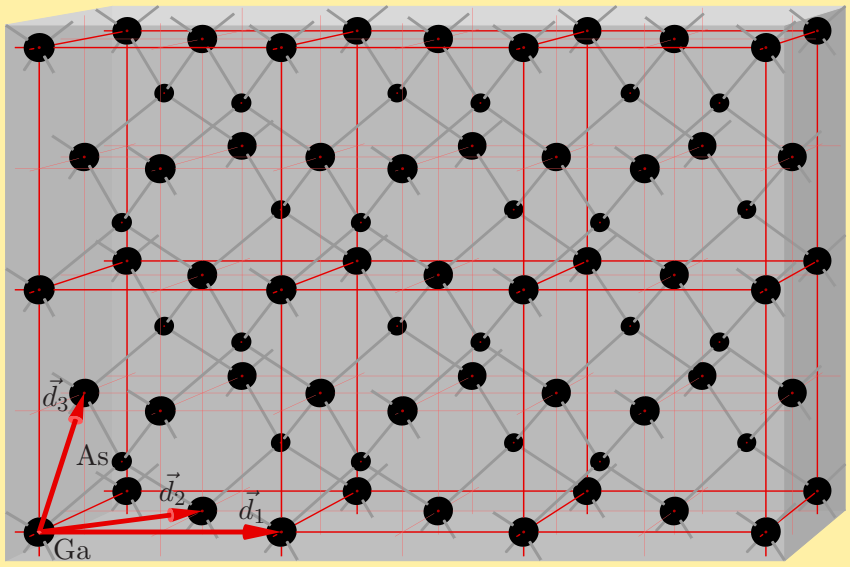


Figure 6.28: Schematic of the zinc blende (ZnS) crystal relevant to important semiconductors including silicon.

It is customary to think of crystals as being build up out of simple building blocks called "unit cells." The conventional unit cells for the zinc blende crystal are the little cubes outlined by the thicker red lines in figure 6.28. Note in particular that you can find gallium atoms at each corner of these little cubes, as well as in the center of each face of them. That makes zinc blende an example of what is called a "face-centered cubic" lattice. For obvious reasons, everybody abbreviates that to FCC.

You can think of the unit cells as subdivided further into 8 half-size cubes, as indicated by the thinner red lines. There is an arsenic atom in the center of every other of these smaller cubes.

The simple one-dimensional Kronig & Penney model assumed that the crystal was periodic with a period d_x. For real three-dimensional crystals, there is not just one period, but three. More precisely, there are three so-called "primitive translation vectors" $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$. A set of primitive translation vectors for the FCC crystal is shown in figure 6.28. If you move around by *whole* multiples of these vectors, you arrive at points that look identical to your starting point.

For example, if you start at the center of a gallium atom, you will again be at the center of a gallium atom. And you can step to whatever gallium atom you like in this way. At least as long as the whole multiples are allowed to be both positive and negative. In particular, suppose you start at the gallium atom with the Ga label in figure 6.28. Then $\vec{d}_1$ allows you to step to any other gallium atom on the same line going towards the right and left. Vector $\vec{d}_2$ allows you to step to the next or previous line in the same horizontal plane. And vector $\vec{d}_3$ allows you to step to the next higher or lower horizontal plane.

The choice of primitive translation vectors is not unique. In particular, many sources prefer to draw the vector $\vec{d}_1$ towards the gallium atom in the front face center rather than to the one at the right. That is more symmetric, but moving around with them gets harder to visualize. Then you would have to step over $\vec{d}_1$, $\vec{d}_2$, and $-\vec{d}_3$ just to reach the atom to the right.

You can use the primitive translation vectors also to mentally *create* the zinc blende crystal. Consider the pair of atoms with the Ga and As labels in figure 6.28. Suppose that you put a copy of this pair at every point that you can reach by stepping around with the primitive translation vectors. Then you get the complete zinc blende crystal. The pair of atoms is therefore called a "basis" of the zinc blende crystal.

This also illustrates another point. The choice of unit cell for a given crystal structure is not unique. In particular, the parallelepiped with the primitive translation vectors as sides can be used as an alternative unit cell. Such a unit cell has the smallest possible volume, and is called a primitive cell.

The crystal structure of silicon and germanium, as well as diamond, is identical to the zinc blende structure, but all atoms are of the same type. This crystal structure is appropriately called the diamond structure. The basis is still a two-atom pair, even if the two atoms are now the same. Interestingly enough, it is not possible to create the diamond crystal by distributing copies of a single atom. Not as long as you step around with only three primitive translation vectors.

For the one-dimensional Kronig & Penney model, there was only a single wave number k_x that characterized the quantum states. For a three-dimensional crystal, there is a three-dimensional wave number vector $\vec{k}$ with components k_x, k_y, and k_z. That is just like for the free-electron gas in three dimensions as discussed in earlier sections.

In the Kronig & Penney model, the wave numbers could be reduced to a finite interval

$$-\frac{\pi}{d_x} \leqslant k_x < \frac{\pi}{d_x}$$

This interval was called the first Brillouin zone. Wave numbers outside this zone are equivalent to ones inside. The general rule was that wave numbers a whole multiple of $2\pi/d_x$ apart are equivalent.

In three dimensions, the first Brillouin zone is no longer a one-dimensional interval but a three-dimensional volume. And the separations over which wave number vectors are equivalent are no longer so simple. Instead of simply taking an inverse of the period d_x, as in $2\pi/d_x$, you have to take an inverse of the matrix formed by the three primitive translation vectors $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$. Next you have to identify the wave number vectors closest to the origin that are enough to describe all quantum states. If you do all that for the FCC crystal, you will end up with the first Brillouin zone shown in figure 6.29. It is shaped like a cube with its 8 corners cut off.

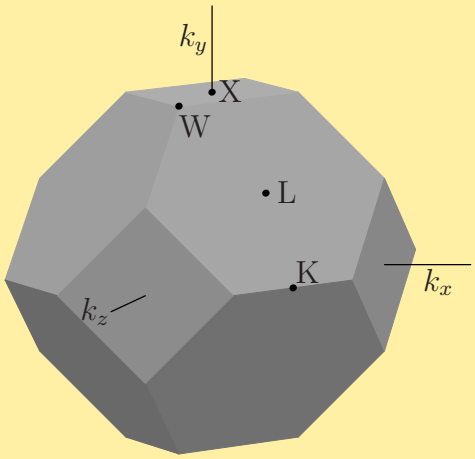


Figure 6.29: First Brillouin zone of the FCC crystal.

The shape of the first Brillouin zone is important for understanding graphs of three-dimensional spectra. Every single point in the first Brillouin zone corresponds to multiple Bloch waves, each with its own energy. To plot all those energies is not possible; it would require a four-dimensional plot. Instead, what is done is plot the energies along representative lines. Such plots will here be indicated as one-dimensional energy bands. Note however that they are one-dimensional bands of true three-dimensional crystals. They are not just Kronig & Penney model bands.

Typical points between which one-dimensional bands are drawn are indicated in figure 6.29. You and I would probably name such points something like F (face), E (edge), and C (corner), with a clarifying subscript as needed. However, physicists come up with names like K, L, W, and X, and declare them standard. The center of the Brillouin zone is the origin, where the wave number vector is zero. Normal people would therefore indicate it as O or 0. However, physicists are not normal people. They indicate the origin by Γ because the shape of this Greek letter reminds them of a gallows. Physicists just love gallows humor.

Computed one-dimensional energy bands between the various points in the Brillouin zone can be found in the plot to the left in figure 6.30. The plot is for germanium. The zero level of energy was chosen as the top of the valence band. The various features of the plot agree well with other experimental and computational data.

The earlier spectrum for germanium in figure 6.26 showed only the part within the little frame in figure 6.30. That part is for the line between zero wave number and the point L in figure 6.29. Unlike figure 6.30, the earlier spectrum figure 6.26 showed both negative and positive wave numbers, as its left and right halves. On the other hand, the earlier spectrum showed only the highest one-dimensional valence band and the lowest one-dimensional conduction band. It was sufficient to show the top of the valence band and the bottom of the

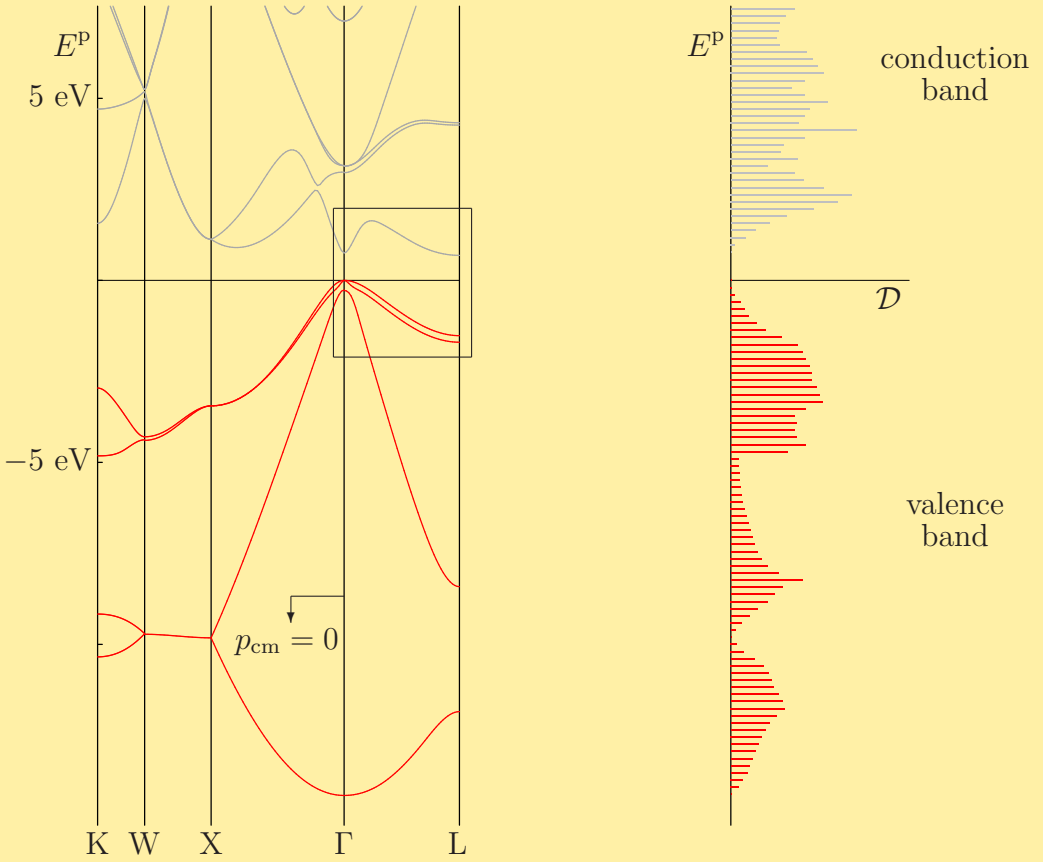


Figure 6.30: Sketch of a more complete spectrum of germanium. (Based on results of the VASP 5.2 commercial computer code.)

conduction band, but little else. As figure 6.30 shows, there are actually four different types of Bloch waves in the valence band. The energy range of each of the four is within the range of the combined valence band.

The complete valence band, as well as the lower part of the conduction band, is sketched in the spectrum to the right in figure 6.30. It shows the energy plotted against the density of states $\mathcal{D}$. Note that the computed density of states for the conduction electrons is a mess when seen over its complete range. It is nowhere near parabolic as it would be for electrons in empty space, figure 6.1. Similarly the density of states applicable to the valence band holes is nowhere near an inverted parabola over its complete range. However, typically only about 1/40th of an eV below the top of the valence band and above the bottom of the conduction band is relevant for applications. That is very small on the scale of the figure.

An interesting feature of figure 6.30 is that two different energy bands merge at the top of the valence band. These two bands have the same energy at the top of the valence band, but very different curvature. And according to the earlier

subsection 6.22.3, that means that they have different effective mass. Physicists therefore speak of "light holes" and "heavy holes" to keep the two types of quantum states apart. Typically even the heavy holes have effective masses less than the true electron mass, [29, pp. 214-216]. Diamond is an exception.

The spectrum of silicon is not that different from germanium. However, the bottom of the conduction band is now on the line from the origin Γ to the point X in figure 6.29.

Key Points

- Silicon and germanium have the same crystal structure as diamond. Gallium arsenide has a generalized version, called the zinc blende structure.
- The spectra of true three-dimensional crystals are considerably more complex than those of the one-dimensional Kronig & Penney model.
- In three dimensions, the period turns into three primitive translation vectors.
- The first Brillouin zone becomes three-dimensional.
- There are light holes and heavy holes at the top of the valence band of typical semiconductors.

6.23 Semiconductors

Semiconductors are at the core of modern technology. This section discusses some basic properties of semiconductors that will be needed to explain how the various semiconductor applications work. The main semiconductor manipulation that must be described in this section is "doping," adding a small amount of impurity atoms.

If semiconductors did not conduct electricity, they would not be very useful. Consider first the pure, or "intrinsic," semiconductor. The vicinity of the band gap in its spectrum is shown to the left in figure 6.31. The vertical coordinate shows the energy E^{p} of the single-electron quantum states. The horizontal coordinate shows the density of states $\mathcal{D}$, the number of quantum states per unit energy range. Recall that there are no quantum states in the band gap. States occupied by electrons are shown in red. At room temperature there are some thermally excited electrons in the conduction band. They left behind some holes in the valence band. Both the electrons and the holes can provide electrical conduction.

Time for a reality check. The number of such electrons and holes is very much smaller than the figure indicates. The number ι_{e} of electrons per quantum state is given by the Fermi-Dirac distribution (6.19). In the conduction band, that

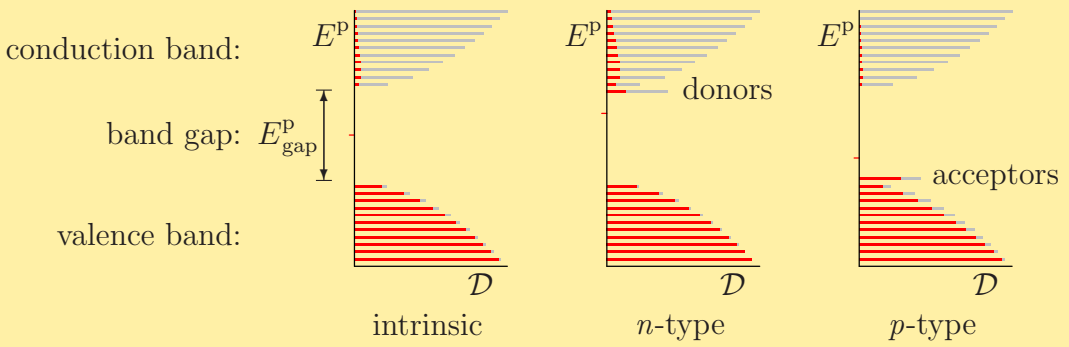


Figure 6.31: Vicinity of the band gap in the spectra of intrinsic and doped semiconductors. The amounts of conduction band electrons and valence band holes have been vastly exaggerated to make them visible.

may be simplified to the Maxwell-Boltzmann one (6.21) because the number of electrons in the conduction band is small. The average number of electrons per state in the conduction band is then:

$$\boxed{\iota_{\mathrm{e}} = e^{-(E^{\mathrm{p}}-\mu)/k_{\mathrm{B}}T}} \tag{6.33}$$

Here T is the absolute temperature, k_{B} is the Boltzmann constant, and μ is the chemical potential, also known as the Fermi level. The Fermi level is shown by a red tick mark in figure 6.31.

For an intrinsic semiconductor, the Fermi level is about in the middle of the band gap. Therefore the average number of electrons per quantum state at the bottom of the conduction band is

$$\text{Bottom of the conduction band:} \quad \iota_{\mathrm{e}} = e^{-E^{\mathrm{p}}_{\mathrm{gap}}/2k_{\mathrm{B}}T}$$

At room temperature, $k_{\mathrm{B}}T$ is about 0.025 eV while for silicon, the band gap energy is about 1.12 eV. That makes ι_{e} about $2\ 10^{-10}$. In other words, only about 1 in 5 billion quantum states in the lower part of the conduction band has an electron in it. And it is even less higher up in the band. A figure cannot show a fraction that small; there are just not enough atoms on a page.

So it is not surprising that pure silicon conducts electricity poorly. It has a resistivity of several thousand ohm-m where good metals have on the order of 10^{-8}. Pure germanium, with a smaller band gap of 0.66 eV, has a much larger ι_{e} of about $3\ 10^{-6}$ at the bottom of the conduction band. Its resistivity is correspondingly lower at about half an ohm-m. That is still many orders of magnitude larger than for a metal.

And the number of conduction electrons becomes much smaller still at cryogenic temperatures. If the temperature is a frigid 150 K instead of a 300 K room temperature, the number of electrons per state in silicon drops by another factor of a billion. That illustrates one important rule:

You cannot just forget about temperature to understand semiconductors.

Usually, you like to analyze the ground state at absolute zero temperature of your system, because it is easier. But that simply does not work for semiconductors.

The number of holes per state in the valence band may be written in a form similar to that for the electrons in the conduction band:

$$\boxed{\iota_{\rm h} = e^{-(\mu - E^{\rm p})/k_{\rm B}T}} \tag{6.34}$$

Note that in the valence band the energy is less than the Fermi level μ, so that the exponential is again very small. The expression above may be checked by noting that whatever states are not filled with electrons are holes, so $\iota_{\rm h} = 1 - \iota_{\rm e}$. If you plug the Fermi-Dirac distribution into that, you get the expression for $\iota_{\rm h}$ above as long as the number of holes per state is small.

From a comparison of the expressions for the number of particles per state $\iota_{\rm e}$ and $\iota_{\rm h}$ it may already be understood why the Fermi level μ is approximately in the middle of the band gap. If the Fermi level is exactly in the middle of the band gap, $\iota_{\rm e}$ at the bottom of the conduction band is the same as $\iota_{\rm h}$ at the top of the valence band. Then there is the same number of electrons per state at the bottom of the conduction band as holes per state at the top of the valence band. That is about as it should be, since the total number of electrons in the conduction band must equal the total number of holes in the valence band. The holes in the valence band is where the electrons in the conduction band came from.

Note that figure 6.31 is misleading in the sense that it depicts the same density of states $\mathcal{D}$ in the conduction band as in the valence band. In reality, the number of states per unit energy range in the conduction band could easily be twice that at the corresponding location in the valence band. It seems that this should invalidate the above argument that the Fermi level μ must be in the middle of the band gap. But it does not. To change the ratio between $\iota_{\rm e}$ and $\iota_{\rm h}$ by a factor 2 requires a shift in μ of about 0.01 eV at room temperature. That is very small compared to the band gap. And the shift would be much smaller still closer to absolute zero temperature. At absolute zero temperature, the Fermi level must move to the exact middle of the gap.

That illustrates another important rule of thumb for semiconductors:

Keep your eyes on the thermal exponentials. Usually, their variations dwarf everything else.

If $E^{\rm p}$ or μ changes just a little bit, $e^{-(E^{\rm p}-\mu)/k_{\rm B}T}$ changes dramatically.

(For gallium arsenide, the difference between the densities of states for holes and electrons is much larger than for silicon or germanium. That makes the shift in Fermi level at room temperature more substantial.)

The Fermi level may be directly computed. Expressions for the total number of conduction electrons per unit volume and the total number of holes per unit volume are, {D.30}:

$$\boxed{i_{\rm e} = 2\left(\frac{m_{\rm eff,e} k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} e^{-(E_{\rm c}^{\rm p}-\mu)/k_{\rm B}T} \qquad i_{\rm h} = 2\left(\frac{m_{\rm eff,h} k_{\rm B} T}{2\pi\hbar^2}\right)^{3/2} e^{-(\mu-E_{\rm v}^{\rm p})/k_{\rm B}T}} \tag{6.35}$$

Here $E_{\rm c}^{\rm p}$ and $E_{\rm v}^{\rm p}$ are the energies at the bottom of the conduction band, respectively the top of the valence band. The appropriate effective masses for electrons and holes to use in these expressions are comparable to the true electron masses for silicon and germanium. Setting the two expressions above equal allows μ to be computed.

The first exponential in (6.35) is the value of the number of electrons per state $\iota_{\rm e}$ at the bottom of the conduction band, and the second exponential is the number of holes per state $\iota_{\rm h}$ at the top of the valence band. The bottom line remains that semiconductors have much too few current carriers to have good conductivity.

That can be greatly improved by what is called doping the material. Suppose you have a semiconductor like germanium, that has 4 valence electrons per atom. If you replace a germanium atom in the crystal by a stray atom of a different element that has 5 valence electrons, then that additional electron is mismatched in the crystal structure. It can easily become dislocated and start roving through the conduction band. That allows additional conduction to occur. Even at very small concentrations, such impurity atoms can make a big difference. For example, you can increase the conductivity of germanium by a factor of a thousand by replacing 1 in a million germanium atoms by an arsenic one.

Because such valence-5 impurity atoms add electrons to the conduction band, they are called "donors." Because electrical conduction occurs by the negatively charged additional electrons provided by the doping, the doped semiconductor is called "n-type."

Alternatively, you can replace germanium atoms by impurity atoms that have only 3 valence electrons. That creates holes that can accept valence band electrons with a bit of thermal energy. Therefore such impurity atoms are called "acceptors." The holes in the valence band from which the electrons were taken allow electrical conduction to occur. Because the holes act like positively charged particles, the doped semiconductor is called "p-type."

Silicon has 4 valence band electrons just like germanium. It can be doped similarly.

Now consider an n-type semiconductor in more detail. As the center of figure 6.31 indicates, the effect of the donor atoms is to add a spike of energy states just below the conduction band. At absolute zero temperature, these states are filled with electrons and the conduction band is empty. And at absolute

zero, the Fermi level is always in between filled and empty states. So the Fermi level is now in the narrow gap between the spike and the conduction band. It illustrates that the Fermi level of a semiconductor can jump around wildly at absolute zero.

But what happens at absolute zero is irrelevant to a room temperature semiconductor anyway. At room temperature the Fermi level is typically as shown by the tick mark in figure 6.31. The Fermi level has moved up a lot compared to the intrinsic semiconductor, but it still stays well below the donor states.

If the Fermi level would still be in the middle of the band gap like for the undoped material, then there would be very few electrons in the donor states. But all the electrons that are in the donor states at absolute zero temperature cannot just disappear into nothing. And they cannot go into the intrinsic states. If the Fermi level does not change, the intrinsic states still have the same number of electrons as before the doping.

So the Fermi level cannot be in the middle of the band gap. And the Fermi level going down makes the missing electron problem worse; then there are even less electrons in the donor states and conduction band, and even more holes in the valence band.

The Fermi level must go up, significantly. For one, that will reduce the number of holes in the valence band. However, since the number of such holes is so tiny compared to the nuber of donor electrons, that does not help much. Much more importantly, the Fermi level μ going up will increase the number of electrons not just in the donor states, but also and especially in the lowest conduction band states, (6.33). The missing electron problem gets resolved: the electrons missing from the donor states are now in conduction band states. (Or to be picky, a very few of them have gone in valence band holes.)

Do note that while the Fermi level must go up, it cannot move too close to the donor states either. For assume the contrary, that the Fermi level is really close to the donor states. Then the donor states will be largely filled with electrons. But at room temperature the gap between the donor states and the conduction band is comparable to $k_{\mathrm{B}}T$. Therefore, if the donor states are largely filled with electrons, then the states at the bottom of the conduction band contain significant numbers of electrons too. Since there are so many of these conduction states compared to a typical number of donor states, the number of electrons in the conduction states would dwarf the number of electrons missing from the donor states. And that would mean that now there would be far too many electrons. The Fermi level must go up but stay low enough that the number of electrons per state ι_{e} stays small in both the donor states and conduction band. That is as sketched in figure 6.31.

If more donors are added, the Fermi level will move up more. Light doping may be on the order of 1 impurity atom in a 100 million, heavy doping 1 in 10,000. If the donor atoms get too close together, their electrons start to

interact. If that happens the spike of donor states broadens into a band extending to the conduction band, and you end up with a metallic "degenerate" semiconductor. For example, low temperature measurements show that phosphor donors turn silicon metallic at about 1 phosphor atom per 15 000 silicon ones. It may seem strange that impurity electrons at such a small concentration could interact at all. But note that 1 impurity in 15 000 atoms means that each $25 \times 25 \times 25$ cube of silicon atoms has one phosphor atom. On average the phosphor atoms are only about 25 atom spacings apart. In addition, the orbit of the very loosely bound donor electron is really far from the positively charged donor atom compared to the crystal spacing.

The p-type material is analyzed pretty much the same as n-type, with holes taking the place of electrons and acceptors the place of donors.

As already mentioned, the upward shift in the Fermi level in the n-type material has another effect besides providing lots of electrons in the conduction band. It decimates the already miserably small number of holes in the valence band that the undoped semiconductor had. That means that virtually all electrical conduction will now be performed by electrons, not holes. The electrons in n-type material are therefore called the "majority carriers" and the holes the "minority carriers."

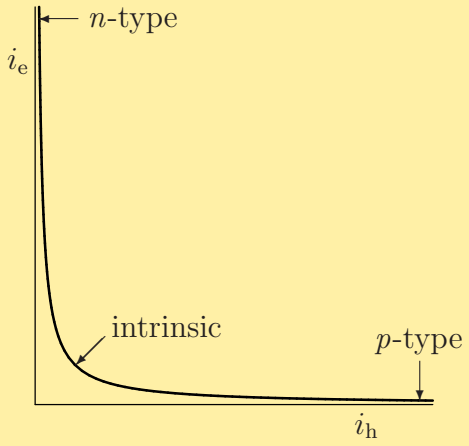


Figure 6.32: Relationship between conduction electron density and hole density. Intrinsic semiconductors have neither much conduction electrons nor holes.

The fact that raising the amount of conduction band electrons lowers the amount of valence band holes may be verified mathematically from (6.35). That equation implies that the product of the electron and hole densities is constant at a given temperature:

$$\boxed{i_e i_h = 4 \left(\frac{\sqrt{m_{\text{eff,e}} m_{\text{eff,e}}} k_B T}{2\pi\hbar^2} \right)^3 e^{-E^p_{\text{gap}}/k_B T}} \tag{6.36}$$

This relationship is called the "law of mass action" since nonexperts would be able to make sense out of "electron-hole density relation." And if you come to think of it, what is wrong with the name? Doesn't pretty much everything in physics come down to masses performing actions? That includes semiconductors too!

The relationship is plotted in figure 6.32. It shows that a high number of conduction electrons implies a very low number of holes. Similarly a p-type material with a high number of holes will have very few conduction electrons.

The law of mass action can also be understood from more classical arguments. That is useful since band theory has its limits. The classical picture is as follows: In thermal equilibrium, the semiconductor is bathed in blackbody radiation. A very small but nonzero fraction of the photons of this radiation have energies above the band gap. These will move valence band electrons to the conduction band, thus creating electron-hole pairs. In equilibrium, this creation of electron-hole pairs must be balanced by the removal of an identical amount of electron-hole pairs. The removal of a pair occurs through "recombination," in which an conduction band electron drops back into a valence band hole, eliminating both. The rate of recombinations will be proportional to the product of the densities of electrons and holes. Indeed, for a given number of holes, the more electrons there are, the more will be able to find holes under suitable conditions for recombination. And vice-versa with electrons and holes swapped. Equating a creation rate of electron-hole pairs by photons, call it A, to a removal rate of the form $Bi_{\mathrm{e}}i_{\mathrm{h}}$ shows that the product $i_{\mathrm{e}}i_{\mathrm{h}}$ equals the constant A/B. This constant will depend primarily on the Maxwell-Boltzmann factor $e^{-E^{\mathrm{p}}_{\mathrm{gap}}/k_{\mathrm{B}}T}$ that limits the number of photons that have sufficient energy to create pairs.

This classical picture also provides an intuitive explanation why adding both donors and acceptors to a semiconductor does not double the amount of current carriers over just one type of doping alone. Quite the opposite. As figure 6.32 shows, if the number of holes becomes comparable to the number of electrons, there are not many of either one. The semiconductor behaves again like an intrinsic one. The reason is that adding, say, some acceptors to an n-type material has the primary effect of making it much easier for the conduction band electrons to find valence band holes to recombine with. It is said that the added acceptors "compensate" for the donors.

Key Points

- Doping a semiconductor with donor atoms greatly increases the number of electrons in the conduction band. It produces an n-type semiconductor.
- Doping a semiconductor with acceptor atoms greatly increases the number of holes in the valence band. It produces an p-type semiconductor.

- The minority carrier gets decimated.
- The Fermi level is in the band gap, and towards the side of the majority carrier.
- There is compensation in doping. In particular, if there are about the same numbers of electrons and holes, then there are not many of either.

6.24 The *P-N* Junction

The *p-n* junction is the work horse of semiconductor applications. This section explains its physical nature, and why it can act as a current rectifier, among other things.

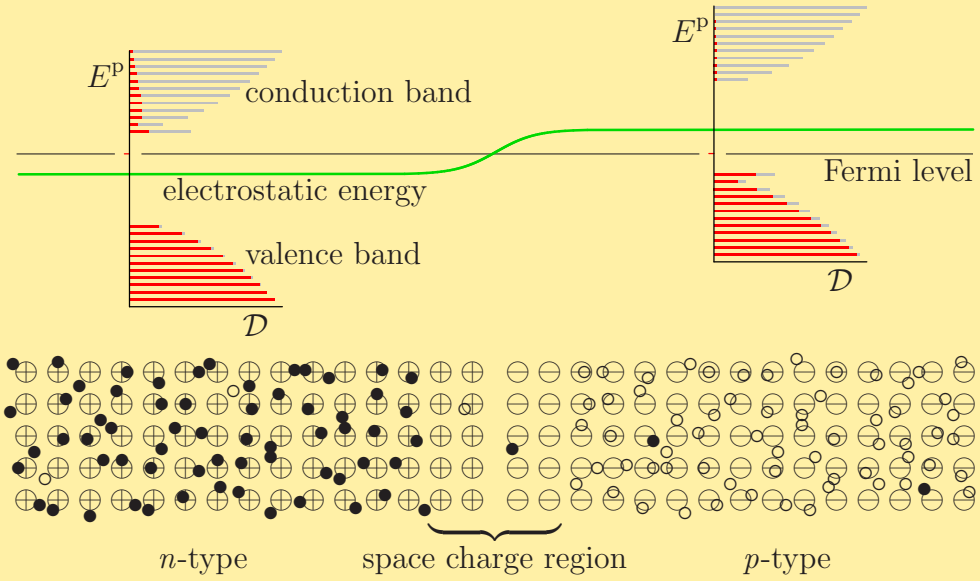


Figure 6.33: The *p-n* junction in thermal equilibrium. Top: energy spectra. Quantum states with electrons in them are in red. The mean electrostatic energy of the electrons is in green. Below: Physical schematic of the junction. The dots are conduction electrons and the small circles holes. The encircled plus signs are donor atoms, and the encircled minus signs acceptor atoms. (Donors and acceptors are not as regularly distributed, nor as densely, as this greatly simplified schematic suggests.)

A *p-n* junction is created by doping one side of a semiconductor crystal *n* type and the other side *p* type. As illustrated at the bottom of figure 6.33, the *n* side has a appreciable amount of conduction electrons, shown as black dots. These electrons have been provided by donor atoms. The donor atoms,

having given up one of their negatively charged electrons, have become positively charged and are shown as encircled plus signs.

The p side has a appreciable number of holes, quantum states that have lost their electrons. The holes are shown as small circles in the figure. Since a negatively charged electron is missing at a hole, the hole behaves as a positively charged particle. The missing electrons have been absorbed by acceptor atoms. These atoms have therefore acquired a negative charge and are shown by encircled minus signs.

The atoms are stuck in the crystal and cannot move. Electrical conduction takes place by means of motion of the electrons and holes. But under normal conditions, significant electrical conduction can only occur in one direction. That makes the p-n junction into a "diode," a current rectifier.

To see the basic reason is not difficult. In the so-called "forward" direction that allows a significant current, both the electrons in the n side and the holes in the p side flow towards the junction between the n and p sides. (Note that since electrons are negatively charged, they move in the direction opposite to the current.) In the vicinity of the junction, the incoming n-side electrons can drop into the incoming p-side holes. Phrased more formally, the electrons recombine with the holes. That can readily happen. A forward current flows freely if a suitable "forward-biased" voltage is applied.

However, if a "reverse-biased" voltage is applied, then normally very little current will flow. For a significant current in the reverse direction, both the electrons in the n side and the holes in the p side would have to flow away from the junction. So new conduction electrons and holes would have to be created near the junction to replace them. But random thermal motion can create only a few. Therefore there is negligible current.

While this simple argument explains why a p-n junction can act as a diode, it is not sufficient. It does not explain the true response of the current to a voltage. It also does not explain other applications of p-n junctions, such as transistors, voltage stabilizers, light-emitting diodes, solar cells, etcetera.

It turns out that in the forward direction, the recombination of the incoming electrons and holes is severely hindered by an electrostatic barrier that develops at the contact surface between the n-type and p-type material. This barrier is known as the "built-in potential." It is shown in green in figure 6.33.

Consider first the p-n junction in thermal equilibrium, when there is no current. The junction is shown in the lower part of figure 6.33. The n side has an excess amount of conduction electrons. The negative charge of these electrons is balanced by the positively charged donor atoms. Similarly, the p side has an excess amount of holes. The positive charge of these holes is balanced by the negatively charged acceptor atoms.

At the junction, due to random thermal motion the n-side electrons would want to diffuse into the p side. Similarly the p-side holes would want to diffuse into the n side. But that cannot go on indefinitely. These diffusion processes

cause a net negative charge to flow out of the n side and a net positive charge out of the p side. That produces the electrostatic barrier; it repels further n-side electrons from the p side and p-side holes from the n side.

The barrier takes the physical form of a double layer of positive charges next to negative charges. This layer is called the "space charge region." It is illustrated in figure 6.33. Double layers are common at contact surfaces between different solids. However, the one at the p-n junction is somewhat unusual as it consists of ionized donor and acceptor atoms. There are preciously few electrons and holes in the space charge region, and therefore the charges of the donors and acceptors are no longer offset by the electrons, respectively holes.

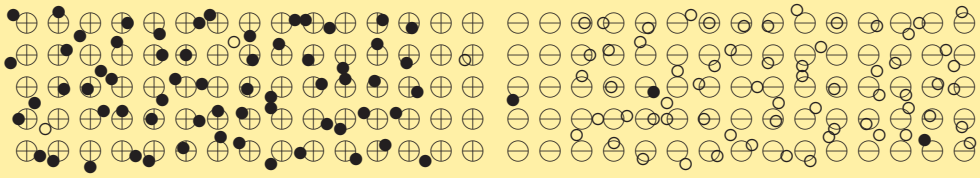
The reason for the lack of electrons and holes in the space charge region may be understood from figure 6.32: when the numbers of electrons and holes become comparable, there are not many of either. The lack of electrons and holes explains why the space charge region is also known as the "depletion layer."

The double layer is relatively thick. It has to be, to compensate for the fact that the fraction of atoms that are donors or acceptors is quite small. A typical thickness is 10^{-6} m, but this can vary greatly with doping level and any applied external voltage.

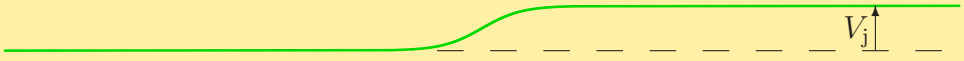
An n-side electron that tries to make it through the space charge region is strongly pulled back by the positive donors behind it and pushed back by the negative acceptors in front of it. Therefore there is a step-up in the electrostatic potential energy of an electron going through the region. This increase in potential energy is shown in green in figure 6.33. It raises the electron energy levels in the p side relative to the n side. In particular, it makes the chemical potentials, or Fermi levels, of the two sides equal. It has to do so; differences in chemical potential produce net electron diffusion, section 6.16. For the diffusion to stop, the chemical potential must become everywhere the same.

There is still some flow of electrons and holes through the junction, even in the absence of net current. It is due to random thermal motion. To simplify its description, it will be assumed that there is no significant recombination of electrons and holes while they pass through the space charge region, nor creation of new electrons and holes. That is a standard assumption, but by no means trivial. It requires great purification of the semiconductor. Crystal defects can act as "recombination centers," locations that help the electrons and holes recombine. For example, if you try to simply press separate n and p crystals together to create a p-n junction, it will not work. It will have far too many defects where the crystals meet. A proper recombination of electrons and holes should take place near the junction, but mostly outside the space charge region.

Consider now first the thermal flow of electrons and holes through the junction when there is no net current. It is sketched in figure 6.34*a*. All those n-side electrons would love to diffuse into the p side, but the electrostatic barrier is



(a) No voltage applied:



junction crossings by electrons:

$$-j_{\mathrm{e,min}} \quad -j_{\mathrm{e,maj}}$$

junction crossings by holes:

$$j_{\mathrm{h,maj}} \quad j_{\mathrm{h,min}}$$

(b) Forward biased:



junction crossings by electrons:

$$-j_{\mathrm{e,min}} \quad -j_{\mathrm{e,maj}}$$

junction crossings by holes:

$$j_{\mathrm{h,maj}} \quad j_{\mathrm{h,min}}$$

net current:

$$-j_{\mathrm{e,net}} + -j_{\mathrm{h,net}} = -j_{\mathrm{net}}$$

(c) Reverse biased:



junction crossings by electrons:

$$-j_{\mathrm{e,min}} \quad -j_{\mathrm{e,maj}}$$

junction crossings by holes:

$$j_{\mathrm{h,maj}} \quad j_{\mathrm{h,min}}$$

net current:

$$-j_{\mathrm{e,net}} + -j_{\mathrm{h,net}} = -j_{\mathrm{net}}$$

Figure 6.34: Schematic of the operation of an p-n junction.

holding them back. Only very few electrons have enough energy to make it through. The required amount of energy is the electrostatic energy increase over the junction. That energy will be called V_j. For n-side electrons to make it through the barrier, they need to have at least that much energy above the bottom of the n-side conduction band. The relative amount of electrons at those energy levels is primarily determined by the Maxwell-Boltzmann factor (6.33). It implies that there are a factor $e^{-V_\mathrm{j}/k_\mathrm{B}T}$ less electrons per quantum state with the additional energy V_j than there are at the bottom of the conduction band.

The crossings of these few very lucky electrons produce a miniscule current through the junction. It is indicated as $j_\mathrm{e,maj}$ in figure 6.34*a*. The electrons are called the majority carriers in the n side because there are virtually no holes in that side to carry current. Note also that the figure shows the negative currents for electrons, because that gives the direction that the electrons actually move. The currents in this discussion will be assumed to be per unit junction area, which explains why the symbol j is used instead of I. A junction twice as large produces double the current, all else being the same. All else being the same includes ignoring edge effects.

The miniscule current of the n-side majority electrons is balanced by an equally miniscule but opposite current $j_\mathrm{e,min}$ produced by p-side minority electrons that cross into the n side. Although the p side has very few conduction band electrons, the number of electrons per state is still the same as that of n-side electrons with enough energy to cross the barrier. And note that for the p-side electrons, there is no barrier. If they diffuse into the space charge region, the electrostatic potential will instead help them along into the n side.

For holes the story is equivalent. Because they have the opposite charge from the electrons, the same barrier that keeps the n-side electrons out of the p side also keeps the p-side holes out of the n side.

The bottom line is that there is no net current. And there should not be; otherwise you would have a battery that worked for free. Batteries must be powered by a chemical reaction.

But now suppose that a "forward-bias" external voltage φ is applied that lowers the barrier by an amount $e\varphi_\mathrm{j}$. What happens then is shown in figure 6.34*b*. The n-side majority electrons will now come pouring over the lowered barrier, and so will the p-side majority holes. Indeed, the Maxwell-Boltzmann factor for the majority carriers that can get through the barrier increases by a factor $e^{e\varphi_\mathrm{j}/k_\mathrm{B}T}$. That is a very large factor if the voltage change is bigger than about 0.025 volt, since $k_\mathrm{B}T$ is about 0.025 eV at normal temperatures. The currents of majority carriers explode, as sketched in the figure. And therefore, so does the net current.

The currents of minority carriers do not change appreciably. Whatever minority carriers diffuse into the space charge region still all pass through it. Note that the Fermi levels of the n and p sides do no longer match up when there is a current. If there is a current, the system is not in thermal equilibrium.

Figure 6.34*c* shows the case that a reverse bias voltage is applied. The reverse voltage increases the barrier for the majority carriers. The number that still have enough energy to cross the junction gets decimated to essentially zero. All that remains is a residual small reverse current of minority carriers through the junction.

Based on this discussion, it is straightforward to write a ballpark expression for the net current through the junction:

$$\boxed{j = j_0 e^{e\varphi_j/k_B T} - j_0} \tag{6.37}$$

The final term is the net reverse current due to the minority carriers. According to the above discussion, that current does not change with the applied voltage. The other term is the net forward current due to the majority carriers. According to the above discussion, it differs from the minority current primarily by a Maxwell-Boltzmann exponential. The energy in the exponential is the electrostatic energy due to the external voltage difference across the junction.

For forward bias the exponential explodes, producing significant current. For reverse bias, the exponential is essentially zero and only the small reverse minority current is left.

Equation (6.37) is known as the "Shockley diode equation." It works well for germanium but not quite that well for silicon. Silicon has a much larger band gap. That makes the minority currents much smaller still, which is good. But the correspondingly small reversed-biased and slightly forward-biased currents are sensitive to depletion layer electron-hole generation, respectively recombination. A fudge factor called the "ideality factor" is often added to the argument of the exponential to improve agreement.

Even for germanium, the Shockley diode equation applies only over a limited range. The equation does not include the resistance of the semiconductor. If the current increases rapidly, the voltage drop due to resistance does too, and it should be added to the voltage drop φ_j over the junction. That will eventually make the current versus voltage relation linear instead of exponential. And if the reverse voltage is too large, phenomena discussed in section 6.26 show up.

Key Points

- The p-n junction is the interface between an n-type and a p-type side of a semiconductor crystal.
- Under normal conditions, it will only conduct a significant current in one direction, called the forward direction.
- In the forward direction both the n-side electrons and the p-side holes move towards the junction.
- The Shockley diode equation describes the current versus voltage relation of p-n junctions, but only in a limited range.

At the junction a space-charge region exists. It provides a barrier for the majority carriers. However, it accelerates the minority carriers passing through the junction.

6.25 The Transistor

A second very important semiconductor device besides the p-n diode is the transistor. While the p-n diode allows currents to be blocked in one direction, the transistor allows currents to be regulated.

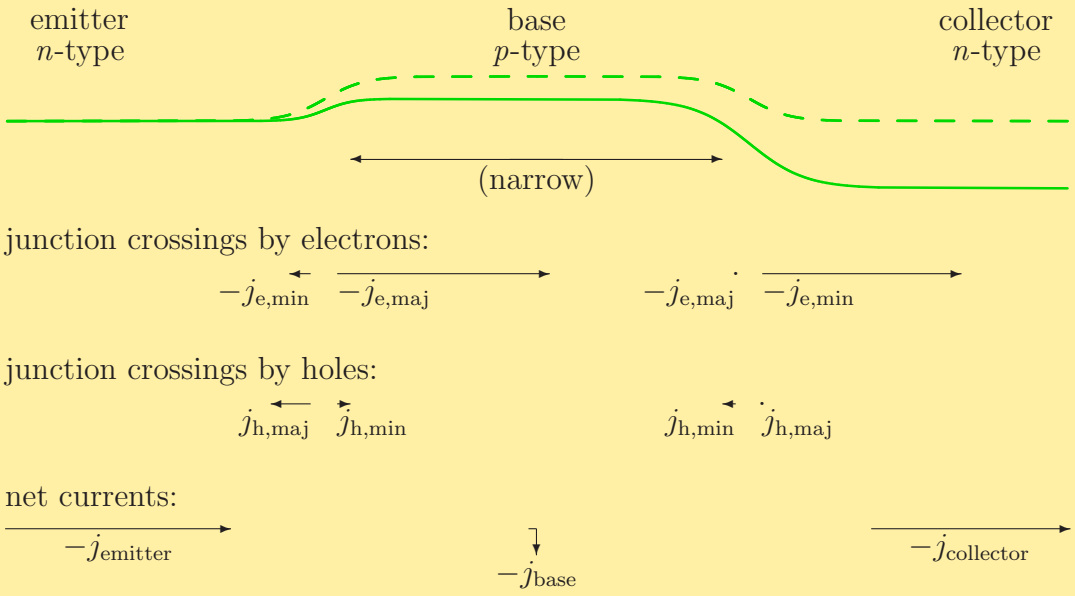


Figure 6.35: Schematic of the operation of an n-p-n transistor.

For example, an n-p-n transistor allows the current of electrons through an n-type semiconductor to be controlled. A schematic is shown in figure 6.35. Electrons flow through the transistor from one side, called the "emitter," to the other side, called the "collector."

To control this current, a very narrow region of p-type doping is sandwiched in between the two sides of n-type doping. This p-type region is called the "base." If the voltage at the base is varied, it regulates the current between emitter and collector.

Of course, when used in a circuit, electrodes are soldered to the emitter and collector, and a third one to the base. The transistor then allows the current between the emitter and collector electrodes to be controlled by the voltage of the base electrode. At the same time, a well-designed transistor will divert almost none of the current being regulated to the base electrode.

The transistor works on the same principles as the p-n junction of the previous section, with one twist. Consider first the flow of electrons through the device, as shown in figure 6.35. The junction between emitter and base is operated at a forward-bias voltage difference. Therefore, the majority electrons of the n-type emitter pour through it in great numbers. By the normal logic, these electrons should produce a current between the emitter and base electrodes.

But here comes the twist. The p region is made extremely thin, much smaller than its transverse dimensions and even much smaller than the diffusion distance of the electrons. Essentially all electrons that pour through the junction blunder into the second junction, the one between base and collector. Now this second junction is operated at a reverse-bias voltage. That produces a strong electric field that sweeps the electrons forcefully into the collector. (Remember that since the electrons are considered to be minority carriers in the base, they get sweeped through the junction by the electric field rather than stopped by it.)

As a result, virtually all electrons leaving the emitter end up as an electron flow to the collector electrode instead of to the base one as they should have. The stupidity of these electrons explains why the base voltage can regulate the current between emitter and collector without diverting much of it. Further, as seen for the p-n junction, the amount of electrons pouring through the junction from emitter to base varies very strongly with the base voltage. Small voltage changes at the base can therefore decimate or explode the electron flow, and almost all of it goes to the collector.

There is one remaining problem, however. The forward bias of the junction between emitter and base also means that the majority holes in the base pour through the junction towards the emitter. And that is strictly a current between the emitter and base electrodes. The holes cannot come from the collector, as the collector has virtually none. The hole current is therefore bad news. Fortunately, if you dope the p-type base only lightly, there are not that many majority holes, and virtually all current through the emitter to base junction will be carried by electrons.

A p-n-p transistor works just like an n-p-n-one, but with holes taking the place of electrons. There are other types of semiconductor transistors, but they use similar ideas.

Key Points

- A transistor allows current to be regulated.

6.26 Zener and Avalanche Diodes

Section 6.24 explained that normally no significant current will pass through a p-n junction in the reverse direction. The basic reason can be readily explained

in terms of the schematic of the p-n junction figure 6.33. A significant reverse current would require that the majority n-side conduction electrons and p-side holes both move away from the junction. That would require the creation of significant amounts of electron-hole pairs at the junction to replenish those that leave. Normally that will not happen.

But if the reverse voltage is increased enough, the diode can break down and a significant reverse current can indeed start to flow. That can be useful for voltage stabilization purposes.

Consider figure 6.33. One thing that can happen is that electrons in the valence band on the p side end up in the conduction band on the n side simply because of their quantum uncertainty in position. That process is called "tunneling." Diodes in which tunneling happens are called "Zener diodes."

The process requires that the energy spectrum at one location is raised sufficiently that its valence band reaches the level of the conduction band at another location. And the two locations must be extremely close together, as the quantum uncertainty in position is very small. Now it is the electrostatic potential, shown in green in figure 6.33, that raises the p-side spectra relative to the n-side ones. To raise a spectrum significantly relative to one very nearby requires a very steep slope to the electrostatic potential. And that in turn requires heavy doping and a sufficiently large reverse voltage to boost the built-in potential.

Once tunneling becomes a measurable effect, the current increases extremely rapidly with further voltage increases. That is a consequence of the fact that the strength of tunneling involves an exponential function, chapter 7.13 (7.74). The fast blow-up of current allows Zener diodes to provide a very stable voltage difference. The diode is put into a circuit that puts a nonzero tunneling current through the diode. Even if the voltage source in the circuit gets perturbed, the voltage drop across the Zener will stay virtually unchanged. Changes in voltage drops will remain restricted to other parts of the circuit; a corresponding change over the Zener would need a much larger change in current.

There is another way that diodes can break down under a sufficiently large reverse voltage. Recall that even under a reverse voltage there is still a tiny current through the junction. That current is due to the minority carriers, holes from the n side and conduction electrons from the p side. However, there are very few holes in the n side and conduction electrons in the p side. So normally this current can be ignored.

But that can change. When the minority carriers pass through the space charge region at the junction, they get accelerated by the strong electric field that exists there. If the reverse voltage is big enough, the space charge region can accelerate the minority carriers so much that they can knock electrons out of the valence band. The created electrons and holes will then add to the current.

Now consider the following scenario. A minority electron passes through the space charge region. Near the end of it, the electron has picked up enough

energy to knock a fellow electron out of the valence band. The two electrons continue on into the n side. But the created hole is swept by the electric field in the opposite direction. It goes back into the space charge region. Traveling almost all the way through it, near the end the hole has picked up enough energy to knock an electron out of the valence band. The created conduction electron is swept by the electric field in the opposite direction of the two holes, back into the space charge region... The single original minority electron has set off an avalanche of new conduction electrons and holes. The current explodes.

A diode designed to survive this is an "avalanche diode." Avalanche diodes are often loosely called Zener diodes, because the current explodes in a similar way. However, the physics is completely different.

Key Points

- Unlike the idealized theory suggests, under suitable conditions significant reverse currents can be made to pass through p-n junctions.
- It allows voltage stabilization.

6.27 Optical Applications

This section gives a concise overview of optical physics ranging from the x-ray spectrum of solids to semiconductor devices such as solar cells and light-emitting diodes.

6.27.1 Atomic spectra

Lone atoms have discrete electron energy levels, figure 6.19. An electron can transition from one of these levels to another by emitting or absorbing a photon of light. The energy of the photon is given by the difference in energy between the levels. Therefore emitted and absorbed photons have very specific energies, and corresponding very specific frequencies and wave lengths.

If they are in the visible range, they have very specific colors. The visible range of light corresponds to photons with energies from about 1.6 eV (red) to 3.2 eV (violet). In terms of the wave length of the light, the range is from about 390 nm (violet) to 760 nm (red).

A basic example is the red photon emitted in an E_3 to E_2 Balmer transition of a hydrogen atom, figure 4.8. Its energy is 1.89 eV and its wave length is 656 nm. In general, when the light emitted by excited lone atoms is sent through a prism, it separates into a few discrete thin beams of specific colors. The colors are characteristic for the type of atom that emitted the light.

Lone atoms can also absorb photons from light that passes them by. The same wave lengths that they can emit, they can also absorb. Absorbing a photon

puts the atoms in an excited state of higher energy. They may then subsequently emit a photon identical to the absorbed one in a different direction. Or they may lose their excitation energy in a transition between different energy levels, producing a photon of a different wave length. Or they may lose the energy in collisions. Either one eliminates the original photon altogether. For example, an excited hydrogen atom in the E_2 state might absorb a 656 nm photon to reach the E_3 state. Then it may transition directly back to the E_1 ground state. One 656 nm photon has then been eliminated.

In 1817 Fraunhofer gave a list of dark lines in the spectrum of sunlight. His list included the red E_2 to E_3 Balmer line, as well as the blue-green E_2 to E_4 one. It was eventually discovered that light at these frequencies is absorbed by the hydrogen atoms in the solar atmosphere. Other lines were due to absorption by other atoms like helium, sodium, calcium, titanium, and iron. The atoms present in the solar atmosphere could be identified without having to actually go there in a space ship. Since the days of Fraunhofer, spectroscopy has become one of the most important sources of information about the large-scale universe.

A typical solar spectrum also includes absorption lines due to molecules like oxygen and water vapor in the atmosphere of the earth. Molecular spectra tend to be more complicated than atomic ones, especially in the infrared region. That is due to relative motion of the different nuclei. The spectra are also more complicated due to the larger number of electrons involved.

Key Points

- Lone atoms and molecules emit and absorb light at specific wave lengths.
- It allows atoms and molecules to be recognized in the lab or far out in space.

6.27.2 Spectra of solids

Solids have electron energy levels arranged into continuous bands, figure 6.19. Therefore solids do not emit discrete wave lengths of light like lone atoms do. When light from solids is sent through a prism, the light will spread out into bands of gradually changing color. That is called "broadband" radiation.

To be sure, transitions involving the inner atomic electrons in solids still produce radiation at discrete wave lengths. The reason is that the energies of the inner electrons are not significantly different from the discrete values of the corresponding lone atoms. But because these energies are so much larger in magnitude, the produced radiation is in the X-ray range, not in the visible light range.

Key Points

- Solids can emit and absorb electromagnetic radiation in continuous bands.
- The X-ray range of the inner electrons is still discrete.

6.27.3 Band gap effects

As noted above, the light from solids is not limited to discrete wave lengths like that of lone atoms. But it is not true that solids can emit and absorb all wave lengths. In particular, a perfect crystal of an insulator with a large-enough band gap will be transparent to visible light. Take diamond as an example. Its valence band is completely filled with electrons but its conduction band is empty, as sketched in figure 6.36. A photon of light with enough energy can use its energy to take take an electron out of the valence band and put it into the conduction band. That leaves a hole behind in the valence band and eliminates the photon. However, to do this requires that the photon has at least the band gap energy of diamond, which is 5.5 eV. The photons of visible light have energies from about 1.6 eV to 3.2 eV. That is not enough. Visible light simply does not have enough energy to be absorbed by diamond electrons. Therefore a perfect diamond is transparent. Visible light passes through it unabsorbed.

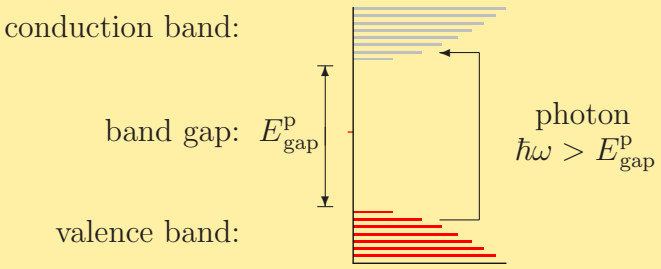


Figure 6.36: Vicinity of the band gap in the electron energy spectrum of an insulator. A photon of light with an energy greater than the band gap can take an electron from the valence band to the conduction band. The photon is absorbed in the process.

By this reasoning, all perfect crystals will be transparent if their band gap exceeds 3.2 eV. But actually, the energy of the photon can be somewhat *less* than the band gap and it may still be able to excite electrons. The model of energy states for noninteracting electrons that underlies spectra such as figure 6.36 is not perfect. The band gap in a spectrum is really the energy to create a conduction band electron and a valence band hole that do not interact. But the electron is negatively charged, and the hole acts as a positive particle. The two attract each other and can therefore form a bound state called an "exciton." The energy of the photon needed to create an exciton is less than the band gap

by the binding energy of the exciton. There is some additional slack due to variations in this binding energy. In the simplest model, the energy levels of lone excitons would be discrete like those of the hydrogen atom. However, they broaden considerably in the less than ideal environment of the solid.

If visible-light photons do not have enough energy to form electron-hole pairs nor excitons, the perfect crystal will be transparent. If the blue side of the visible spectrum has enough energy to excite electrons, the crystal will be colored reddish, since those components of light will remain unabsorbed.

Key Points

- A perfect crystal of a solid with a large enough band gap will be transparent.
- An exciton is a bound state of an electron and a hole.

6.27.4 Effects of crystal imperfections

It should be pointed out that in real life, the colors of most nonmetals are caused by crystal imperfections. For example, in ionic materials there may be a vacancy where a negative ion is missing. Since the vacancy has a net positive charge, an electron can be trapped inside it. That is called an "F-center." Because its energy levels are relatively small, such a center can absorb light in the visible range. Besides vacancies, chemical impurities are another common cause of optical absorption. A complete description of all the different types of crystal imperfections and their effects is beyond the scope of this book.

Key Points

- The colors of most nonmetals are caused by crystal imperfections.
- An electron bound to a vacancy in a ionic crystal is a called an F-center.

6.27.5 Photoconductivity

For a nonmetal with a sufficiently narrow band gap, photons of light may have enough energy to take electrons to the conduction band. Then both the electrons in the conduction band, as well as the holes that they leave behind in the valence band, can participate in electrical conduction through the solid. Increased electrical conductivity due to light is called "photoconductivity." It is used for a variety of light sensing devices and for Xerox copiers.

Note that excitons cannot directly produce electrical conduction, as the complete exciton is electrically neutral. However, excitons can create charge carriers

by interacting with crystal imperfections. Or photons with energies less than the band gap can do so themselves. In general, the mechanisms underlying photoconductivity are highly complex and strongly affected by crystal imperfections.

Key Points

- Photoconductivity is the increase in conductivity of nonmetals when photons of light create additional charge carriers.

6.27.6 Photovoltaic cells

In the vicinity of a p-n junction in a semiconductor crystal, light can do much more than just increase conductivity. It can *create* electricity. That is the principle of the "photovoltaic cell." These cells are also known as solar cells if the source of light is sunlight.

To understand how they work, consider the schematic of a p-n junction in figure 6.33. Suppose that the crystal is exposed to light. If the photons of light have more energy than the band gap, they can knock electrons out of the valence band. For example, silicon has a band gap of about 1.12 eV. And as noted above, the photons of visible light have energies from about 1.6 eV to 3.2 eV. So a typical photon of sunlight has plenty of energy to knock a silicon electron out of the valence band.

That produces a conduction band electron and a valence band hole. The two will move around randomly due to thermal motion. If they are close enough to the junction, they will eventually stumble into its space charge region, figure 6.33. The electric field in this region will forcefully sweep electrons to the n side and holes to the p side. Therefore, if the p-n junction is exposed to a continuous stream of light, there will be a continuous flow of new electrons to the n side and new holes to the p side. This creates a usable electric voltage difference between the two sides: the excess n-side electrons are willing to pass through an external load to recombine with the p-side holes.

There are limitations for the efficiency of the creation of electricity. The excess energy that the absorbed photons have above the band gap ends up as heat instead of as electrical power. And photons with insufficient energy to create electron-hole pairs do not contribute. Having p-n junctions with different band gaps absorb different wave lengths of the incoming light can significantly improve efficiency.

Key Points

- Photovoltaics is the creation of electricity by photons. Solar cells are an important example.

6.27.7 Light-emitting diodes

In the photovoltaic effect, light creates electricity. But the opposite is also possible. A current across a p-n junction can create light. That is the principle of the "light-emitting diode" (LED) and the "semiconductor laser."

Consider again the schematic of a p-n junction in figure 6.33. When a forward voltage is applied across the junction, n-side electrons stream into the p side. These electrons will eventually recombine with the prevailing holes in the p side. Simply put, the conduction electrons drop into the valence band holes. Similarly, p-side holes stream into the n side and eventually recombine with the prevailing electrons at that side. Each recombination releases a net amount of energy that is at least equal to the band gap energy. In a suitably chosen semiconductor, the energy can come out as light.

As section 6.22.4 discussed, silicon or germanium are not really suitable. They are what is called "indirect band gap" semiconductors. For these the energy is much more likely to come out as heat rather than light. Using various tricks, silicon can be made to emit some light, but the efficiency is low. LEDs normally use "direct band gap" semiconductors. The classical direct gap material is gallium arsenide, which produced the first patented infrared LED. To emit visible light, the band gap should exceed about 1.6 eV. Indeed, as noted earlier, the photons of visible light range from about 1.6 eV (red) to 3.2 eV (violet). That relates the band gap of the LED to its color. (For indirect gap semiconductors a phonon is involved, section 6.22.4, but its energy is small.) Gallium arsenide, with its 1.4 eV direct band gap emits infrared light with an average wave length of 940 nm. A 1.4 eV photon has a wave length of 885 nm. Diamond, with its 5.5 eV indirect band gap emits some ultraviolet light with an average wave length of 235 nm. A 5.5 eV photon has a wave length of 225 nm.

By the addition of a suitable optical cavity, a "diode laser" can be constructed that emits coherent light. The cavity lets the photons bounce a few times around through the region with the conduction electrons and holes. Now it is one of the peculiar symmetries of quantum mechanics that photons are not just good in taking electrons out of the valence band, they are also good at putting them back in. Because of energy conservation, the latter produces more photons than there were already; therefore it is called stimulated emission. Of course, bouncing the photons around might just get them absorbed again. But stimulated emission can win out over absorption if most electrons at the top of the valence band have been excited to the bottom of the conduction band. That is called a "population inversion." Such a situation can be achieved using a strong current across the junction. Under these conditions a photon may produce another photon through stimulated emission, then the two photons go on to stimulate the emission of still more photons, and so on in a runaway process. The result is coherent light because of the common origin of all the photons.

The idea of lasers is discussed in more detail in chapter 7.7.

Key Points

- A LED creates light due to the recombination of electrons and holes near a p-n junction. Normally, the semiconductor has a direct band gap.
- A laser diode adds an optical cavity to create coherent light.

6.28 Thermoelectric Applications

Thermoelectric effects can be used to make solid-state refrigeration devices, or to sense temperature differences, or to convert thermal energy directly into electricity. This section explains the underlying principles.

There are three different thermoelectric effects. They are named the Peltier, Seebeck, and Thomson effects after the researchers who first observed them. Thomson is better known as Kelvin.

These effects are not at all specific to semiconductors. However semiconductors are particularly suitable for thermoelectric applications. The reason is that the nature of the current carriers in semiconductors can be manipulated. That is done by doping the material as described in section 6.23. In an n-type doped semiconductor, currents are carried by mobile electrons. In a p-type doped semiconductor, the currents are carried by mobile holes, quantum states from which electrons are missing. Electrons are negatively charged particles, but holes act as positively charged ones. That is because a negatively charged electron is missing from a hole.

6.28.1 Peltier effect

Thermoelectric cooling can be achieved through what is called the "Peltier effect." The top part of figure 6.37 shows a schematic of a Peltier cooler. The typical device consists of blocks of a semiconductor like bismuth telluride that are alternately doped n-type and p-type. The blocks are electrically connected by strips of a metal like copper.

The connections are made such that when a current is passed through the device, both the n-type electrons and the p-type holes move towards the same side of the device. For example, in figure 6.37 both electrons and holes move to the top of the device. The current however is upward in the p-type blocks and downward in the n-type blocks. (Since electrons are negatively charged, their current is in the direction opposite to their motion.) The same current that enters a metal strip from one block leaves the strip again through the other block.

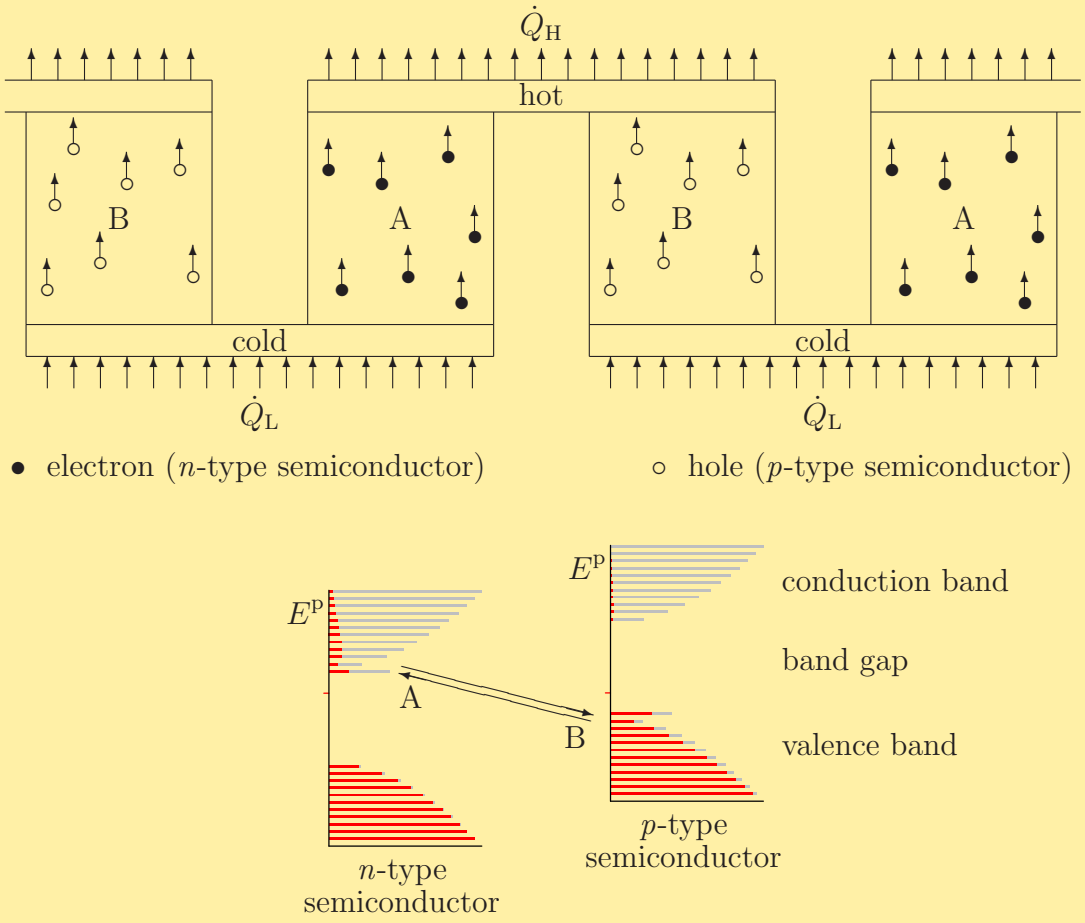


Figure 6.37: Peltier cooling. Top: physical device. Bottom: Electron energy spectra of the semiconductor materials. Quantum states filled with electrons are shown in red.

Consider now a metal strip at the top of the device in figure 6.37. Such a strip needs to take in a stream of conduction-band electrons from an n-type semiconductor block A. It must drop the same number of electrons into the valence-band holes coming in from a p-type semiconductor block B to eliminate them. As illustrated by the top arrow between the spectra at the bottom of figure 6.37, this lowers the energy of the electrons. Therefore energy is released, and the top strips get hot.

However, a bottom strip needs to take electrons out of the valence band of a p-type semiconductor B to create the outgoing holes. It needs to put these electrons into the conduction band of an n-type semiconductor A. That requires energy, so the bottom strips lose energy and cool down. You might think of it as evaporative cooling: the bottom strips have to give up their electrons with the highest thermal energy.

The net effect is that the Peltier cooler acts as a heat pump that removes

heat from the cold side and adds it to the hot side. It can therefore provide refrigeration at the cold side. At the time of writing, Peltier coolers use a lot more power to operate than a refrigerant-based device of the same cooling capability. However, the device is much simpler, and is therefore more suitable for various small applications. And it can easily regulate temperatures; a simple reversal of the current turns the cold side into the hot side.

Note that while the Peltier device connects p and n type semiconductors, it does not act as a diode. In particular, even in the bottom strips there is no need to raise electrons over the band gap of the semiconductor to create the new electrons and holes. Copper does not have a band gap.

It is true that the bottom strips must take electrons out of the p-type valence band and put them into the n-type conduction band. However, as the spectra at the bottom of figure 6.37 show, the energy needed to do so is much less than the band gap. The reason is that the p-type spectrum is raised relative to the n-type one. That is an effect of the electrostatic potential energies that are different in the two semiconductors. Even in thermal equilibrium, the spectra are at unequal levels. In particular, in equilibrium the electrostatic potentials adjust so that the chemical potentials, shown as red tick marks in the spectra, line up. The applied external voltage then decreases the energy difference even more.

The analysis of Peltier cooling can be phrased more generally in terms of properties of the materials involved. The "Peltier coefficient" $\mathcal{P}$ of a material is defined as the heat flow produced by an electric current, taken per unit current.

$$\boxed{\mathcal{P} \equiv \frac{\dot{Q}}{I}} \tag{6.38}$$

Here I is the current through the material and $\dot{Q}$ the heat flow it causes. Phrased another way, the Peltier coefficient is the thermal energy carried per unit charge. That gives it SI units of volts.

Now consider the energy balance of a top strip in figure 6.37. An electric current I_{AB} flows from material A to material B through the strip. (This current is negative as shown, but that is not important for the general formula.) The current brings along a heat flux $\dot{Q}_{\mathrm{A}} = \mathcal{P}_{\mathrm{A}} I_{\mathrm{AB}}$ from material A that flows into the strip. But a different heat flux $\dot{Q}_{\mathrm{B}} = \mathcal{P}_{\mathrm{B}} I_{\mathrm{AB}}$ leaves the strip through material B. The difference between what comes in and what goes out is what remains inside the strip to heat it:

$$\boxed{\dot{Q} = -\left(\mathcal{P}_{\mathrm{B}} - \mathcal{P}_{\mathrm{A}}\right) I_{\mathrm{AB}}} \tag{6.39}$$

This equation is generally valid; A and B do not need to be semiconductors. The difference in material Peltier coefficients is called the Peltier coefficient of the junction.

For the top strips in figure 6.37, $I_{\rm AB}$ is negative. Also, as discussed below, the n-type $\mathcal{P}_{\rm A}$ will be negative and the p-type $\mathcal{P}_{\rm B}$ positive. That makes the net heat flowing into the strip positive as it should be. Note also that the opposite signs of n-type and p-type Peltier coefficients really help to make the net heat flow as big as possible.

If there is a temperature gradient in the semiconductors in addition to the current, and there will be, it too will create a heat flow, {A.11}. This heat flow can be found using what is known as Fourier's law. It is bad news as it removes heat from the hot side and conducts it to the cold side.

A more quantitative understanding of the Peltier effect can be obtained using some ballpark Peltier coefficients. Consider again the spectra in figure 6.37. In the n-type semiconductor, each conduction electron has an energy per unit charge of about

$$\mathcal{P}_{n\text{ type}} \sim \frac{E^{\rm p}}{-e} = \frac{E_{\rm c}^{\rm p} + \frac{3}{2}k_{\rm B}T - \mu}{-e}$$

Here $-e$ in the denominator is the charge of the electron, while $E_{\rm c}^{\rm p}$ in the numerator is the energy at the bottom of the conduction band. It has been assumed that a typical electron in the conduction band has an additional random thermal energy equal to the classical value $\frac{3}{2}k_{\rm B}T$. Further the chemical potential, or Fermi level, μ has been taken as the zero level of energy.

The reason for doing the latter has to do with the fact that in thermal equilibrium, all solids in contact have the same chemical potential. That makes the chemical potential a convenient reference level of energy. The idea can be described graphically in terms of the spectra of figure 6.37. In the spectra, the chemical potential is indicated by the red tick marks on the vertical axes. Now consider again the energy change in transferring electrons between the n- and p-type materials. What determines it is how much the n-type electrons are higher in energy than the chemical potential and how much electrons put in the p-type holes are lower than it. (This assumes that the current remains small enough that the chemical potentials in the two semiconductors stay level. Otherwise theoretical description would become much more difficult.)

As this picture suggests, for the holes in the p-type semiconductor, the energy should be taken to be increasing downwards in the electron spectrum. It takes more energy to create a hole by taking an electron up to the Fermi level if the hole is lower in the spectrum. Therefore the Peltier coefficient of the p-doped semiconductor is

$$\mathcal{P}_{p\text{ type}} \sim \frac{E^{\rm p}}{e} = \frac{\mu - E_{\rm v}^{\rm p} + \frac{3}{2}k_{\rm B}T}{e}$$

where $E_{\rm v}^{\rm p}$ is the electron energy at the top of the valence band. Because holes act as positively charged particles, the Peltier coefficient of a p-type semiconductor

is positive. On the other hand, the Peltier coefficient of an n-type semiconductor is negative because of the negative charge in the denominator.

Note that both formulae are just ballparks. The thermal energy dragged along by a current is not simply the thermal equilibrium distribution of electron energy. The average thermal kinetic energy per current carrier to be used turns out to differ somewhat from $\frac{3}{2}k_{\mathrm{B}}T$. The current is also associated with a flow of phonons; their energy should be added to the thermal energy that is carried directly by the electrons or holes, {A.11}. Such issues are far beyond the scope of this book.

It is however interesting to compare the above semiconductor ballparks to one for metals:

$$\mathcal{P}_{\mathrm{metal}} \sim -\frac{2\pi^2}{9}\frac{\frac{3}{2}k_{\mathrm{B}}T}{E_{\mathrm{F}}^{\mathrm{p}}}\frac{\frac{3}{2}k_{\mathrm{B}}T}{e}$$

This ballpark comes from assuming the spectrum of a free-electron gas, {A.11}. The final ratio is easily understood as the classical thermal kinetic energy $\frac{3}{2}k_{\mathrm{B}}T$ per unit charge e. The ratio in front of it is the thermal energy divided by the Fermi energy $E_{\mathrm{F}}^{\mathrm{p}}$. As discussed in section 6.10, this fraction is much less than one. Its presence can be understood from the exclusion principle: as illustrated in figure 6.15, only a small fraction of the electrons pick up thermal energy in a metal.

The ballpark above implies that the Peltier coefficient of a metal is very much less than that of a doped semiconductor. It should however be noted that while the ballpark does give the rough order of magnitude of the Peltier coefficients of metals, they tend to be noticeably larger. Worse, there are quite a few metals whose Peltier coefficient is positive, unlike the ballpark above says.

To some extent, the lower Peltier coefficients of metals are compensated for by their larger electrical conductivity. A nondimensional figure of merit can be defined for thermoelectric materials as, {A.11}:

$$\frac{\mathcal{P}^2\sigma}{T\kappa}$$

where T is a typical operating absolute temperature. This figure of merit shows that a large Peltier coefficient is good, quadratically so, but so is a large electrical conductivity σ and a low thermal conductivity κ. Unfortunately, metals also conduct heat well.

Key Points

- In the Peltier effect, a current produces cooling or heating when it passes through the contact area between two solids.
- The heat released is proportional to the current and the difference in Peltier coefficients of the materials.
- Connections between oppositely-doped semiconductors work well.

6.28.2 Seebeck effect

Thermoelectric temperature sensing and power generation can be achieved by what is known as the "Seebeck effect." It is in some sense the opposite of the Peltier effect of the previous subsection.

Consider the configuration shown in figure 6.38. Blocks of n-type and p-type doped semiconductors are electrically connected at their tops using a copper strip. Copper strips are also attached to the bottoms of the semiconductor blocks. Unlike for the Peltier device, no external voltage source is attached. In the pure Seebeck effect, the bottom strips are electrically not in contact at all. So there is no current through the device. It is what is called an open-circuit configuration.

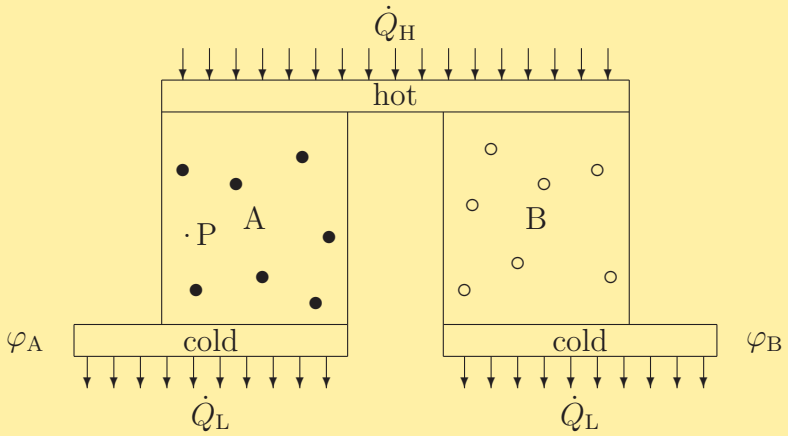


Figure 6.38: An example Seebeck voltage generator.

To achieve the Seebeck effect, heat from an external heat source is added to the top copper strip. That heats it up. Heat is allowed to escape from the bottom strips to, say, cooling water. This heat flow pattern is the exact opposite of the one for the Peltier cooler. If heat went out of the strips of your Peltier cooler at the cold side, it would melt your ice cubes.

But the Peltier cooler requires an external voltage to be supplied to keep the device running. The opposite happens for the Seebeck generator of figure 6.38. The device itself turns into a electric power supply. A voltage difference develops spontaneously between the bottom two strips.

That voltage difference can be used to determine the temperature of the top copper strip, assuming that the bottom strips are kept at a known temperature. A device that measures temperatures this way is called a "thermocouple."

Alternatively, you can extract electrical power from the voltage difference between the two bottom terminals. In that case the Seebeck device acts as a

"thermoelectric generator." Of course, to extract power you need to allow some current to flow. That will reduce the voltage below the pure Seebeck value.

To describe why the device works physically is not that easy. To understand the basic idea, consider an arbitrary point P in the n-type semiconductor, as indicated in figure 6.38. Imagine yourself standing at this point, shrunk down to microscopic dimensions. Due to random heat motion, conduction electrons come at you randomly from both above and below. However, those coming from above are hotter and so they come towards you at a higher speed. Therefore, assuming that all else is the same, there is a net electron current downwards at your location. Of course, that cannot go on, because it moves negative charge down, charging the lower part of the device negative and the top positive. This will create an electric field that slows down the hot electrons going down and speeds up the cold electrons going up. The voltage gradient associated with this electric field is the Seebeck effect, {A.11}.

In the Seebeck effect, an incremental temperature change $\mathrm{d}T$ in a material causes a corresponding change in voltage $\mathrm{d}\varphi$ given by:

$$\mathrm{d}\varphi_\mu = -\mathcal{S}\mathrm{d}T$$

The subscript on φ_μ indicates that the intrinsic chemical potential of the material must be included in addition to the electrostatic potential φ. In other words, φ_μ is the total chemical potential per unit electron charge. The constant $\mathcal{S}$ is a material coefficient depending on material and temperature.

This coefficient is sometimes called the "Seebeck coefficient." However, it is usually called the "thermopower" or "thermoelectric power." These names are much better, because the Seebeck coefficient describes an open-circuit voltage, in which no power is produced. It has units of V/K. It is hilarious to watch the confused faces of those hated nonspecialists when a physicist with a straight face describes something that is not, and cannot be, a power as the "thermopower."

The net voltage produced is the integrated total voltage change over the lengths of the two materials. If T_H is the temperature of the top strip and T_L that of the bottom ones, the net voltage can be written as:

$$\boxed{\varphi_\mathrm{B} - \varphi_\mathrm{A} = \int_{T_\mathrm{L}}^{T_\mathrm{H}} (\mathcal{S}_\mathrm{B} - \mathcal{S}_\mathrm{A})\,\mathrm{d}T} \tag{6.40}$$

This is the voltage that will show up on a voltmeter connected between the bottom strips. Note that there is no need to use the chemical potential φ_μ in this expression: since the bottom strips are both copper and at the same temperature, their intrinsic chemical potentials are identical.

The above equation assumes that the copper strips conduct heat well enough that their temperature is constant, (or alternatively, that materials A and B are in direct contact with each other at their top edges and with the voltmeter at

their bottom edges). Otherwise you would need to add an integral over the copper.

Note from the above equation that, given the temperature $T_{\rm L}$ of the bottom strips, the voltage only depends on the temperature $T_{\rm H}$ of the top strip. In terms of figure 6.38, the detailed way that the temperature varies with height is not important, just that the end values are T_H and T_L. That is great for your thermocouple application, because the voltage that you get only depends on the temperature at the tip of the thermocouple, the one you want to measure. It is not affected by whatever is the detailed temperature distribution in the two leads going to and from the tip. (As long as the material properties stay constant in the leads, that is. The temperature dependence of the Seebeck coefficients is not a problem.)

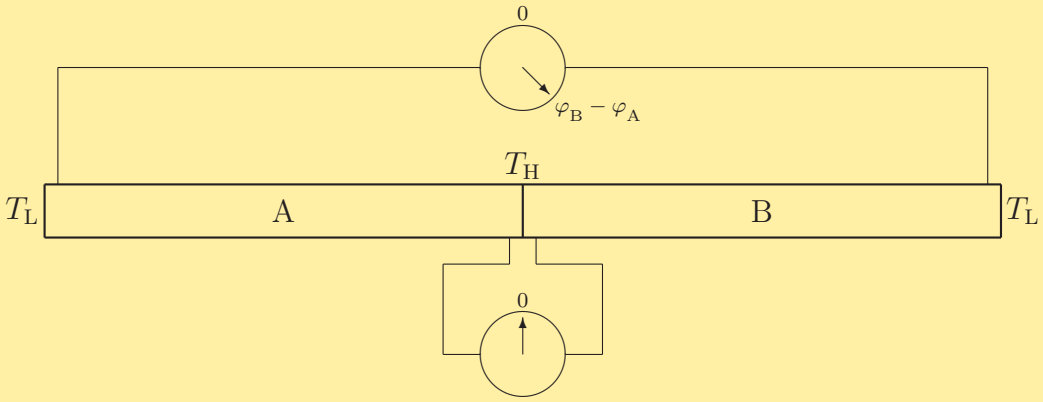


Figure 6.39: The Galvani potential jump over the contact surface does not produce a usable voltage.

It is sometimes suggested, even by some that surely know better like [22, p. 14-9], that the Seebeck potential is due to jumps in potential at the contact surfaces. To explain the idea, consider figure 6.39. In this figure materials A and B have been connected directly in order to simplify the ideas. It turns out that the mean electrostatic potential inside material A immediately before the contact surface with material B is different from the mean electrostatic potential inside material B immediately after the contact surface. The difference is called the Galvani potential. It is due to the charge double layer that exists at the contact surface between different solids. This charge layer develops to ensure that the chemical potentials are the same at both sides of the contact surface. Equality of chemical potentials across contact surfaces is a requirement for thermal equilibrium. Electrostatic potentials can be different.

If you try to measure this Galvani potential directly, like with the bottom voltmeter in figure 6.39, you fail. The reason is that there are also Galvani potential jumps between materials A and B and the leads of your voltmeter. Assume for simplicity that the leads of your voltmeter are both made of copper.

Because the chemical potentials are pairwise equal across the contact surfaces, all four chemical potentials are the same, including the two in the voltmeter leads. Therefore, the actual voltmeter can detect no difference between its two leads and gives a zero reading.

Now consider the top voltmeter in figure 6.39. This voltmeter does measure a voltage. Also in this case, the contact surfaces between the leads of the voltmeter and materials A and B are at a different temperature $T_{\rm L}$ than the temperature $T_{\rm H}$ of the contact surface between materials A and B. The suggestion is therefore sometimes made that changes in the Galvani potentials due to temperature differences produce the measured voltage. That would explain very neatly why the measured voltage only depends on the temperatures of the contact surfaces. Not on the detailed temperature distributions along the lengths of the materials.

It may be neat, but unfortunately it is also all wrong. The fact that the dependence on the temperature distribution drops out of the final result is just a mathematical coincidence. As long as the changes in intrinsic chemical potential can be ignored, the Galvani potential jumps still sum to zero. Not to the measured potential. After all, in that case the voltage changes over the lengths of the materials are the same as the chemical potential changes. And because they already sum to the measured voltage, there is nothing left for the Galvani jumps. Consider for example the free-electron gas model of metals. While its intrinsic chemical potential does change with temperature, {D.62}, that change is only one third of the potential change produced by the Seebeck coefficient given in addendum {A.11}. Galvani potential changes then sum to only a third of the measured potential. No, there is no partial credit.

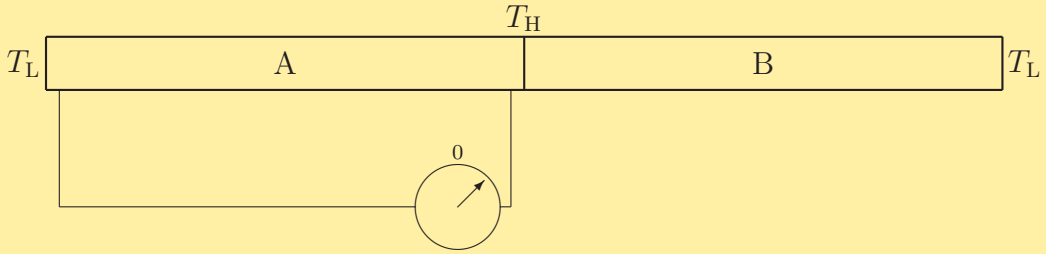


Figure 6.40: The Seebeck effect is not directly measurable.

It should also be pointed out that the Seebeck effect of a material is not directly measurable. Figure 6.40 illustrates an attempt to directly measure the Seebeck effect of material A. Unfortunately, the only thing that changes compared to figure 6.39 is that the two leads of the voltmeter take over the place of material B. Unless the two leads are attached to points of equal temperature, they are an active part of the total Seebeck effect measured. (Superconductors should have their Seebeck coefficient zero. However, finding superconductors that still are superconductors if they are in thermal contact with real-life temperatures is an obvious issue.)

Kelvin discovered that you can find the Seebeck coefficient $\mathcal{S}$ from the Peltier coefficient $\mathcal{P}$ simply by dividing by the absolute temperature. Unfortunately, the Peltier coefficient is not directly measurable either. Its effect too requires a second material to be present to compare against. It does show, however, that good materials for the Peltier effect are also good materials for the Seebeck effect.

You might wonder where the charges that transfer between the hot and cold sides in the Seebeck effect end up. In thermal equilibrium, the interiors of solids need to stay free of net electric charge, or a current would develop to eliminated the charge difference. But in the Seebeck effect, the solids are not in thermal equilibrium. It is therefore somewhat surprising that the interiors do remain free of net charge. At least, they do if the temperature variations are small enough, {A.11}. So the charges that transfer between hot and cold, and so give rise to the Seebeck potential difference, end up at the surfaces of the solids. Not in the interior. Even in the Seebeck effect.

Key Points

- The Seebeck effect produces a usable voltage from temperature differences.
- It requires two different materials in electrical contact to span the temperature difference.
- The voltage is the difference in the integrals of the Seebeck coefficients of the two materials with respect to temperature.
- The Seebeck coefficient is usually called thermopower because it is not power.

6.28.3 Thomson effect

The "Thomson effect," or "Kelvin heat," describes the heat release in a material with a current through it. This heat release is directly measurable. That is unlike the Peltier and Seebeck effects, for which only the net effect of two different materials can be measured. Since the Peltier and Seebeck coefficients can be computed from the Thomson one, in principle the Thomson effect allows all three thermoelectric coefficients to be found without involving a second material.

Thomson, who later became lord Kelvin, showed that the net energy accumulation per unit volume in a bar of material with a current through it can be written as:

$$\boxed{\dot{e} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\kappa \frac{\mathrm{d}T}{\mathrm{d}x}\right) + \frac{j^2}{\sigma} - \mathcal{K} j \frac{\mathrm{d}T}{\mathrm{d}x}} \tag{6.41}$$

Here x is the position along the bar, T is the temperature, j is the current per unit area, and κ and σ are the thermal and electrical conductivities. The first term in the right hand side is heat accumulation due to Fourier's law of heat conduction. The second term is the Joule heating that keeps your resistance heater working. The final term is the thermoelectric Thomson effect or Kelvin heat. (The term "Kelvin effect" is not used because it is already in common use for something else.) The coefficient $\mathcal{K}$ is called the "Kelvin coefficient" or "Thomson coefficient." A derivation from the general equations of thermoelectrics is given in addendum {A.11}.

It may be noted that for devices in which the Thomson effect is important, the figure of merit introduced earlier becomes less meaningful. In such cases, a second nondimensional number based on the Kelvin coefficient will also affect device performance.

The other two thermoelectric coefficients can be computed from the Kelvin one using the Kelvin, or Thomson, relationships {A.11}:

$$\boxed{\frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\ln T} = \mathcal{K} \qquad \mathcal{P} = \mathcal{S}T} \tag{6.42}$$

By integrating $\mathcal{K}$ with respect to $\ln T$ you can find the Seebeck coefficient and from that the Peltier one.

That requires of course that you find the Kelvin coefficient over the complete temperature range. But you only need to do it for one material. As soon as you accurately know the thermoelectric coefficients for one material, you can use that as the reference material to find Peltier and Seebeck coefficients for every other material. Lead is typically used as the reference material, as it has relatively low thermoelectric coefficients.

Of course, if it turns out that the data on your reference material are not as accurate as you thought they were, it would be very bad news. It will affect the accuracy of the thermoelectric coefficients of every other material that you found using this reference material. A prediction on whether such a thing was likely to happen for lead could be derived from what is known as Murphy's law.

Key Points

- The Thomson effect, or Kelvin heat, describes the internal heating in a material with a current going through it. More precisely, it describes the part of this heating that is due to interaction of the current with the temperature changes.
- Unlike the Peltier and Seebeck coefficients, the Kelvin (Thomson) coefficient can be measured without involving a second material.
- The Kelvin (Thomson) relations allow you to compute the Peltier and Seebeck coefficients from the Kelvin one.

Chapter 7

Time Evolution

Abstract

The evolution of systems in time is less important in quantum mechanics than in classical physics, since in quantum mechanics so much can be learned from the energy eigenvalues and eigenfunctions. Still, time evolution is needed for such important physical processes as the creation and absorption of light and other radiation. And many other physical processes of practical importance are simplest to understand in terms of classical physics. To translate a typical rough classical description into correct quantum mechanics requires an understanding of unsteady quantum mechanics.

The chapter starts with the introduction of the Schrödinger equation. This equation is as important for quantum mechanics as Newton's second law is for classical mechanics. A formal solution to the equation can be written immediately down for most systems of interest.

One direct consequence of the Schrödinger equation is energy conservation. Systems that have a definite value for their energy conserve that energy in the simplest possible way: they just do not change at all. They are stationary states. Systems that have uncertainty in energy do evolve in a nontrivial way. But such systems do still conserve the probability of each of their possible energy values.

Of course, the energy of a system is only conserved if no devious external agent is adding or removing energy. In quantum mechanics that usually boils down to the condition that the Hamiltonian must be independent of time. If there is a nasty external agent that does mess things up, analysis may still be possible if that agent is a slowpoke. Since physicists do not know how to spell slowpoke, they call this the adiabatic approximation. More precisely, they call it adiabatic because they know how to spell adiabatic, but not what it means.

The Schrödinger equation is readily used to describe the evolution of expectation values of physical quantities. This makes it possible to show

that Newton's equations are really an approximation of quantum mechanics valid for macroscopic systems. It also makes it possible to formulate the popular energy-time uncertainty relationship.

Next, the Schrödinger equation does not just explain energy conservation. It also explains where other conservation laws such as conservation of linear and angular momentum come from. For example, angular momentum conservation is a direct consequence of the fact that space has no preferred direction.

It is then shown how these various conservation laws can be used to better understand the emission of electromagnetic radiation by say an hydrogen atom. In particular, they provide conditions on the emission process that are called selection rules.

Next, the Schrödinger equation is used to describe the detailed time evolution of a simple quantum system. The system alternates between two physically equivalent states. That provides a model for how the fundamental forces of nature arise. It also provides a model for the emission of radiation by an atom or an atomic nucleus.

Unfortunately, the model for emission of radiation turns out to have some problems. These require the consideration of quantum systems involving two states that are not physically equivalent. That analysis then finally allows a comprehensive description of the interaction between atoms and the electromagnetic field. It turns out that emission of radiation can be stimulated by radiation that already exists. That allows for the operation of masers and lasers that dump out macroscopic amounts of monochromatic, coherent radiation.

The final sections discuss examples of the nontrivial evolution of simple quantum systems with infinite numbers of states. Before that can be done, first the so-far neglected eigenfunctions of position and linear momentum must be discussed. Position eigenfunctions turn out to be spikes, while linear momentum eigenfunctions turn out to be waves. Particles that have significant and sustained spatial localization can be identified as "packets" of waves. These ideas can be generalized to the motion of conduction electrons in crystals.

The motion of such wave packets is then examined. If the forces change slowly on quantum scales, wave packets move approximately like classical particles do. Under such conditions, a simple theory called the WKB approximation applies.

If the forces vary more rapidly on quantum scales, more weird effects are observed. For example, wave packets may be repelled by attractive forces. On the other hand, wave packets can penetrate through barriers even though classically speaking, they do not have enough energy to do so. That is called tunneling. It is important for various applications. A

simple estimate for the probability that a particle will tunnel through a barrier can be obtained from the WKB approximation.

Normally, a wave packet will be partially transmitted and partially reflected by a finite barrier. That produces the weird quantum situation that the same particle is going in two different directions at the same time. From a more practical point of view, scattering particles from objects is a primary technique that physicists use to examine nature.

7.1 The Schrödinger Equation

In Newtonian mechanics, Newton's second law states that the linear momentum changes in time proportional to the applied force; $\mathrm{d}m\vec{v}/\mathrm{d}t = m\vec{a} = \vec{F}$. The equivalent in quantum mechanics is the Schrödinger equation, which describes how the wave function evolves. This section discusses this equation, and a few of its immediate consequences.

7.1.1 The equation

The Schrödinger equation says that the time derivative of the wave function is obtained by applying the Hamiltonian on it. More precisely:

$$\boxed{\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = H\Psi} \tag{7.1}$$

An equivalent and earlier formulation of quantum mechanics was given by Heisenberg, {A.12}. However, the Schrödinger equation tends to be easier to deal with, especially in nonrelativistic applications. An integral version of the Schrödinger equation that is sometimes convenient is in {A.13}.

The Schrödinger equations is nonrelativistic. The simplest relativistic version is called the Klein-Gordon equation. A discussion is in addendum {A.14}. However, relativity introduces a fundamentally new issue: following Einstein's mass-energy equivalence, particles may be created out of pure energy or destroyed. To deal with that, you typically need a formulation of quantum mechanics called quantum field theory. A very brief introduction is in addendum {A.15}.

Key Points

- The Schrödinger equation describes the time evolution of the wave function.
- The time derivative is proportional to the Hamiltonian.

7.1.2 Solution of the equation

The solution to the Schrödinger equation can immediately be given for most cases of interest. The only condition that needs to be satisfied is that the Hamiltonian depends only on the state the system is in, and not explicitly on time. This condition is satisfied in all cases discussed so far, including the particle in a box, the harmonic oscillator, the hydrogen and heavier atoms, and the molecules, so the following solution applies to them all:

> *To satisfy the Schrödinger equation, write the wave function Ψ in terms of whatever are the energy eigenfunctions $\psi_{\vec{n}}$ of the Hamiltonian,*
>
> $$\Psi = c_{\vec{n}_1}(t)\psi_{\vec{n}_1} + c_{\vec{n}_2}(t)\psi_{\vec{n}_2} + \ldots = \sum_{\vec{n}} c_{\vec{n}}(t)\psi_{\vec{n}} \tag{7.2}$$
>
> *Then the coefficients $c_{\vec{n}}$ must evolve in time as complex exponentials:*
>
> $$\boxed{c_{\vec{n}}(t) = c_{\vec{n}}(0)e^{-\mathrm{i}E_{\vec{n}}t/\hbar}} \tag{7.3}$$
>
> *for every combination of quantum numbers $\vec{n}$.*

In short, you get the wave function for arbitrary times by taking the initial wave function and shoving in additional factors $e^{-\mathrm{i}E_{\vec{n}}t/\hbar}$. The initial values $c_{\vec{n}}(0)$ of the coefficients are not determined from the Schrödinger equation, but from whatever initial condition for the wave function is given. As always, the appropriate set of quantum numbers $\vec{n}$ depends on the problem.

Consider how this works out for the electron in the hydrogen atom. Here each spatial energy state ψ_{nlm} is characterized by the three quantum numbers n, l, m, chapter 4.3. However, there is a spin-up version $\psi_{nlm}\uparrow$ of each state in which the electron has spin magnetic quantum number $m_s = \frac{1}{2}$, and a spin-down version $\psi_{nlm}\downarrow$ in which $m_s = -\frac{1}{2}$, chapter 5.5.1. So the states are characterized by the set of four quantum numbers

$$\vec{n} \equiv (n, l, m, m_s)$$

The most general wave function for the hydrogen atom is then:

$$\Psi(r,\theta,\phi,S_z,t) =$$

$$\sum_{n=1}^{\infty}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} c_{nlm\frac{1}{2}}(0)e^{-\mathrm{i}E_n t/\hbar}\psi_{nlm}(r,\theta,\phi)\uparrow + c_{nlm,\text{-}\frac{1}{2}}(0)e^{-\mathrm{i}E_n t/\hbar}\psi_{nlm}(r,\theta,\phi)\downarrow$$

Note that each eigenfunction has been given its own coefficient that depends exponentially on time. (The summation limits come from chapter 4.3.)

The given solution in terms of eigenfunctions covers most cases of interest, but as noted, it is not valid if the Hamiltonian depends explicitly on time. That

possibility arises when there are external influences on the system; in such cases the energy does not just depend on what state the system itself is in, but also on what the external influences are like at the time.

Key Points

- Normally, the coefficients of the energy eigenfunctions must be proportional to $e^{-\mathrm{i}E_{\vec{n}}t/\hbar}$.

7.1.2 Review Questions

1. The energy of a photon is $\hbar\omega$ where ω is the classical frequency of the electromagnetic field produced by the photon. So what is $e^{-\mathrm{i}E_{\vec{n}}t/\hbar}$ for a photon? Are you surprised by the result?

   Solution schrodsol-a

2. For the one-dimensional harmonic oscillator, the energy eigenvalues are

   $$E_n = \frac{2n+1}{2}\omega$$

   Write out the coefficients $c_n(0)e^{-\mathrm{i}E_n t/\hbar}$ for those energies.

   Now classically, the harmonic oscillator has a natural frequency ω. That means that whenever ωt is a whole multiple of 2π, the harmonic oscillator is again in the same state as it started out with. Show that the coefficients of the energy eigenfunctions have a natural frequency of $\frac{1}{2}\omega$; $\frac{1}{2}\omega t$ must be a whole multiple of 2π for the coefficients to return to their original values.

   Solution schrodsol-b

3. Write the full wave function for a one-dimensional harmonic oscillator. Formulae are in chapter 4.1.2.

   Solution schrodsol-c

7.1.3 Energy conservation

The Schrödinger equation implies that the energy of a system is conserved, assuming that there are no external influences on the system.

To see why, consider the general form of the wave function:

$$\Psi = \sum_{\vec{n}} c_{\vec{n}}(t)\psi_{\vec{n}} \qquad c_{\vec{n}}(t) = c_{\vec{n}}(0)e^{-\mathrm{i}E_{\vec{n}}t/\hbar}$$

According to chapter 3.4, the square magnitudes $|c_{\vec{n}}|^2$ of the coefficients of the energy eigenfunctions give the probability for the corresponding energy. While the coefficients vary with time, their square magnitudes do not:

$$|c_{\vec{n}}(t)|^2 \equiv c^*_{\vec{n}}(t)c_{\vec{n}}(t) = c^*_{\vec{n}}(0)e^{\mathrm{i}E_{\vec{n}}t/\hbar}c_{\vec{n}}(0)e^{-\mathrm{i}E_{\vec{n}}t/\hbar} = |c_{\vec{n}}(0)|^2$$

So the probability of measuring a given energy level does not vary with time either. That means that energy is conserved.

For example, a wave function for a hydrogen atom at the excited energy level E_2 might be of the form:

$$\Psi = e^{-\mathrm{i}E_2 t/\hbar}\psi_{210}\uparrow$$

(This corresponds to an assumed initial condition in which all coefficients c_{nlmm_s} are zero except $c_{2101} = 1$.) The square magnitude of the exponential is one, so the energy of this excited atom will stay E_2 with 100% certainty for all time. The energy of the atom is conserved.

This is an important example, because it also illustrates that an excited atom will stay excited for all time if left alone. That is an apparent contradiction because, as discussed in chapter 4.3, the above excited atom will eventually emit a photon and transition back to the ground state. Even if you put it in a sealed box whose interior is at absolute zero temperature, it will still decay.

The explanation for this apparent contradiction is that an atom is never truly left alone. Simply put, even at absolute zero temperature, quantum uncertainty in energy allows an electromagnetic photon to pop up that perturbs the atom and causes the decay. (To describe more precisely what happens is a major objective of this chapter.)

Returning to the unperturbed atom, you may wonder what happens to energy conservation if there is uncertainty in energy. In that case, what does not change with time are the probabilities of measuring the possible energy levels. As an arbitrary example, the following wave function describes a case of an unperturbed hydrogen atom whose energy has a 50/50 chance of being measured as E_1, (-13.6 eV), or as E_2, (-3.4 eV):

$$\Psi = \frac{1}{\sqrt{2}}e^{-\mathrm{i}E_1 t/\hbar}\psi_{100}\downarrow + \frac{1}{\sqrt{2}}e^{-\mathrm{i}E_2 t/\hbar}\psi_{210}\uparrow$$

The 50/50 probability applies regardless how long the wait is before the measurement is done.

You can turn the observations of this subsection also around. If an external effect changes the energy of a system, then clearly the probabilities of the individual energies must change. So then the coefficients of the energy eigenfunctions cannot be simply vary exponentially with time as they do for the unperturbed systems discussed above.

Key Points

- Energy conservation is a fundamental consequence of the Schrödinger equation.
- An isolated system that has a given energy retains that energy.

Even if there is uncertainty in the energy of an isolated system, still the probabilities of the various energies do not change with time.

7.1.4 Stationary states

The quest for the dynamical implications of the Schrödinger equation must start with the simplest case. That is the case in which there is only a single energy eigenfunction involved. Then the wave function is of the form

$$\Psi = c_{\vec{n}}(0)e^{-\mathrm{i}E_{\vec{n}}t/\hbar}\psi_{\vec{n}}$$

Such states are called "stationary states." Systems in their ground state are of this type.

To see why these states are called stationary, note first of all that the energy of the state is $E_{\vec{n}}$ for all time, with no uncertainty.

But energy is not the only thing that does not change in time. According to the Born interpretation, chapter 3.1, the square magnitude of the wave function of a particle gives the probability of finding the particle at that position and time. Now the square magnitude of the wave function above is

$$|\Psi|^2 = |\psi_{\vec{n}}|^2$$

Time has dropped out in the square magnitude; the probability of finding the particle is the same for all time.

For example, consider the case of the particle in a pipe of chapter 3.5. If the particle is in the ground state, its wave function is of the form

$$\Psi = c_{111}(0)e^{-\mathrm{i}E_{111}t/\hbar}\psi_{111}$$

The precise form of the function ψ_{111} is not of particular interest here, but it can be found in chapter 3.5.

The relative probability for where the particle may be found can be shown as grey tones:



Figure 7.1: The ground state wave function looks the same at all times.

The bottom line is that this picture is the same for all time.

If the wave function is purely the first excited state ψ_{211}, the corresponding picture looks for all time like:



Figure 7.2: The first excited state at all times.

And it is not just position that does not change. Neither do linear or angular momentum, kinetic energy, etcetera. That can be easily checked. The probability for a specific value of any physical quantity is given by

$$|\langle\alpha|\Psi\rangle|^2$$

where α is the eigenfunction corresponding to the value. (If there is more than one eigenfunction with that value, sum their contributions.) The exponential drops out in the square magnitude. So the probability does not depend on time.

And if probabilities do not change, then neither do expectation values, uncertainties, etcetera. No physically meaningful quantity changes with time.

Hence it is not really surprising that none of the energy eigenfunctions derived so far had any resemblance to the classical Newtonian picture of a particle moving around. Each energy eigenfunction by itself is a stationary state. There is no change in the probability of finding the particle regardless of the time that you look. So how could it possibly resemble a classical particle that is at different positions at different times?

To get time variations of physical quantities, states of different energy must be combined. In other words, there must be uncertainty in energy.

Key Points

- States of definite energy are stationary states.
- To get nontrivial time variation of a system requires uncertainty in energy.

7.1.5 The adiabatic approximation

The previous subsections discussed the solution for systems in which the Hamiltonian does not explicitly depend on time. Typically that means isolated systems, unaffected by external effects, or systems for which the external effects are relatively simple. If the external effects produce a time-dependent Hamiltonian, things get much messier. You cannot simply make the coefficients of the eigenfunctions vary exponentially in time as done in the previous subsections.

However, dealing with systems with time-dependent Hamiltonians can still be relatively easy if the Hamiltonian varies sufficiently slowly in time. Such systems are quasi-steady ones.

So physicists cannot call these systems quasi-steady; that would give the secret away to these hated nonspecialists and pesky students. Fortunately, physicists were able to find a much better name. They call these systems "adiabatic." That works much better because the word "adiabatic" is a well-known term in thermodynamics: it indicates systems that evolve *fast* enough that heat conduction with the surroundings can be ignored. So, what better name to use also for quantum systems that evolve *slowly* enough that they stay in equilibrium with their surroundings? No one familiar with even the most basic thermodynamics will ever guess what it means.

As a simple example of an adiabatic system, assume that you have a particle in the ground state in a box. Now you change the volume of the box by a significant amount. The question is, will the particle still be in the ground state after the volume change? Normally there is no reason to assume so; after all, either way the energy of the particle will change significantly. However, the "adiabatic theorem" says that if the change is performed slowly enough, it will. The particle will indeed remain in the ground state, even though that state slowly changes into a completely different form.

If the system is in an energy state other than the ground state, the particle will stay in that state as it evolves during an adiabatic process. The theorem does assume that the energy is nondegenerate, so that the energy state is unambiguous. More sophisticated versions of the analysis exist to deal with degeneracy and continuous spectra.

A derivation of the theorem can be found in {D.34}. Some additional implications are in addendum {A.16}. The most important practical application of the adiabatic theorem is without doubt the Born-Oppenheimer approximation, which is discussed separately in chapter 9.2.

Key Points

- If the properties of a system in its ground state are changed, but slowly, the system will remain in the changing ground state.
- More generally, the "adiabatic" approximation can be used to analyze slowly changing systems.
- No, it has nothing to do with the normal use of the word "adiabatic."

7.2 Time Variation of Expectation Values

The time evolution of systems may be found using the Schrödinger equation as described in the previous section. However, that requires the energy eigenfunctions to be found. That might not be easy.

For some systems, especially for macroscopic ones, it may be sufficient to figure out the evolution of the expectation values. An expectation value of a

physical quantity is the average of the possible values of that quantity, chapter 4.4. This section will show how expectation values may often be found without finding the energy eigenfunctions. Some applications will be indicated.

The Schrödinger equation requires that the expectation value $\langle a \rangle$ of any physical quantity a with associated operator A evolves in time as:

$$\boxed{\frac{\mathrm{d}\langle a \rangle}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar}\langle [H, A] \rangle + \left\langle \frac{\partial A}{\partial t} \right\rangle} \tag{7.4}$$

A derivation is in {D.35}. The commutator $[H, A]$ of A with the Hamiltonian was defined in chapter 4.5 as $HA - AH$. The final term in (7.4) is usually zero, since most (simple) operators do not explicitly depend on time.

The above evolution equation for expectation values does not require the energy eigenfunctions, but it does require the commutator.

Note from (7.4) that if an operator A commutes with the Hamiltonian, i.e. $[H, A] = 0$, then the expectation value of the corresponding quantity a will not vary with time. Actually, that is just the start of it. Such a quantity has eigenfunctions that are also energy eigenfunctions, so it has the same time-conserved statistics as energy, section 7.1.4. The uncertainty, probabilities of the individual values, etcetera, do not change with time either for such a variable.

One application of equation (7.4) is the so-called "virial theorem" that relates the expectation potential and kinetic energies of energy eigenstates, {A.17}. For example, it shows that harmonic oscillator states have equal potential and kinetic energies. And that for hydrogen states, the potential energy is minus two times the kinetic energy.

Two other important applications are discussed in the next two subsections.

Key Points

- A relatively simple equation that describes the time evolution of expectation values of physical quantities exists. It is fully in terms of expectation values.
- Variables which commute with the Hamiltonian have the same time-independent statistics as energy.
- The virial theorem relates the expectation kinetic and potential energies for important systems.

7.2.1 Newtonian motion

The purpose of this section is to show that even though Newton's equations do not apply to very small systems, they are correct for macroscopic systems.

The trick is to note that for a macroscopic particle, the position and momentum are very precisely defined. Many unavoidable physical effects, such as

incident light, colliding air atoms, earlier history, etcetera, will narrow down position and momentum of a macroscopic particle to great accuracy. Heisenberg's uncertainty relationship says that they must have uncertainties big enough that $\Delta p_x \Delta x \geqslant \frac{1}{2}\hbar$, but $\hbar$ is far too small for that to be noticeable on a macroscopic scale. Normal light changes the momentum of a rocket ship in space only immeasurably little, but it is quite capable of locating it to excellent accuracy.

With little uncertainty in position and momentum, both can be approximated accurately by their expectation values. So the evolution of macroscopic systems can be obtained from the evolution equation (7.4) for expectation values given in the previous subsection. Just work out the commutator that appears in it.

Consider one-dimensional motion of a particle in a potential $V(x)$ (the three-dimensional case goes exactly the same way). The Hamiltonian H is:

$$H = \frac{\widehat{p}_x^2}{2m} + V(x)$$

where $\widehat{p}_x$ is the linear momentum operator and m the mass of the particle.

Now according to evolution equation (7.4), the expectation position $\langle x \rangle$ changes at a rate:

$$\frac{\mathrm{d}\langle x \rangle}{\mathrm{d}t} = \left\langle \frac{\mathrm{i}}{\hbar}[H, \widehat{x}] \right\rangle = \left\langle \frac{\mathrm{i}}{\hbar}\left[\frac{\widehat{p}_x^2}{2m} + V(x), \widehat{x}\right] \right\rangle \tag{7.5}$$

Recalling the properties of the commutator from chapter 4.5, $[V(x), \widehat{x}] = 0$, since multiplication commutes. Further, according to the rules for manipulation of products and the canonical commutator

$$[\widehat{p}_x^2, \widehat{x}] = \widehat{p}_x[\widehat{p}_x, \widehat{x}] + [\widehat{p}_x, \widehat{x}]\widehat{p}_x = -\widehat{p}_x[\widehat{x}, \widehat{p}_x] - [\widehat{x}, \widehat{p}_x]\widehat{p}_x = -2\mathrm{i}\hbar\widehat{p}_x$$

So the rate of change of expectation position becomes:

$$\frac{\mathrm{d}\langle x \rangle}{\mathrm{d}t} = \left\langle \frac{p_x}{m} \right\rangle \tag{7.6}$$

This is exactly the Newtonian expression for the change in position with time, because Newtonian mechanics defines p_x/m to be the velocity. However, it is in terms of expectation values.

To figure out how the expectation value of momentum varies, the commutator $[H, \widehat{p}_x]$ is needed. Now $\widehat{p}_x$ commutes, of course, with itself, but just like it does not commute with $\widehat{x}$, it does not commute with the potential energy $V(x)$. The generalized canonical commutator (4.62) says that $[V, \widehat{p}_x]$ equals $-\hbar\partial V/\mathrm{i}\partial x$. As a result, the rate of change of the expectation value of linear momentum becomes:

$$\frac{\mathrm{d}\langle p_x \rangle}{\mathrm{d}t} = \left\langle -\frac{\partial V}{\partial x} \right\rangle \tag{7.7}$$

This is Newton's second law in terms of expectation values: Newtonian mechanics defines the negative derivative of the potential energy to be the force, so the right hand side is the expectation value of the force. The left hand side is equivalent to mass times acceleration.

The fact that the expectation values satisfy the Newtonian equations is known as "Ehrenfest's theorem."

For a quantum-scale system, however, it should be cautioned that even the expectation values do not truly satisfy Newtonian equations. Newtonian equations use the force at the expectation value of position, instead of the expectation value of the force. If the force varies nonlinearly over the range of possible positions, it makes a difference.

There is a alternative formulation of quantum mechanics due to Heisenberg that is like the Ehrenfest theorem on steroids, {A.12}. Here the operators satisfy the Newtonian equations.

Key Points

- Newtonian physics is an approximate version of quantum mechanics for macroscopic systems.
- The equations of Newtonian physics apply to expectation values.

7.2.2 Energy-time uncertainty relation

The Heisenberg uncertainty relationship provides an intuitive way to understand the various weird features of quantum mechanics. The relationship says $\Delta p_x \Delta x \geqslant \frac{1}{2}\hbar$, chapter 4.5.3. Here Δp_x is the uncertainty in a component of the momentum of a particle, and Δx is the uncertainty in the corresponding component of position.

Now special relativity considers the energy E divided by the speed of light c to be much like a zeroth momentum coordinate, and ct to be much like a zeroth position coordinate, chapter 1.2.4 and 1.3.1. Making such substitutions transforms Heisenberg's relationship into the so-called "energy-time uncertainty relationship:"

$$\boxed{\Delta E \, \Delta t \geqslant \frac{1}{2}\hbar} \tag{7.8}$$

There is a difference, however. In Heisenberg's original relationship, the uncertainties in momentum and positions are mathematically well defined. In particular, they are the standard deviations in the measurable values of these quantities. The uncertainty in energy in the energy-time uncertainty relationship can be defined similarly. The problem is what to make of that "uncertainty in time" Δt. The Schrödinger equation treats time fundamentally different from space.

One way to address the problem is to look at the typical evolution time of the expectation values of quantities of interest. Using careful analytical arguments along those lines, Mandelshtam and Tamm succeeded in giving a meaningful definition of the uncertainty in time, {A.18}. Unfortunately, its usefulness is limited.

Ignore it. Careful analytical arguments are for wimps! Take out your pen and cross out "Δt." Write in "any time difference you want." Cross out "ΔE" and write in "any energy difference you want." As long as you are at it anyway, also cross out "$\geqslant$" and write in "=." This can be justified because both are mathematical symbols. And inequalities are so vague anyway. You have now obtained the popular version of the Heisenberg energy-time uncertainty equality:

$$\boxed{\text{any energy difference you want} \times \text{any time difference you want} = \tfrac{1}{2}\hbar} \quad (7.9)$$

This is an extremely powerful equation that can explain anything in quantum physics involving any two quantities that have dimensions of energy and time. Be sure, however, to only publicize the cases in which it gives the right answer.

Key Points

- The energy-time uncertainty relationship is a generalization of the Heisenberg uncertainty relationship. It relates uncertainty in energy to uncertainty in time. What uncertainty in time means is not obvious.
- If you are not a wimp, the answer to that problem is easy.

7.3 Conservation Laws and Symmetries

Physical laws like conservation of linear and angular momentum are important. For example, angular momentum was key to the solution of the hydrogen atom in chapter 4.3. More generally, conservation laws are often the central element in the explanation for how simple systems work. And conservation laws are normally the most trusted and valuable source of information about complex, poorly understood, systems like atomic nuclei.

It turns out that conservation laws are related to fundamental "symmetries" of physics. A symmetry means that you can do something that does not make a difference. For example, if you place a system of particles in empty space, far from anything that might affect it, it does not make a difference where exactly you put it. There are no preferred locations in empty space; all locations are equivalent. That symmetry leads to the law of conservation of linear momentum. A system of particles in otherwise empty space conserves its total amount of linear momentum. Similarly, if you place a system of particles in empty space, it

does not make a difference under what angle you put it. There are no preferred directions in empty space. That leads to conservation of angular momentum. See addendum {A.19} for the details.

Why is the relationship between conservation laws and symmetries important? One reason is that it allows for other conservation laws to be formulated. For example, for conduction electrons in solids all locations in the solid are not equivalent. For one, some locations are closer to nuclei than others. Therefore linear momentum of the electrons is not conserved. (The total linear momentum of the complete solid is conserved in the absence of external forces. In other words, if the solid is in otherwise empty space, it conserves its total linear momentum. But that does not really help for describing the motion of the conduction electrons.) However, if the solid is crystalline, its atomic structure is periodic. Periodicity is a symmetry too. If you shift a system of conduction electrons in the interior of the crystal over a whole number of periods, it makes no difference. That leads to a conserved quantity called "crystal momentum," {A.19}. It is important for optical applications of semiconductors.

Even in empty space there are additional symmetries that lead to important conservation laws. The most important example of all is that it does not make a difference at what time you start an experiment with a system of particles in empty space. The results will be the same. That symmetry with respect to time shift gives rise to the law of conservation of energy, maybe the most important conservation law in physics.

In a sense, time-shift symmetry is already "built-in" into the Schrödinger equation. The equation does not depend on what time you take to be zero. Any solution of the equation can be shifted in time, assuming a Hamiltonian that does not depend explicitly on time. So it is not really surprising that energy conservation came rolling out of the Schrödinger equation so easily in section 7.1.3. The time shift symmetry is also evident in the fact that states of definite energy are stationary states, section 7.1.4. They change only trivially in time shifts. Despite all that, the symmetry of nature with respect to time shifts is a bit less self-evident than that with respect to spatial shifts, {A.19}.

As a second example of an additional symmetry in empty space, physics works, normally, the same when seen in the mirror. That leads to a very useful conserved quantity called "parity." Parity is somewhat different from momentum. While a component of linear or angular momentum can have any value, parity can only be 1, called "even," or -1, called "odd." Also, while the contributions of the parts of a system to the total momentum components *add* together, their contributions to parity *multiply* together, {A.19}. That is why the ± 1 parities of the parts of a system can combine together into a corresponding system parity that is still either 1 or -1.

(Of course, there can be uncertainty in parity just like there can be uncertainty in other quantities. But the measurable values are either 1 or -1.)

Despite having only two possible values, parity is still very important. In the

emission and absorption of electromagnetic radiation by atoms and molecules, parity conservation provides a very strong restriction on which electronic transitions are possible. And in nuclear physics, it greatly restricts what nuclear decays and nuclear reactions are possible.

Another reason why the relation between conservation laws and symmetries is important is for the information that it produces about physical properties. For example, consider a nucleus that has zero net angular momentum. Because of the relationship between angular momentum and angular symmetry, such a nucleus looks the same from all directions. It is spherically symmetric. Therefore such a nucleus does not respond in magnetic resonance imaging. That can be said without knowing all the complicated details of the motion of the protons and neutrons inside the nucleus. So-called spin ½ nuclei have the smallest possible nonzero net angular momentum allowed by quantum mechanics, with components that can be $\pm\frac{1}{2}\hbar$ in a given direction. These nuclei do respond in magnetic resonance imaging. But because they depend in a relatively simple way on the direction from which they are viewed, their response is relatively simple.

Similar observations apply for complete atoms. The hydrogen atom is spherically symmetric in its ground state, figure 4.9. Although that result was derived ignoring the motion and spin of the proton and the spin of the electron, the hydrogen atom remains spherically symmetric even if these effects are included. Similarly, the normal helium atom, with two electrons, two protons, and two neutrons, is spherically symmetric in its ground state. That is very useful information if you want, say, an ideal gas that is easy to analyze. For heavier noble gases, the spherical symmetry is related to the "Ramsauer effect" that makes the atoms almost completely transparent to electrons of a certain wave length.

As you may guess from the fact that energy eigenstates are stationary, conserved quantities normally have definite values in energy eigenstates, {A.19.3}. (An exception may occur when the energy does not depend on the value of the conserved quantity.) For example, nuclei, lone atoms, and lone molecules normally have definite net angular momentum and parity in their ground state. Excited states too will have definite angular momentum and parity, although the values may be different from the ground state.

It is also possible to derive physical properties of particles from their symmetry properties. As an example, addendum {A.20} derives the spin and parity of an important class of particles, including photons, that way.

Finally, the relation between conservation laws and symmetries gives more confidence in the conservation laws. For example, as mentioned nuclei are still poorly understood. It might therefore seem reasonable enough to conjecture that maybe the nuclear forces do not conserve angular momentum. And indeed, the force between the proton and neutron in a deuteron nucleus does not conserve orbital angular momentum. But it is quite another matter to suppose that the forces do not conserve the net angular momentum of the nucleus, in-

cluding the spins of the proton and neutron. That would imply that empty space has some inherent preferred direction. That is much harder to swallow. Such a preferred direction has never been observed, and there is no known mechanism or cause that would give rise to it. So physicists are in fact quite confident that nuclei do conserve angular momentum just like everything else does. The deuteron conserves its net angular momentum if you include the proton and neutron spins in the total.

It goes both ways. If there is unambiguous evidence that a supposedly conserved quantity is not truly conserved, then nature does not have the corresponding symmetry. That says something important about nature. This happened for the mirror symmetry of nature. If you look at a person in a mirror, the heart is on the other side of the chest. On a smaller scale, the molecules that make up the person change in their mirror images. But physically the person in the mirror would function just fine. (As long as you do not try to mix mirror images of biological molecules with nonmirror images, that is.) In principle, evolution could have created the mirror image of the biological systems that exist today. Maybe it did on a different planet. The electromagnetic forces that govern the mechanics of biological systems obey the exact same laws when nature is seen in the mirror. So does the force of gravity that keeps the systems on earth. And so does the so-called strong force that keeps the atomic nuclei together.

Therefore it was long believed that nature behaved in exactly the same way when seen in the mirror. That then leads to the conserved quantity called parity. But eventually, in 1956, Lee and Yang realized that the decay of a certain nuclear particle by means of the so-called weak force does not conserve parity. As a result, it had to be accepted also that nature does not always behave in exactly the same way when seen in the mirror. That was confirmed experimentally by Wu and her coworkers in 1957. (In fact, while other experimentalists like Lederman laughed at the ideas of Lee and Yang, Wu spend eight months of hard work on the risky proposition of confirming them. If she had been a man, she would have been given the Nobel Prize along with Lee and Yang. However, Nobel Prize committees have usually recognized that giving Nobel Prizes to women might interfere with their domestic duties.)

Fortunately, the weak force is not important for most applications, not even for many involving nuclei. Therefore conservation of parity usually remains valid to an excellent approximation.

Mirroring corresponds mathematically to an inversion of a spatial coordinate. But it is mathematically much cleaner to invert the direction of all three coordinates, replacing every position vector $\vec{r}$ by $-\vec{r}$. That is called "spatial inversion." Spatial inversion is cleaner since no choice of mirror is required. That is why many physicists reserve the term "parity transformation" exclusively to spatial inversion. (Mathematicians do not, since inversion does not work in strictly two-dimensional systems, {A.19}.) A normal mirroring is equivalent to spatial inversion followed by a rotation of 180° around the axis normal to the

chosen mirror.

Inversion of the time coordinate is called "time reversal." That can be thought of as making a movie of a physical process and playing the movie back in reverse. Now if you make a movie of a macroscopic process and play it backwards, the physics will definitely not be right. However, it used to be generally believed that if you made a movie of the microscopic physics and played it backwards, it would look fine. The difference is really not well understood. But presumably it is related to that evil demon of quantum mechanics, the collapse of the wave function, and its equally evil macroscopic alter ego, called the second law of thermodynamics. In any case, as you might guess it is somewhat academic. If physics is not completely symmetric under reversal of a spatial coordinate, why would it be under reversal of time? Special relativity has shown the close relationship between spatial and time coordinates. And indeed, it was found that nature is not completely symmetric under time reversal either, even on a microscopic scale.

There is a third symmetry involved in this story of inversion. It involves replacing every particle in a system by the corresponding antiparticle. For every elementary particle, there is a corresponding antiparticle that is its exact opposite. For example, the electron, with electric charge $-e$ and lepton number 1, has an antiparticle, the positron, with charge e and lepton number -1. (Lepton number is a conserved quantity much like charge is.) Bring an electron and a positron together, and they can totally annihilate each other, producing two photons. The net charge was zero, and is still zero. Photons have no charge. The net lepton number was zero, and is still zero. Photons are not leptons and have zero lepton number.

All particles have antiparticles. Protons have antiprotons, neutrons antineutrons, etcetera. Replacing every particle in a system by its antiparticle produces almost the same physics. You can create an antihydrogen atom out of an antiproton and a positron that seems to behave just like a normal hydrogen atom does.

Replacing every particle by its antiparticle is not called particle inversion, as you might think, but "charge conjugation." That is because physicists recognized that "charge inversion" would be all wrong; a lot more changes than just the charge. And the particle involved might not even have a charge, like the neutron, with no net charge but a baryon number that inverts, or the neutrino, with no charge but a lepton number that inverts. So physicists figured that if "charge inversion" is wrong anyway, you may as well replace "inversion" by "conjugation." That is not the same as inversion, but it was wrong anyway, and conjugation sounds much more sophisticated and it alliterates.

The bottom line is that physics is almost, but not fully, symmetric under spatial inversion, time inversion, and particle inversion. However, physicists currently believe that if you apply all three of these operations together, then the resulting physics is indeed truly the same. There is a theorem called the CPT

theorem, (charge, parity, time), that says so under relatively mild assumptions. One way to look at it is to say that systems of antiparticles are the mirror images of systems of normal particles that move backwards in time.

At the time of writing, there is a lot of interest in the possibility that nature may in fact not be exactly the same when the CPT transformations are applied. It is hoped that this may explain why nature ended up consisting almost exclusively of particles, rather than antiparticles.

Symmetry transformations like the ones discussed above form mathematical "groups." There are infinitely many different angles that you can rotate a system over or distances that you can translate it over. What is mathematically particularly interesting is how group members combine together into different group members. For example, a rotation followed by another rotation is equivalent to a single rotation over a combined angle. You can even eliminate a rotation by following it by one in the opposite direction. All that is nectar to mathematicians.

The inversion transformations are somewhat different in that they form finite groups. You can either invert or not invert. These finite groups provide much less detailed constraints on the physics. Parity can only be 1 or -1. On the other hand, a component of linear or angular momentum must maintain one specific value out of infinitely many possibilities. But even these constraints remain restricted to the total system. It is the complete system that must maintain the same linear and angular momentum, not the individual parts of it. That reflects that the same rotation angle or translation distance applies for all parts of the system.

Advanced relativistic theories of quantum mechanics postulate symmetries that apply on a local (point by point) basis. A simple example relevant to quantum electrodynamics can be found in addendum {A.19}. Such symmetries narrow down what the physics can do much more because they involve separate parameters at each individual point. Combined with the massive antisymmetrization requirements for fermions, they allow the physics to be deduced in terms of a few remaining numerical parameters. The so-called "standard model" of relativistic quantum mechanics postulates a combination of three symmetries of the form

$$\mathrm{U}(1) \times \mathrm{SU}(2) \times \mathrm{SU}(3)$$

In terms of linear algebra, these are complex matrices that describe rotations of complex vectors in 1, 2, respectively 3 dimensions. The "S" on the latter two matrices indicates that they are special in the sense that their determinant is 1. The first matrix is characterized by 1 parameter, the angle that the single complex numbers are rotated over. It gives rise to the photon that is the single carrier of the electromagnetic force. The second matrix has 3 parameters, corresponding to the 3 so-called "vector bosons" that are the carriers of the weak nuclear force. The third matrix has 8 parameters, corresponding to the 8

"gluons" that are the carriers of the strong nuclear force.

There is an entire branch of mathematics, "group theory," devoted to how group properties relate to the solutions of equations. It is essential to advanced quantum mechanics, but far beyond the scope of this book.

Key Points

- Symmetries of physics give rise to conserved quantities.
- These are of particular interest in obtaining an understanding of complicated and relativistic systems. They can also aid in the solution of simple systems.
- Translational symmetry gives rise to conservation of linear momentum. Rotational symmetry gives rise to conservation of angular momentum.
- Spatial inversion replaces every position vector $\vec{r}$ by $-\vec{r}$. It produces a conserved quantity called parity.
- There are kinks in the armor of the symmetries under spatial inversion, time reversal, and "charge conjugation." However, it is believed that nature is symmetric under the combination of all three.

7.4 Conservation Laws in Emission

Conservation laws are very useful for understanding emission or absorption of radiation of various kinds, as well as nuclear reactions, collision processes, etcetera. As an example, this section will examine what conservation laws say about the spontaneous emission of a photon of light by an excited atom. While this example is relatively simple, the concepts discussed here apply in essentially the same way to more complex systems.

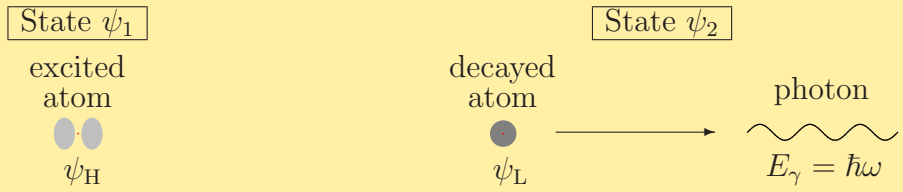


Figure 7.3: Crude concept sketch of the emission of an electromagnetic photon by an atom. The initial state is left and the final state is right.

Figure 7.3 gives a sketch of the emission process. The atom is initially in an high energy, or excited, state that will be called ψ_H. After some time, the atom releases a photon and returns a lower energy state that will be called ψ_L. As a simple example, take an hydrogen atom. Then the excited atomic state

could be the "$2p_z$" ψ_{210} state, chapter 4.3. The final atomic state will then be the "1s" ground state ψ_{100}.

The emitted photon has an energy given by the Planck-Einstein relation

$$E_\gamma = \hbar\omega$$

where ω is the frequency of the electromagnetic radiation corresponding to the photon. Note that γ (gamma, think gamma decay) is the standard symbol used to indicate a photon.

Key Points

- Atoms can transition to a lower electronic energy level while emitting a photon of electromagnetic radiation.
- The Planck-Einstein relation gives the energy of a photon in terms of its frequency.

7.4.1 Conservation of energy

The first conservation law that is very useful for understanding the emission process is conservation of energy. The final atom and photon should have the exact same energy as the initial excited atom. So the difference between the atomic energies $E_{\rm H}$ and $E_{\rm L}$ must be the energy $\hbar\omega$ of the photon. Therefore, the emitted photon must have a very precise frequency ω. That means that it has a very precise color. For example, for the $2p_z$ to 1s transition of a hydrogen atom, the emitted photon is a very specific ultraviolet color.

It should be pointed out that the frequency of the emitted photon does have a very slight variation. The reason can be understood from the fact that the excited state decays at all. Energy eigenstates should be stationary, section 7.1.4.

> *The very fact that a state decays shows that it is not truly an energy eigenstate.*

The big problem with the analysis of the hydrogen atom in chapter 4.3 was that it ignored any ambient radiation that the electron might be exposed to. It turns out that there is always some perturbing ambient radiation, even if the atom is inside a black box at absolute zero temperature. This is related to the fact that the electromagnetic field has quantum uncertainty. Advanced quantum analysis is needed to take that into account, {A.23}. Fortunately, the uncertainty in energy is extremely small for the typical applications considered here.

As a measure of the uncertainty in energy of a state, physicists often use the so-called "natural width"

$$\boxed{\Gamma = \frac{\hbar}{\tau}} \tag{7.10}$$

Here τ is the mean lifetime of the state, the average time it takes for the photon to be emitted.

The claim that this width gives the uncertainty in energy of the state is usually justified using the all-powerful energy-time uncertainty equality (7.9). A different argument will be given at the end of section 7.6.1. In any case, the bottom line is that Γ does indeed give the observed uncertainty in energy for isolated atoms, [52, p. 139], and for nuclei, [31, p. 40, 167].

As an example, the hydrogen atom $2p_z$ state has a lifetime of 1.6 nanoseconds. (The lifetime can be computed using the data in addendum {A.25.8}.) That makes its width about 4 10^{-7} eV. Compared to the 10 eV energy of the emitted photon, that is obviously extremely small. Energy conservation in atomic transitions may not be truly exact, but it is definitely an excellent approximation.

Still, since a small range of frequencies can be emitted, the observed line in the emission spectrum is not going to be a mathematically exact line, but will have a small width. Such an effect is known as "spectral line broadening."

The natural width of a state is usually only a small part of the actual line broadening. If the atom is exposed to an incoherent ambient electromagnetic field, it will increase the uncertainty in energy. (The evolution of atoms in an incoherent electromagnetic field will be analyzed in {D.41}.) Frequent interactions with surrounding atoms or other perturbations will also increase the uncertainty in energy, in part for reasons discussed at the end of section 7.6.1. And anything else that changes the atomic energy levels will of course also change the emitted frequencies.

An important further effect that causes spectral line deviations is atom motion, either thermal motion or global gas motion. It produces a Doppler shift in the radiation. This is not necessarily bad news in astronomy; line broadening can provide an hint about the temperature of the gas you are looking at, while line displacement can provide a hint of its overall motion away from you.

It may also be mentioned that the natural width is not always small. If you start looking at excited nuclear particles, the uncertainty in energy can be enormous. Such particles may have an uncertainty in energy that is of the order of 10% of their relativistic rest mass energy. And as you might therefore guess, they are hardly stationary states. Typically, they survive for only about 10^{-23} seconds after they are created. Even moving at a speed comparable to the speed of light, such particles will travel only a distance comparable to the diameter of a proton before disintegrating.

Generally speaking, the shorter the lifetime of a state, the larger its uncertainty in energy, and vice-versa.

(To be fair, physicists do not actually manage to see these particles during their infinitesimal lifetime. Instead they infer the lifetime from the variation in energy of the resulting state.)

Key Points

- In a transition, the difference in atomic energy levels gives the energy, and so the frequency, of the emitted photon.
- Unstable states have some uncertainty in energy, but it is usually very small. For extremely unstable particles, the uncertainty can be a lot.
- The width of a state is $\Gamma = \hbar/\tau$ with τ the mean lifetime. It is a measure for the minimum observed variation in energy of the final state.

7.4.2 Combining angular momenta and parities

Conservation of angular momentum and parity is easily stated:

The angular momentum and parity of the initial atomic state must be the same as the combined angular momentum and parity of the final atomic state and photon.

The question is however, how do you combine angular momenta and parity values? Even combining angular momenta is not trivial, because angular momenta are quantized.

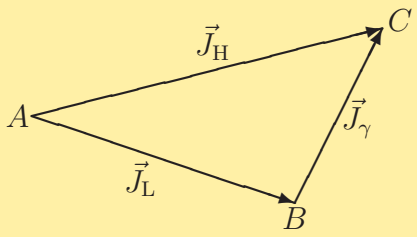


Figure 7.4: Addition of angular momenta in classical physics.

To get an idea of how angular momenta combine, first consider what would happen in classical physics. Conservation of angular momentum would say that

$$\vec{J}_{\rm H} = \vec{J}_{\rm L} + \vec{J}_\gamma$$



Figure 7.5: Longest and shortest possible final atomic angular momenta in classical physics.

Here $\vec{J}_{\rm H}$ is the angular momentum vector of the initial high energy atomic state, and $\vec{J}_{\rm L}$ and $\vec{J}_\gamma$ are those of the final low energy atomic state and the emitted photon. The conservation law is shown graphically in figure 7.4.

Now consider what possible lengths the vector $\vec{J}_{\rm L}$ of the final atomic state can have. As figure 7.4 shows, the length of $\vec{J}_{\rm L}$ is the distance between the starting points of the other two vectors. So the maximum length occurs when the two vectors point in opposite direction, with their noses touching, like to the left in figure 7.5. In that case, the length of $\vec{J}_{\rm L}$ is the sum of the lengths of the other two vectors. The minimum length for $\vec{J}_{\rm L}$ occurs when the other two vectors are in the same direction, still pointing at the same point, like to the right in figure 7.5. In that case the length of $\vec{J}_{\rm L}$ is the difference in length between the other two vectors.

All together:

$$\text{classical physics:} \qquad |J_{\rm H} - J_\gamma| \leqslant J_{\rm L} \leqslant J_{\rm H} + J_\gamma$$

Note that the omission of a vector symbol indicates that the length of the vector is meant, rather than the vector itself. The second inequality is the famous "triangle inequality." (The first inequality is a rewritten triangle inequality for the longer of the two vectors in the absolute value.) The bottom line is that according to classical physics, the length of the final atomic angular momentum can take any value in the range given above.

However, in quantum mechanics angular momentum is quantized. The length of an angular momentum vector $\vec{J}$ must be $\sqrt{j(j+1)}\hbar$. Here the "azimuthal quantum number" j must be a nonnegative integer or half of one. Fortunately, the triangle inequality above still works if you replace lengths by azimuthal quantum numbers. To be precise, the possible values of the final atomic angular momentum quantum number are:

$$\boxed{j_{\rm L} = |j_{\rm H}{-}j_\gamma|,\ |j_{\rm H}{-}j_\gamma|+1,\ \ldots,\ j_{\rm H}{+}j_\gamma{-}1,\text{ or } j_{\rm H}{+}j_\gamma} \tag{7.11}$$

In other words, the possible values of $j_{\rm L}$ increase from $|j_{\rm H}{-}j_\gamma|$ to $j_{\rm H}{+}j_\gamma$ in steps of 1. To show that angular momentum quantum numbers satisfy the triangle inequality in this way is not trivial; that is a major topic of chapter 12.

Classical physics also says that components of vectors can be added and subtracted as ordinary numbers. Quantum physics agrees, but adds that for nonzero angular momentum only one component can be certain at a time. That

is usually taken to be the z-component. Also, the component cannot have any arbitrary value; it must have a value of the form $m\hbar$. Here the "magnetic quantum number" m can only have values that range from $-j$ to j in increments of 1.

If you can add and subtract components of angular momentum, then you can also add and subtract magnetic quantum numbers. After all, they are only different from components by a factor $\hbar$. Therefore, the conservation of angular momentum in the z-direction becomes

$$m_{\rm L} = m_{\rm H} - m_\gamma$$

Putting in the possible values of the magnetic quantum number of the photon gives for the final atomic magnetic quantum number:

$$\boxed{m_{\rm L} = m_{\rm H}{-}j_\gamma,\ m_{\rm H}{-}j_\gamma{+}1,\ \ldots,\ m_{\rm H}{+}j_\gamma{-}1,\ \text{or}\ m_{\rm H}{+}j_\gamma} \tag{7.12}$$

To be sure, $m_{\rm L}$ is also constrained by the fact that its magnitude cannot exceed $j_{\rm L}$.

Next consider conservation of parity. Recall from section 7.3 that parity is the factor by which the wave function changes when the positive direction of all three coordinate axes is inverted. That replaces every position vector $\vec{r}$ by $-\vec{r}$. Parity can have only two values, 1 or -1. Parity is commonly indicated by π, which is the Greek letter for "p." Parity starts with a p and may well be Greek. Also, the symbol avoids confusion, assuming that π is not yet used for anything else in science.

Conservation of parity means that the initial and final parities must be equal. The parity of the initial high energy atom must be the same as the combined parity of the final low energy atom and photon:

$$\pi_{\rm H} = \pi_{\rm L}\pi_\gamma$$

Note that parity is a multiplicative quantity. You get the combined parity of the final state by multiplying the parities of atom and photon; you do not add them.

(Just think of the simplest possible wave function of two particles, $\Psi = \psi_1(\vec{r}_1)\psi_2(\vec{r}_2)$. If ψ_1 changes by a factor π_1 when $\vec{r}_1 \to -\vec{r}_1$ and ψ_2 changes by a factor π_2 when $\vec{r}_2 \to -\vec{r}_2$, then the total wave function Ψ changes by a factor $\pi_1\pi_2$. Actually, it is angular momentum, not parity, that is the weird case. The reason that angular momenta must be added together instead of multiplied together is because angular momentum is defined by taking a logarithm of the "natural" conserved quantity. For details, see addendum {A.19}.)

The parity of the atom is related to the orbital angular momentum of the electron, and in particular to its azimuthal quantum number l. If you check out the example spherical harmonics in table 4.3, you see that those with even

values of l only contain terms that are square in the position coordinates. So these states do not change when $\vec{r}$ is replaced by $-\vec{r}$. In other words, they change by a trivial factor 1. That makes the parity 1, or even, or positive. The spherical harmonics for odd l change sign when $\vec{r}$ is replaced by $-\vec{r}$. In other words, they get multiplied by a factor -1. That makes the parity -1, or odd, or negative. These observations apply for all values of l, {D.14}.

The parity can therefore be written for any value of l as

$$\boxed{\pi = (-1)^l} \tag{7.13}$$

This is just the parity due to orbital angular momentum. If the particle has negative intrinsic parity, you need to multiply by another factor -1. However, an electron has positive parity, as does a proton. (Positrons and antiprotons have negative parity. That is partly a matter of convention. Conservation of parity would still work if it was the other way around.)

It follows that parity conservation in the emission process can be written as

$$\boxed{(-1)^{l_{\rm H}} = (-1)^{l_{\rm L}} \pi_\gamma} \tag{7.14}$$

Therefore, if the parity of the photon is even, (i.e. 1), then $l_{\rm H}$ and $l_{\rm L}$ are both even or both odd. In other words, the atomic parity stays unchanged. If the parity of the photon is odd, (i.e. -1), then one of $l_{\rm H}$ and $l_{\rm L}$ is even and the other odd. The atomic parity flips over.

To apply the obtained conservation laws, the next step must be to figure out the angular momentum and parity of the photon.

Key Points

- The rules for combining angular momenta and parities were discussed.
- Angular momentum and parity conservation lead to constraints on the atomic emission process given by (7.11), (7.12), and (7.14).

7.4.3 Transition types and their photons

The conservation laws of angular momentum and parity restrict the emission of a photon by an excited atom. But for these laws to be useful, there must be information about the spin and parity of the photon.

This section will just state various needed photon properties. Derivations are given in {A.21.7} for the brave. In any case, the main conclusions reached about the photons associated with atomic transitions will be verified by more detailed analysis of transitions in later sections.

There are two types of transitions, electric ones and magnetic ones. In electric transitions, the electromagnetic field produced by the photon at the atom is primarily electric, {A.21.7}. In magnetic transitions, it is primarily magnetic. Electric transitions are easiest to understand physically, and will be discussed first.

A photon is a particle with spin $s_\gamma = 1$ and intrinsic parity -1. Also, assuming that the size of the atom is negligible, in the simplest model the photon will have zero orbital angular momentum around the center of the atom. That is most easily understood using classical physics: a particle that travels along a line that comes out of a point has zero angular momentum around that point. For equivalent quantum arguments, see {N.10} or {A.21.7}. It means in terms of quantum mechanics that the photon has a quantum number l_γ of orbital angular momentum that is zero. That makes the total angular momentum quantum number j_γ of the photon equal to the spin s_γ, 1.

The normal, efficient kind of atomic transition does in fact produce a photon like that. Since the term "normal" is too normal, such a transition is called "allowed." For reasons that will eventually be excused for in section 7.7.2, allowed transitions are also more technically called "electric dipole" transitions. According to the above, then, the photon net angular momentum and parity are:

$$\boxed{\text{for electric dipole transitions:} \qquad j_\gamma = 1 \qquad \pi_\gamma = -1} \tag{7.15}$$

Transitions that cannot happen according to the electric dipole mechanism are called "forbidden." That does not mean that these transitions cannot occur at all; just forbid your kids something. But they are much more awkward, and therefore normally very much slower, than allowed transitions.

One important case of a forbidden transition is one in which the atomic angular momentum changes by 2 or more units. Since the photon has only 1 unit of spin, in such a transition the photon must have nonzero orbital angular momentum. Transitions in which the photon has more than 1 unit of net angular momentum are called "multipole transitions." For example, in a "quadrupole" transition, the net angular momentum of the photon $j_\gamma = 2$. In an "octupole" transition, $j_\gamma = 3$ etcetera. In all these transitions, the photon has at least $j_\gamma - 1$ units of orbital angular momentum.

To roughly understand how orbital angular momentum arises, reconsider the sketch of the emission process in figure 7.3. As shown, the photon has no orbital angular momentum around the center of the atom, classically speaking. But the photon does not have to come from exactly the center of the atom. If the atom has a typical radius R, then the photon could come from a point at a distance comparable to R away from the center. That will give it an orbital angular momentum of order Rp around the center, where p is the linear momentum of the photon. And according to relativity, (1.2), the photon's momentum is related to its energy, which is in turn related to its frequency by the Planck-

Einstein relation. That makes the classical orbital angular momentum of the photon of order $R\hbar\omega/c$. But c/ω is the wave length λ of the photon, within a factor 2π. That factor is not important for a rough estimate. So the typical classical orbital angular momentum of the photon is

$$L \sim \frac{R}{\lambda}\hbar$$

The fraction is typically small. For example, the wave length λ of visible light is about 5 000 Å and the size R of an atom is about an Å. So the orbital angular momentum above is a very small fraction of $\hbar$.

But according to quantum mechanics, the orbital angular momentum *cannot* be a small fraction of $\hbar$. If the quantum number l_γ is zero, then so is the orbital angular momentum. And if l_γ is 1, then the orbital angular momentum is $\sqrt{2}\hbar$. There is nothing in between. The above classical orbital angular momentum should be understood to mean that there is quantum uncertainty in orbital angular momentum. That the photon has almost certainly zero orbital angular momentum, but that there remains a small probability of $l_\gamma = 1$. In particular, if you take the ratio R/λ to be the coefficient of the $l_\gamma = 1$ state, then the probability of the photon coming out with $l_\gamma = 1$ is the square of that, $(R/\lambda)^2$. That will be a very small probability. But still, there is a slight probability that the net photon angular momentum j_γ will be increased from 1 to 2 by a unit's worth of orbital angular momentum. That will then produce a quadrupole transition. And of course, two units of orbital angular momentum can increase the net photon angular momentum to $j_\gamma = 3$, the octupole level. But that reduces the probability by another factor $(R\lambda)^2$, so don't hold you breath for these higher order multipole transitions to occur.

(If the above random mixture of unjustified classical and quantum arguments is too unconvincing, there is a quantum argument in {N.10} that may be more believable. If you are brave, see {A.21.7} for a precise analysis of the relevant photon momenta and their probabilities in an interaction with an atom or nucleus. But the bottom line is that the above ideas do describe what happens in transition processes. That follows from a complete analysis of the transition process, as discussed in later sections and notes like {A.25} and {D.39}.)

So far, only electric multipole transitions have been discussed, in which the electromagnetic field at the atom is primarily electric. In magnetic multipole transitions however, it is primarily magnetic. In a "magnetic dipole" transition, the photon comes out with one unit of net angular momentum just like in an electric dipole one. However, the parity of the photon is now even:

$$\boxed{\text{for magnetic dipole transitions:} \qquad j_\gamma = 1 \qquad \pi_\gamma = 1} \tag{7.16}$$

You might wonder how the positive parity is possible if the photon has negative intrinsic parity and no orbital angular momentum. The reason is that

in a magnetic dipole transition, the photon does have a unit of orbital angular momentum. Recall from the previous subsection that it is quite possible for one unit of spin and one unit of orbital angular momentum to combine into still only one unit of net angular momentum.

In view of the crude discussion of orbital angular momentum given above, this may still seem weird. How come that an atom of vanishing size does suddenly manage to readily produce a unit of orbital angular momentum in a magnetic dipole transition? The basic reason is that the magnetic field acts in some way as if it has one unit of orbital angular momentum less than the photon, {A.21.7}. It is unpexpectedly strong at the atom. This allows a magnetic atom state to "get a solid grip" on a photon state of unit orbital angular momentum. It is somewhat like hitting a rapidly spinning ball with a bat in baseball; the resulting motion of the ball can be weird. And in a sense the orbital angular momentum comes at the expense of the spin; the net angular momentum j_γ of a photon in a magnetic dipole transition will not be 2 despite the orbital angular momentum.

Certainly this sort of complications would not arise if the photon had no spin. Without discussion, the photon is one of the most basic particles in physics. But it is surprisingly complex for such an elementary particle. This also seems the right place to confess to the fact that electric multipole photons have uncertainty in orbital angular momentum. For example, an electric dipole photon has a probability for $l_\gamma = 2$ in addition to $l_\gamma = 0$. However, this additional orbital angular momentum comes courtesy of the internal mechanics, and especially the spin, of the photon. It does *not* give the photon a probability for net angular momentum $j_\gamma = 2$. So it does not really change the given discussion.

All else being the same, the probability of a magnetic dipole transition is normally much smaller than an electric dipole one. The principal reason is that the magnetic field is really a relativistic effect. That can be understood, for example, from how the magnetic field popped up in the description of the relativistic motion of charged particles, chapter 1.3.2. So you would expect the effect of the magnetic field to be minor unless the atomic electron or nucleon involved in the transition has a kinetic energy comparable to its rest mass energy. Indeed, it turns out that the probability of a magnetic transition is smaller than an electric one by a factor of order T/mc^2, where T is the kinetic energy of the particle and mc^2 its rest mass energy, {A.25.4}. For the electron in a hydrogen atom, and for the outer electrons in atoms in general, this ratio is very much less than one. The same holds for the nucleons in nuclei. It follows that magnetic dipole transitions will normally take place much slower than electric dipole ones.

In magnetic multipole transitions, the photon receives additional angular momentum. Like for electric multipole transitions, there is one additional unit of angular momentum for each additional multipole order. And there is a corresponding slow down of the transitions.

Table 7.1 gives a summary of the photon properties in multipole transitions.

	j_γ	π_γ	slow down
Eℓ :	ℓ	$(-1)^\ell$	$(R/\lambda)^{2\ell-2}$
Mℓ :	ℓ	$(-1)^{\ell-1}$	$(T/mc^2)(R/\lambda)^{2\ell-2}$
photons do not have zero net angular momentum; $\ell = j_\gamma \geqslant 1$			

Table 7.1: Properties of photons emitted in electric and magnetic multipole transitions.

It is conventional to write electric multipole transitions as Eℓ and magnetic ones as Mℓ where ℓ, (or L, but never j), is the net photon angular momentum j_γ. So an electric dipole transition is E1 and a magnetic dipole one M1. In agreement with the previous section, each unit increase in the orbital angular momentum produces an additional factor -1 in parity.

The column "slow down" gives an order of magnitude estimate by what factor a transition is slower than an electric dipole one, all else being equal. Note however that all else is definitely not equal, so these factors should not be used even for ballparks.

There are some official ballparks for atomic nuclei based on a more detailed analysis. These are called the Weisskopf and Moszkowski estimates, chapter 14.20.4 and in particular addendum {A.25.8}. But even there you should not be surprised if the ballpark is off by orders of magnitude. These estimates do happen to work fine for the nonrelativistic hydrogen atom, with appropriate adjustments, {A.25.8}.

The slow down factors T/mc^2 and $(R/\lambda)^2$ are often quite comparable. That makes the order of slow down of magnetic dipole transitions similar to that of electric quadrupole transitions. To see the equivalence of the slow-down factors, rewrite them as

$$\left(\frac{R}{\lambda}\right)^2 = \frac{1}{\hbar^2 c^2} R^2 (\hbar\omega)^2 \qquad \Longleftrightarrow \qquad \frac{T}{mc^2} = \frac{1}{\hbar^2 c^2} R^2 T \frac{\hbar^2}{mR^2}$$

where the 2π in the wave length was again put back. For an atom, the energy of the emitted photon $\hbar\omega$ is often comparable to the kinetic energy T of the outer electrons, and the final ratio in the equations above is a rough estimate for that kinetic energy. It follows that the two slow down factors are comparable. Another way of looking at the similarity between magnetic dipole and electric quadrupole transitions will be given in {D.39}.

Note from the table that electric quadrupole and magnetic dipole transitions have the same parity. That means that they may compete directly with each

other on the same transition, provided that the atomic angular momentum does not change more than one unit in that transition.

For nuclei, the photon energy tends to be significantly less than the nucleon kinetic energy. That is one reason that the Weisskopf estimates have the electric quadrupole transitions a lot slower than magnetic dipole ones for typical transitions. Also note that the kinetic energy estimate above does not include the effect of the exclusion principle. Exclusion raises the true kinetic energy if there are multiple identical particles in a given volume.

There is another issue that should be mentioned here. Magnetic transitions have a tendency to underperform for simple systems like the hydrogen atom. For these systems, the magnetic field has difficulty making effective use of spin in changing the atomic or nuclear structure. That is discussed in more detail in the next subsection.

One very important additional property must still be mentioned. The photon cannot have zero net angular momentum. Normally it is certainly possible for a particle with spin $s = 1$ and orbital angular momentum quantum number $l = 1$ to be in a state that has zero net angular momentum, $j = 0$. However, a photon is not a normal particle; it is a relativistic particle with zero rest mass that can only move at the speed of light. It turns out that for a photon, spin and orbital angular momentum are not independent, but intrinsically linked. This limitation prevents a state where the photon has zero net angular momentum, {A.21.3}.

There are some effects in classical physics that are related to this limitation. First of all, consider a photon with definite linear momentum. That corresponds to a light wave propagating in a particular direction. Now linear and angular momentum do not commute, so such a photon will not have definite angular momentum. However, the angular momentum component in the direction of motion is still well defined. The limitation on photons is in this case that the photon must either have angular momentum $\hbar$ or $-\hbar$ along the direction of motion. A normal particle of spin 1 could also have zero angular momentum in the direction of motion, but a photon cannot. The two states of definite angular momentum in the direction of motion are called "right- and left-circularly polarized" light, respectively.

Second, for the same type of photon, there are two equivalent states that have definite directions of the electric and magnetic fields. These states have uncertainty in angular momentum in the direction of motion. They are called "linearly polarized" light. These states illustrate that there cannot be an electric or magnetic field component in the direction of motion. The electric and magnetic fields are normal to the direction of motion, and to each other.

More general photons of definite linear momentum may have uncertainty in both of the mentioned properties. But still there is zero probability for zero angular momentum in the direction of motion, and zero probability for a field in the direction of motion.

Third, directly related to the previous case. Suppose you have a charge distribution that is spherically symmetric, but pulsating in the radial direction. You would expect that you would get a fluctuating radial electrical field outside this pulsating charge. But you do not, it does not radiate energy. Such radiation would have the electric field in the direction of motion, and that does not happen. Now consider the transition from the spherically symmetric "2s" state of a hydrogen atom to the spherical symmetric "1s" state. Because of the lack of spherically symmetric radiation, you might guess that this transition is in trouble. And it is; that is discussed in the next subsection.

In fact, the last example is directly related to the missing state of zero angular momentum of the photon. Recall from section 7.3 that angular momentum is related to angular symmetry. In particular, a state of zero angular momentum (if exact to quantum accuracy) looks the same when seen from all directions. The fact that there is no spherically symmetric radiation is then just another way of saying that the photon cannot have zero angular momentum.

Key Points

- Normal atomic transitions are called allowed or electric dipole ones. All others are called forbidden but can occur just fine.
- In electric dipole transitions the emitted photon has angular momentum quantum number $j_\gamma = 1$ and negative parity $\pi_\gamma = -1$.
- In the slower magnetic dipole transitions the photon parity is positive, $\pi_\gamma = 1$.
- Each higher multipole order adds a unit to the photon angular momentum quantum number j_γ and flips over the parity π_γ.
- The higher the multipole order, the slower the transition will be.

7.4.4 Selection rules

As discussed, a given excited atomic state may be able to transition to a lower energy state by emitting a photon. But many transitions from a higher energy state to a lower one simply do not happen. There are so-called "selection rules" that predict whether or not a given transition process is possible. This subsection gives a brief introduction to these rules.

The primary considered system will be the hydrogen atom. However, some generally valid rules are given at the end. It will usually be assumed that the effect of the spin of the electron on its motion can be ignored. That is the same approximation as used in chapter 4.3, and it is quite accurate. Basically, the model system studied is a spinless charged electron going around a stationary proton. Spin will be tacked on after the fact.

The selection rules result from the conservation laws and photon properties as discussed in the previous two subsections. Since the conservation laws are applied to a spinless electron, the angular momentum of the electron is simply its orbital angular momentum. That means that for the atomic states, the angular momentum quantum number j becomes the orbital angular momentum quantum number l. For the emitted photon, the true net angular momentum quantum number ℓ must be used.

Now suppose that the initial high-energy atomic state has an orbital angular momentum quantum number $l_{\rm H}$ and that it emits a photon with angular momentum quantum number ℓ. The question is then what can be said about the orbital angular momentum $l_{\rm L}$ of the atomic state of lower energy after the transition. The answer is given by subsection 7.4.2 (7.11):

$$l_{\rm L} = |l_{\rm H}-\ell|,\ |l_{\rm H}-\ell|+1,\ \ldots,\ l_{\rm H}+\ell-1,\ \text{or}\ l_{\rm H}+\ell \tag{7.17}$$

That leads immediately to a stunning conclusion for the decay of the hydrogen ψ_{200} "2s" state. This state has angular momentum $l_{\rm H} = 0$, as any s state. So the requirement above simplifies to $l_{\rm L} = \ell$. Now recall from the previous subsection that a photon must have ℓ at least equal to 1. So $l_{\rm L}$ must be at least 1. But $l_{\rm L}$ *cannot* be at least 1. The only lower energy state that exists is the ψ_{100} "1s" ground state. It has $l_{\rm L} = 0$. So the 2s state cannot decay!

Never say never, of course. It turns out that if left alone, the 2s state will eventually decay through the emission of two photons, rather than a single one. This takes forever on quantum scales; the 2s state survives for about a tenth of a second rather than maybe a nanosecond for a normal transition. Also, to actually observe the two-photon emission process, the atom must be in high vacuum. Otherwise the 2s state would be messed up by collisions with other particles long before it could decay. Now you see why the introduction to this section gave a 2p state, and not the seemingly more simple 2s one, as a simple example of an atomic state that decays by emitting a photon.

Based on the previous subsection, you might wonder why a second photon can succeed where a unit of photon orbital angular momentum cannot. After all, photons have only two independent spin states, while a unit of orbital angular momentum has the full set of three. The explanation is that in reality you *cannot* add a suitable unit of orbital angular momentum to a photon; the orbital and spin angular momentum of a photon are intrinsically linked. But photons do have complete sets of states with angular momentum $\ell = 1$, {A.21.7}. For two photons, these can combine into zero net angular momentum.

It is customary to "explain" photons in terms of states of definite linear momentum. That is in fact what was done in the final paragraphs of the previous subsection. But it is simplistic. It is definitely impossible to understand how two photons, each missing the state of zero angular momentum along their direction of motion, could combine into a state of zero net angular momentum. In fact, they simply cannot. Linear and orbital angular momentum do not

commute. But photons do not have to be in quantum states of definite linear momentum. They can be, and often are, in quantum superpositions of such states. The states of definite angular momentum are quantum superpositions of infinitely many states of linear momentum in all directions. To make sense out of that, you need to switch to a description in terms of photon states of definite angular, rather than linear, momentum. Those states are listed in {A.21.7}. Unfortunately, they are much more difficult to describe physically than states of definite linear momentum.

It should also be noted that if you include relativistic effects, the 2s state can actually decay to the 2p state that has net angular momentum (spin plus orbital) $\frac{1}{2}$. This 2p state has very slightly lower energy than the 2s state due to a tiny relativistic effect called "Lamb shift," {A.39.4}. But because of the negligible difference in energy, such a transition is even slower than two-photon emission. It takes over 100 years to have a 50/50 probability for the transition.

Also, including relativistic efects, a magnetic dipole transition is possible. An atomic state with net angular momentum $\frac{1}{2}$ (due to the spin) can decay to a state with again net angular momentum $\frac{1}{2}$ by emitting a photon with angular momentum $\ell = 1$, subsection 7.4.2. A magnetic M1 transition is needed in order that the parity stays the same. Unfortunately, in the nonrelativistic approximation an M1 transition does not change the orbital motion; it just flips over the spin. Also, without any energy change the theoretical transition rate will be zero, section 7.6.1.

Relativistic effects remove these obstacles. But since these effects are very small, the one-photon transition does take several days, so it is again much slower than two-photon emission. In this case, it may be useful to think in terms of the complete atom, including the proton spin. The electron and proton can combine their spins into a singlet state with zero net angular momentum or a triplet state with one unit of net momentum. The photon takes one unit of angular momentum away, turning a triplet state into a singlet state or vice-versa. If the atom ends up in a 1s triplet state, it will take another 10 million year or so to decay to the singlet state, the true ground state.

For excited atomic states in general, different types of transitions may be possible. As discussed in the previous subsection, the normal type is called an "allowed," "electric dipole," or E1 transition.

Yes, *every one* of these three names is confusing. Nonallowed transitions, called "forbidden" transitions, are perfectly allowed and they do occur. They are typically just a lot slower. The atomic states between which the transitions occur do not have electric dipole moments. And how many people really know what an electric dipole is? And E1 is just cryptic. E0 would have been more intuitive, as it indicates the level to which the transition is forbidden. Who cares about photon angular momentum?

The one good thing that can be said is that in the electric dipole approximation, the atom does indeed respond to the electric part of the electromagnetic

field. In such transitions the photon comes out with one unit of angular momentum, i.e. $\ell = 1$, and negative parity. Then the selection rules are:

$$\boxed{\text{E1:} \qquad l_{\rm L} = l_{\rm H} \pm 1 \qquad m_{l,{\rm L}} = m_{l,{\rm H}} \text{ or } m_{l,{\rm H}} \pm 1 \qquad m_{s,{\rm L}} = m_{s,{\rm H}}} \tag{7.18}$$

The first rule reflects the possible orbital angular momentum values as given above. To be sure, these values also allow l to stay the same. However, since the parity of the photon is negative, parity conservation requires that the parity of the atom must change, subsection 7.4.2. And that means that the orbital angular momentum quantum number l must change from odd to even or vice-versa. It cannot stay the same.

The second rule gives the possible magnetic quantum numbers. Recall that these are a direct measure for the angular momentum in the chosen z-direction. Since the photon momentum $\ell = 1$, the photon z momentum m_γ can be -1, 0, or 1. So the photon can change the atomic z momentum by up to one unit, as the selection rule says. Note that while the photon angular momentum cannot be zero in the direction of its motion, the direction of motion is not necessarily the z-direction. In essence the photon may be coming off sideways. (The better way of thinking about this is in terms of photon states of definite angular momentum. These can have the angular momentum zero in the z-direction, while the direction of photon motion is uncertain.)

The final selection rule says that the electron spin in the z-direction does not change. That reflects the fact that the electron spin does not respond to an electric field in a nonrelativistic approximation. (Of course, you might argue that in a nonrelativistic approximation, the electron should not have spin in the first place, chapter 12.12.)

Note that ignoring relativistic effects in transitions is a tricky business. Even a small effect, given enough time to build up, might produce a transition where one was not possible before. In a more sophisticated analysis of the hydrogen atom, addendum {A.39}, there is a slight interaction between the orbital angular momentum of the electron and its spin. That is known as spin-orbit interaction. Note that the s states have no orbital angular momentum for the spin to interact with.

As a result of spin-orbit interaction the correct energy eigenfunctions, except the s states, develop uncertainty in the values of both m_l and m_s. In other words, the z components of both the orbital and the spin angular momenta have uncertainty. That implies that the above rules are no longer really right. The energy eigenfunctions do keep definite values for l, representing the magnitude of orbital angular momentum, for j, representing the magnitude of net angular momentum, orbital plus spin, and m_j representing the net angular momentum in the z-direction. In those terms the modified selection rules become

$$\boxed{\text{E1}_{\rm so}\text{:} \qquad l_{\rm L} = l_{\rm H} \pm 1 \qquad j_{\rm L} = j_{\rm H} \text{ or } j_{\rm H} \pm 1 \qquad m_{j,{\rm L}} = m_{j,{\rm H}} \text{ or } m_{j,{\rm H}} \pm 1} \tag{7.19}$$

The last two selection rules above are a direct consequence of angular momentum conservation; since the photon has $\ell = 1$, it can change each atomic quantum number by at most one unit. In the first selection rule, angular momentum conservation could in principle allow a change in l by 2 units. A change in electron spin could add to the photon angular momentum. But parity conservation requires that l changes by an odd amount and 2 is not odd.

If the selection rules are not satisfied, the transition is called forbidden. However, the transition may still occur through a different mechanism. One possibility is a slower magnetic dipole transition, in which the electron interacts with the magnetic part of the electromagnetic field. That interaction occurs because an electron has spin and orbital angular momentum. A charged particle with angular momentum behaves like a little electromagnet and wants to align itself with an ambient magnetic field, chapter 13.4. The selection rules in this case are

$$\boxed{\text{M1:} \qquad l_{\rm L} = l_{\rm H} \qquad m_{l,{\rm L}} = m_{l,{\rm H}} \text{ or } m_{l,{\rm H}} \pm 1 \qquad m_{s,{\rm L}} = m_{s,{\rm H}} \text{ or } m_{s,{\rm H}} \pm 1} \tag{7.20}$$

The reasons are similar to the electric dipole case, taking into account that the photon comes out with positive parity rather than negative. Also, the electron spin definitely interacts with a magnetic field. A more detailed analysis will show that exactly one of the two magnetic quantum numbers m_l and m_s must change, {D.39}.

It must be pointed out that an M1 transition is trivial for an hydrogen atom in the nonrelativistic approximation. All the transition does is change the direction of the orbital or spin angular momentum vector, {D.39}. Not only is this ho-hum, the rate of transitions will be vanishingly small since it depends on the energy release in the transition. The same problem exists more generally for charged particles in radial potentials that only depend on position, {A.25.8}.

Relativistic effects can change this. In particular, in the presence of spin-orbit coupling, the selection rules become

$$\boxed{\text{M1}_{\rm so}\text{:} \qquad l_{\rm L} = l_{\rm H} \qquad j_{\rm L} = j_{\rm H} \text{ or } j_{\rm H}\pm 1 \qquad m_{j,{\rm L}} = m_{j,{\rm H}} \text{ or } m_{j,{\rm H}}\pm 1} \tag{7.21}$$

In this case, it is less obvious why l could not change by 2 units. The basic reason is that the magnetic field wants to rotate the orbital angular momentum vector, rather than change its magnitude, {D.39}. (Note however that that derivation, and this book in general, uses a nonrelativistic Hamiltonian for the interaction between the spin and the magnetic field.) Similar limitations apply for magnetic transitions of higher multipole order, {A.25.5} (A.175).

In higher-order multipole transitions the photon comes out with angular momentum ℓ greater than 1. As the previous subsection noted, this slows down the transitions. The fastest multipole transitions are the electric quadrupole ones. In these transitions the emitted photon has $\ell = 2$ and positive parity.

The selection rules are then

$$\boxed{\text{E2:} \quad l_{\rm L} = l_{\rm H} \text{ or } l_{\rm H}{\pm}2 \qquad m_{l,{\rm L}} = m_{l,{\rm H}} \text{ or } m_{l,{\rm H}}{\pm}1 \text{ or } m_{l,{\rm H}}{\pm}2 \qquad m_{s,{\rm L}} = m_{s,{\rm H}}} \tag{7.22}$$

In addition $l_{\rm H} = l_{\rm L} = 0$ is not possible for such transitions. Neither is $l_{\rm H} = 1$ and $l_{\rm L} = 0$ or vice-versa. Including electron spin, $j_{\rm H} = j_{\rm L} = \frac{1}{2}$ is not possible. The reasons are similar to the ones before.

Magnetic transitions at higher multipole orders have similar problems as the magnetic dipole one. In particular, consider the orbital angular momentum selection rule (7.17) above. The lowest possible multipole order in the nonrelativistic case is

$$\ell_{\rm min} = |l_{\rm H} - l_{\rm L}|$$

Because of parity, that is always an electric multipole transition. (This excludes the case that the orbital angular momenta are equal, in which case the lowest transition is the already discussed trivial magnetic dipole one.)

The bottom line is that magnetic transitions simply cannot compete. Of course, conservation of *net* angular momentum might forbid the electric transition to a given final state. But in that case there will be an equivalent state that differs only in spin to which the electric transition can proceed just fine.

However, for a multi-electron atom or nucleus in an independent-particle model, that equivalent state might already be occupied by another particle. Or there may be enough spin-orbit interaction to raise the energy of the equivalent state to a level that transition to it becomes impossible. In that case, the lowest possible transition will be a magnetic one.

Consider now more general systems than hydrogen atoms. General selection rules for electric Eℓ and magnetic Mℓ transitions are:

$$\boxed{\text{E}\ell\text{:} \quad |j_{\rm H} - \ell| \leqslant j_{\rm L} \leqslant j_{\rm H} + \ell \quad \text{and} \quad \pi_{\rm L} = \pi_{\rm H}(-1)^{\ell} \qquad (\ell \geqslant 1)} \tag{7.23}$$

$$\boxed{\text{M}\ell\text{:} \quad |j_{\rm H} - \ell| \leqslant j_{\rm L} \leqslant j_{\rm H} + \ell \quad \text{and} \quad \pi_{\rm L} = \pi_{\rm H}(-1)^{\ell-1} \qquad (\ell \geqslant 1)} \tag{7.24}$$

These rules rely only on the spin and parity of the emitted photon. So they are quite generally valid for one-photon emission.

If a normal electric dipole transition is possible for an atomic or nuclear state, it will most likely decay that way before any other type of transition can occur. But if an electric dipole transition is forbidden, other types of transitions may appear in significant amounts. If both electric quadrupole and magnetic dipole transitions are possible, they may be competitive. And electric quadrupole transitions can produce two units of change in the atomic angular momentum, rather than just one like the magnetic dipole ones.

Given the initial state, often the question is not what final states are possible, but what transition types are possible given the final state. In that case, the

general selection rules can be written as

$$\boxed{|j_{\rm H} - j_{\rm L}| \leqslant \ell \leqslant j_{\rm H} + j_{\rm L} \quad \text{and} \quad \ell \geqslant 1 \quad \text{and} \quad \pi_{\rm L}\pi_{\rm H}(-1)^\ell = \begin{cases} 1\text{: electric} \\ -1\text{: magnetic} \end{cases}} \tag{7.25}$$

Since transition rates decrease rapidly with increasing multipole order ℓ, normally the lowest value of ℓ allowed will be the important one. That is

$$\boxed{\ell_{\rm min} = |j_{\rm H} - j_{\rm L}| \quad \text{or} \quad 1 \text{ if } j_{\rm H} = j_{\rm L} \quad \text{and} \quad j_{\rm H} = j_{\rm L} = 0 \text{ is not possible.}} \tag{7.26}$$

If parity makes the corresponding transition magnetic, the next-higher order electric transition may well be of importance too.

Key Points

- Normal atomic transitions are called electric dipole ones, or allowed ones, or E1 ones. Unfortunately.
- The quantum numbers of the initial and final atomic states in transitions must satisfy certain selection rules in order for transitions of a given type to be possible.
- If a transition does not satisfy the rules of electric dipole transitions, it will have to proceed by a slower mechanism. That could be a magnetic dipole transition or an electric or magnetic multipole transition.
- A state of zero angular momentum cannot decay to another state of zero angular momentum through any of these mechanisms. For such transitions, two-photon emission is an option.

7.5 Symmetric Two-State Systems

This section will look at the simplest quantum systems that can have nontrivial time variation. They are called symmetric two-state systems. Despite their simplicity, a lot can be learned from them.

Symmetric two-state systems were encountered before in chapter 5.3. They describe such systems as the hydrogen molecule and molecular ion, chemical bonds, and ammonia. This section will show that they can also be used as a model for the fundamental forces of nature. And for the spontaneous emission of radiation by say excited atoms or atomic nuclei.

Two-state systems are characterized by just two basic states; these states will be called ψ_1 and ψ_2. For symmetric two-state systems, these two states must be physically equivalent. Or at least they must have the same expectation energy. And the Hamiltonian must be independent of time.

For example, for the hydrogen molecular ion ψ_1 is the state where the electron is in the ground state around the first proton. And ψ_2 is the state in which it is in the ground state around the second proton. Since the two protons are identical in their properties, there is no physical difference between the two states. So they have the same expectation energy.

The interesting quantum mechanics arises from the fact that the two states ψ_1 and ψ_2 are not energy eigenstates. The ground state of the system, call it $\psi_{\rm gs}$, is a symmetric combination of the two states. And there is also an excited energy eigenstate $\psi_{\rm as}$ that is an antisymmetric combination, chapter 5.3, {N.11}:

$$\psi_{\rm gs} = \frac{\psi_1 + \psi_2}{\sqrt{2}} \qquad \psi_{\rm as} = \frac{\psi_1 - \psi_2}{\sqrt{2}}$$

The above expressions may be inverted to give the states ψ_1 and ψ_2 in terms of the energy states:

$$\psi_1 = \frac{\psi_{\rm gs} + \psi_{\rm as}}{\sqrt{2}} \qquad \psi_2 = \frac{\psi_{\rm gs} - \psi_{\rm as}}{\sqrt{2}}$$

It follows that ψ_1 and ψ_2 are a 50/50 mixture of the low and high energy states. That means that they have uncertainty in energy. In particular they have a 50% chance for the ground state energy $E_{\rm gs}$ and a 50% chance for the elevated energy $E_{\rm as}$.

That makes their expectation energy $\langle E \rangle$ equal to the average of the two energies, and their uncertainty in energy ΔE equal to half the difference:

$$\langle E \rangle = \frac{E_{\rm gs} + E_{\rm as}}{2} \qquad \Delta E = \frac{E_{\rm as} - E_{\rm gs}}{2}$$

The question in this section is how the system evolves in time. In general the wave function is, section 7.1,

$$\Psi = c_{\rm gs} e^{-{\rm i}E_{\rm gs}t/\hbar} \psi_{\rm gs} + c_{\rm as} e^{-{\rm i}E_{\rm as}t/\hbar} \psi_{\rm as}$$

Here $c_{\rm gs}$ and $c_{\rm as}$ are constants that are arbitrary except for the normalization requirement.

However, this section will be more concerned with what happens to the basic states ψ_1 and ψ_2, rather than to the energy eigenstates. So, it is desirable to rewrite the wave function above in terms of ψ_1 and ψ_2 and their properties. That produces:

$$\Psi = e^{-{\rm i}\langle E \rangle t/\hbar} \left[c_{\rm gs} e^{{\rm i}\Delta E t/\hbar} \frac{\psi_1 + \psi_2}{\sqrt{2}} + c_{\rm as} e^{-{\rm i}\Delta E t/\hbar} \frac{\psi_1 - \psi_2}{\sqrt{2}} \right]$$

This expression is of the general form

$$\Psi = c_1 \psi_1 + c_2 \psi_2$$

According to the ideas of quantum mechanics, $|c_1|^2$ gives the probability that the system is in state ψ_1 and $|c_2|^2$ that it is in state ψ_2.

The most interesting case is the one in which the system is in the state ψ_1 at time zero. In that case the probabilities of the states ψ_1 and ψ_2 vary with time as

$$\boxed{|c_1|^2 = \cos^2(\Delta E\, t/\hbar) \qquad |c_2|^2 = \sin^2(\Delta E\, t/\hbar)} \tag{7.27}$$

To verify this, first note from the general wave function that if the system is in state ψ_1 at time zero, the coefficients $c_{\rm gs}$ and $c_{\rm as}$ must be equal. Then identify what c_1 and c_2 are and compute their square magnitudes using the Euler formula (2.5).

At time zero, the above probabilities produce state ψ_1 with 100% probability as they should. And so they do whenever the sine in the second expressions is zero. However, at times at which the cosine is zero, the system is fully in state ψ_2. It follows that the system is oscillating between the states ψ_1 and ψ_2.

Key Points

- Symmetric two-state systems are described by two quantum states ψ_1 and ψ_2 that have the same expectation energy $\langle E\rangle$.
- The two states have an uncertainty in energy ΔE that is not zero.
- The probabilities of the two states are given in (7.27). This assumes that the system is initially in state ψ_1.
- The system oscillates between states ψ_1 and ψ_2.

7.5.1 A graphical example

Consider a simple example of the oscillatory behavior of symmetric two-state systems. The example system is the particle inside a closed pipe as discussed in chapter 3.5. It will be assumed that the wave function is of the form

$$\Psi = \sqrt{\tfrac{4}{5}}e^{-{\rm i}E_{111}t/\hbar}\psi_{111} + \sqrt{\tfrac{1}{5}}e^{-{\rm i}E_{211}t/\hbar}\psi_{211}$$

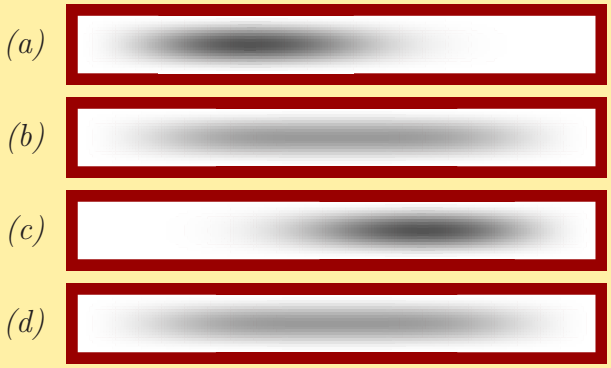
Here ψ_{111} and ψ_{211} are the ground state and the second lowest energy state, and E_{111} and E_{211} are the corresponding energies, as given in chapter 3.5.

The above wave function is a valid solution of the Schrödinger equation since the two terms have the correct exponential dependence on time. And since the two terms have different energies, there is uncertainty in energy.

The relative probability to find the particle at a given position is given by the square magnitude of the wave function. That works out to

$$|\Psi|^2 = \Psi^*\Psi = \tfrac{4}{5}|\psi_{111}|^2 + \tfrac{4}{5}\cos\Big((E_{111}-E_{211})t/\hbar\Big)\psi_{111}\psi_{211} + \tfrac{1}{5}|\psi_{211}|^2$$

Note that this result is time dependent. If there was no uncertainty in energy, which would be true if $E_{111} = E_{211}$, the square wave function would be independent of time.



Animation: `http://www.dommelen.net/quansup/pipemv.gif`

Figure 7.6: A combination of two energy eigenfunctions seen at some typical times.

The probability for finding the particle is plotted at four representative times in figure 7.6. After time *(d)* the evolution repeats at *(a)*. The wave function blob is sloshing back and forth in the pipe. That is much like a classical frictionless particle with kinetic energy would bounce back and forth between the ends of the pipe.

In terms of symmetric two-state systems, you can take the state ψ_1 to be the one in which the blob is at its leftmost position, figure 7.6*(a)*. Then ψ_2 is the state in which the blob is at its rightmost position, figure 7.6*(c)*. Note from the figure that these two states are physically equivalent. And they have uncertainty in energy.

Key Points

- A graphical example of a simple two-state system was give.

7.5.2 Particle exchange and forces

An important two-state system very similar to the simple example in the previous subsection is the hydrogen molecular ion. This ion consists of two protons and one electron.

The molecular ion can show oscillatory behavior very similar to that of the example. In particular, assume that the electron is initially in the ground state around the first proton, corresponding to state ψ_1. In that case, after some time interval Δt, the electron will be found in the ground state around the second

proton, corresponding to state ψ_2. After another time interval Δt, the electron will be back around the first proton, and the cycle repeats. In effect, the two protons play catch with the electron!

That may be fun, but there is something more serious that can be learned. As is, there is no (significant) force between the two protons. However, there is a second similar play-catch solution in which the electron is initially around the second proton instead of around the first. If these two solutions are symmetrically combined, the result is the ground state of the molecular ion. In this state of lowered energy, the protons are bound together. In other words, there is now a force that holds the two protons together:

> *If two particles play catch, it can produce forces between these two particles.*

A "play catch" mechanism as described above is used in more advanced quantum mechanics to explain the forces of nature. For example, consider the correct, relativistic, description of electromagnetism, given by "quantum electrodynamics". In it, the electromagnetic interaction between two charged particles comes about largely through processes in which one particle creates a photon that the other particle absorbs and vice versa. Charged particles play catch using photons.

That is much like how the protons in the molecular ion get bound together by exchanging the electron. Note however that the solution for the ion was based on the Coulomb potential. This potential implies instantaneous interaction at a distance: if, say, the first proton is moved, the electron and the other proton notice this instantaneously in the force that they experience. Classical relativity, however, does not allow effects that propagate at infinite speed. The highest possible propagation speed is the speed of light. In classical electromagnetics, charged particles do not really interact instantaneously. Instead charged particles interact with the electromagnetic field at their location. The electromagnetic field then communicates this to the other charged particles, at the speed of light. The Coulomb potential is merely a simple approximation, for cases in which the particle velocities are much less than the speed of light.

In a relativistic quantum description, the electromagnetic field is quantized into photons. (A concise introduction to this advanced topic is in addendum {A.23}.) Photons are bosons with spin 1. Similarly to classical electrodynamics, in the quantum description charged particles interact with photons at their location. They do not interact directly with other charged particles.

These are three-particle interactions, a boson and two fermions. For example, if an electron absorbs a photon, the three particles involved are the photon, the electron before the absorption, and the electron after the absorption. (Since in relativistic applications particles may be created or destroyed, a particle after an interaction should be counted separately from an identical particle that may exist before it.)

The ideas of quantum electrodynamics trace back to the early days of quantum mechanics. Unfortunately, there was the practical problem that the computations came up with infinite values. A theory that got around this problem was formulated in 1948 independently by Julian Schwinger and Sin-Itiro Tomonaga. A different theory was proposed that same year by Richard Feynman based on a more pictorial approach. Freeman Dyson showed that the two theories were in fact equivalent. Feynman, Schwinger, and Tomonaga received the Nobel prize in 1965 for this work, Dyson was not included. (The Nobel prize in physics is limited to a maximum of three recipients.)

Following the ideas of quantum electrodynamics and pioneering work by Sheldon Glashow, Steven Weinberg and Abdus Salam in 1967 independently developed a particle exchange model for the so called "weak force." All three received the Nobel prize for that work in 1979. Gerardus 't Hooft and Martinus Veltman received the 1999 Nobel Prize for a final formulation of this theory that allows meaningful computations.

The weak force is responsible for the beta decay of atomic nuclei, among other things. It is of key importance for such nuclear reactions as the hydrogen fusion that keeps our sun going. In weak interactions, the exchanged particles are not photons, but one of three different bosons of spin 1: the negatively charged W^-, (think W for weak force), the positively charged W^+, and the neutral Z^0 (think Z for zero charge). You might call them the "massives" because they have a nonzero rest mass, unlike the photons of electromagnetic interactions. In fact, they have gigantic rest masses. The $W^\pm$ have an experimental rest mass energy of about 80 GeV (giga-electron-volt) and the Z^0 about 91 GeV. Compare that with the rest mass energy of a proton or neutron, less than a GeV, or an electron, less than a thousandth of a GeV. However, a memorable name like "massives" is of course completely unacceptable in physics. And neither would be "weak-force carriers," because it is accurate and to the point. So physicists call them the "intermediate vector bosons." That is also three words, but completely meaningless to most people and almost meaningless to the rest, {A.20}. It meets the requirements of physics well.

A typical weak interaction might involve the creation of say a W^- by a quark inside a neutron and its absorption in the creation of an electron and an antineutrino. Now for massive particles like the intermediate vector bosons to be created out of nothing requires a gigantic quantum uncertainty in energy. Following the idea of the energy-time equality (7.9), such particles can only exist for extremely short times. And that makes the weak force of extremely short range.

The theory of "quantum chromedynamics" describes the so-called "strong force" or "color force." This force is responsible for such things as keeping atomic nuclei together.

The color force acts between "quarks." Quarks are the constituents of "baryons" like the proton and the neutron, and of "mesons" like the pions.

In particular, baryons consist of three quarks, while mesons consist of a quark and an antiquark. For example, a proton consists of two so-called "up quarks" and a third "down quark." Since up quarks have electric charge $\frac{2}{3}e$ and down quarks $-\frac{1}{3}e$, the net charge of the proton $\frac{2}{3}e+\frac{2}{3}e-\frac{1}{3}e$ equals e. Similarly, a neutron consists of one up quark and two down quarks. That makes its net charge $\frac{2}{3}e-\frac{1}{3}e-\frac{1}{3}e$ equal to zero. As another example, a so-called π^+ meson consists of an up quark and an antidown quark. An antiparticle has the opposite charge from the corresponding particle, so the charge of the π^+ meson $\frac{2}{3}e+\frac{1}{3}e$ equals e, the same as the proton. Three antiquarks make up an antibaryon. That gives an antibaryon the opposite charge of the corresponding baryon. More exotic baryons and mesons may involve the strange, charm, bottom, and top flavors of quarks. (Yes, there are six of them. You might well ask, "Who ordered that?" as the physicist Rabi did in 1936 upon the discovery of the muon, a heavier version of the electron. He did not know the least of it.)

Quarks are fermions with spin $\frac{1}{2}$ like electrons. However, quarks have an additional property called "color charge." (This color charge has nothing to do with the colors you can see. There are just a few superficial similarities. Physicists love to give complete different things identical names because it promotes such hilarious confusion.) There are three quark "colors" called, you guessed it, red, green and blue. There are also three corresponding "anticolors" called cyan, magenta, and yellow.

Now the electric charge of quarks can be observed, for example in the form of the charge of the proton. But their color charge cannot be observed in our macroscopic world. The reason is that quarks can only be found in "colorless" combinations. In particular, in baryons each of the three quarks takes a different color. (For comparison, on a video screen full-blast red, green and blue produces a colorless white.) Similarly, in antibaryons, each of the antiquarks takes on a different anticolor. In mesons the quark takes on a color and the antiquark the corresponding anticolor. (For example on a video screen, if you define antigreen as magenta, i.e. full-blast red plus blue, then green and antigreen produces again white.)

Actually, it is a bit more complicated still than that. If you had a green and magenta flag, you might call it color-balanced, but you would definitely not call it colorless. At least not in this book. Similarly, a green-antigreen meson would not be colorless, and such a meson does not exist. An actual meson is an quantum superposition of the three possibilities red-antired, green-antigreen, and blue-antiblue. The meson color state is

$$\frac{1}{\sqrt{3}}(r\bar{r}+g\bar{g}+b\bar{b})$$

where a bar indicates an anticolor. Note that the quark has equal probabilities of being observed as red, green, or blue. Similarly the antiquark has equal

probabilities of being observed antired, antigreen, or antiblue, but always the anticolor of the quark.

In addition, the meson color state above is a one-of-a-kind, or "singlet" state. To see why, suppose that, say, the final $b\bar{b}$ term had a minus sign instead of a plus sign. Then surely, based on symmetry arguments, there should also be states where the $g\bar{g}$ or $r\bar{r}$ has the minus sign. And that cannot be true because linear combinations of such states would produce states like the green-antigreen meson that are not colorless. So the only true colorless possibility is the state above, where all three color-anticolor states have the same coefficient. (Do recall that a constant of magnitude one is indeterminate in quantum states. So if all three color-anticolor states had a minus sign, it would still be the same state.)

Similarly, an "rgb" baryon with the first quark red, the second green, and the third blue would be color-balanced but not colorless. So such a baryon does not exist. For baryons there are six different possible color combinations: there are three possibilities for which of the three quarks is red, times two possibilities which of the remaining two quarks is green. An actual baryon is a quantum superposition of these six possibilities. Moreover, the combination is antisymmetric under color exchange:

$$\frac{1}{\sqrt{6}}(rgb - rbg + gbr - grb + brg - bgr)$$

Equivalently, the combination is antisymmetric under quark exchange. That explains why the so-called Δ^{++} delta baryon can exist. This baryon consists of three up quarks in a symmetric spatial ground state and a symmetric spin $\frac{3}{2}$ state, like $\uparrow\uparrow\uparrow$. Because of the antisymmetric color state, the antisymmetrization requirements for the three quarks can be satisfied. The color state above is again a singlet one. In terms of chapter 5.7, it is the unique Slater determinant that can be formed from three states for three particles.

It is believed that baryons and mesons cannot be taken apart into separate quarks to study quarks in isolation. In other words, quarks are subject to "confinement" inside colorless baryons and mesons. The problem with trying to take these apart is that the force between quarks does not become zero with distance like other forces. If you try to take a quark out of a baryon or meson, presumably eventually you will put in enough energy to create a quark-antiquark pair in between. That kills off the quark separation that you thought you had achieved.

The color force between quarks is due to the exchange of so-called "gluons." Gluons are massless bosons with spin 1 like photons. However, photons do not carry electric charge. Gluons do carry color/anticolor combinations. That is one reason that quantum chromedynamics is enormously more difficult than quantum electrodynamics. Photons cannot move electric charge from one fermion to the next. But gluons allow the interchange of colors between quarks.

Also, because photons have no charge, they do not interact with other photons. But since gluons themselves carry color, gluons do interact with other gluons. In fact, both three-gluon and four-gluon interactions are possible. In principle, this makes it conceivable that "glueballs," colorless combinations of gluons, might exist. However, at the time of writing, 2012, only baryons, antibaryons, and mesons have been solidly established.

Gluon-gluon interactions are related to an effective strengthening of the color force at larger distances. Or as physicists prefer to say, to an effective weakening of the interactions at short distances called "asymptotic freedom." This helps a bit because it allows some analysis to be done at very short distances, i.e. at very high energies.

Normally you would expect nine independent color/anticolor gluon states: there are three colors times three anticolors. But in fact only eight independent gluon states are believed to exist. Recall the colorless meson state described above. If a gluon could be in such a colorless state, it would not be subject to confinement. It could then be exchanged between distant protons and neutrons, giving rise to a long-range nuclear force. Since such a force is not observed, it must be concluded that gluons cannot be in the colorless state. So if the nine independent orthonormal color states are taken to be the colorless state plus eight more states orthogonal to it, then only the latter eight states can be observable. In terms of section 7.3, the relevant symmetry of the color force must be SU(3), not U(3).

Many people contributed to the theory of quantum chromedynamics. However Murray Gell-Mann seemed to be involved in pretty much every stage. He received the 1969 Nobel Prize at least in part for his work on quantum chromedynamics. It is also he who came up with the name "quark." The name is really not bad compared to many other terms in physics. However, Gell-Mann is also responsible for not spelling "color" as "qolor." That would have saved countless feeble explanations that, "No, this color has absolutely nothing to do with the color that you see in nature." So far nobody has been able to solve that problem, but David Gross, David Politzer and Frank Wilczek did manage to discover the asymptotic freedom mentioned above. For that they were awarded the 2004 Nobel Prize in Physics.

It may be noted that Gell-Mann initially called the three colors red, white, and blue. Just like the colors of the US flag, in short. Or of the Netherlands and Taiwan, to mention a few others. Huang, [27, p. 167], born in China, with a red and yellow flag, claims red, yellow and green are now the conventional choice. He must live in a world different from ours. Sorry, but the honor of having the color-balanced, (but not colorless), flag goes to Azerbaijan.

The force of gravity is supposedly due to the exchange of particles called "gravitons." They should be massless bosons with spin 2. However, it is hard to experiment with gravity because of its weakness on human scales. The graviton remains unconfirmed. Worse, the exact place of gravity in quantum mechanics

remains very controversial.

Key Points

- The fundamental forces are due to the exchange of particles.
- The particles are photons for electromagnetism, intermediate vector bosons for the weak force, gluons for the color force, and presumably gravitons for gravity.

7.5.3 Spontaneous emission

Symmetric two state systems provide the simplest model for spontaneous emission of radiation by atoms or atomic nuclei. The general ideas are the same whether it is an atom or nucleus, and whether the radiation is electromagnetic (like visible light) or nuclear alpha or beta radiation. But to be specific, this subsection will use the example of an excited atomic state that decays to a lower energy state by releasing a photon of electromagnetic radiation. The conservation laws applicable to this process were discussed earlier in section 7.4. This subsection wants to examine the actual mechanics of the emission process.

First, there are some important terms and concepts that must be mentioned. You will encounter them all the time in decay processes.

The big thing is that decay processes are random. A typical atom in an excited state $\psi_{\rm H}$ will after some time transition to a state of lower energy $\psi_{\rm L}$ while releasing a photon. But if you take a second identical atom in the exact same excited state, the time after which this atom transitions will be different.

Still, the decay process is not completely unpredictable. Averages over large numbers of atoms have meaningful values. In particular, suppose that you have a very large number I of identical excited atoms. Then the "decay rate" is by definition

$$\boxed{\lambda = -\frac{1}{I}\frac{{\rm d}I}{{\rm d}t}} \tag{7.28}$$

It is the relative fraction $-{\rm d}I/I$ of excited atoms that disappears per unit time through transitions to a lower energy state. The decay rate has a precise value for a given atomic state. It is not a random number.

To be precise, the above decay rate is better called the specific decay rate. The actual decay rate is usually defined to be simply $-{\rm d}I/{\rm d}t$. But anyway, decay rate is not a good term to use in physics. It is much too clear. Sometimes the term "spontaneous emission rate" or "transition rate" is used, especially in the context of atoms. But that is even worse. A better and very popular choice is "decay constant." But, while "constant" is a term that can mean anything, it really is still far too transparent. How does "disintegration constant" sound? Especially since the atom hardly disintegrates in the transition? Why not call

it the [specific] "activity," come to think of it? Activity is another of these vague terms. Another good one is "transition probability," because a probability should be nondimensional and λ is per unit time. May as well call it "radiation probability" then. Actually, many references will use a bunch of these terms interchangeably on the same page.

In fact, would it not be a good thing to take the inverse of the decay rate? That allows another term to be defined for essentially the same thing: the [mean] "lifetime" of the excited state:

$$\boxed{\tau \equiv \frac{1}{\lambda}} \tag{7.29}$$

Do remember that this is not really a lifetime. Each individual atom has its own lifetime. (However, if you average the lifetimes of a large number of identical atoms, you will in fact get the mean lifetime above.)

Also, remember, if more than one decay process occurs for the excited state,

Add decay rates, not lifetimes.

The sum of the decay rates gives the total decay rate of the atomic state. The reciprocal of that total is the correct lifetime.

Now suppose that initially there is a large number I_0 of excited atoms. Then the number of excited atoms I left at a later time t is

$$\boxed{I = I_0 e^{-\lambda t}} \tag{7.30}$$

So the number of excited atoms left decays exponentially in time. To check this expression, just check that it is right at time zero and plug it into the definition for the decay rate.

A quantity with a clearer physical meaning than lifetime is the time for about half the nuclei in a given large sample of excited atoms to decay. This time is called the "half-life" $\tau_{1/2}$. From (7.30) and (7.29) above, it follow that the half-life is shorter than the lifetime by a factor $\ln 2$:

$$\boxed{\tau_{1/2} = \tau \ln 2} \tag{7.31}$$

Note that $\ln 2$ is less than one.

The purpose in this subsection is now to understand some of the above concepts in decays using the model of a symmetric two-state system.

The initial state ψ_1 of the system is taken to be an atom in a high-energy atomic state $\psi_{\rm H}$, figure 7.3. The state seems to be an state of definite energy. That would make it a stationary state, section 7.1.4, and hence it would not decay. However, ψ_1 is not really an energy eigenstate, because an atom is always perturbed by a certain amount of ambient electromagnetic radiation. The actual state ψ_1 has therefore some uncertainty in energy ΔE.

The decayed state ψ_2 consists of an atomic state of lowered energy $\psi_{\rm L}$ plus an emitted photon. This state seems to have the same combined energy as the

initial state ψ_1. It too, however, is not really an energy eigenstate. Otherwise it would always have existed. In fact, it has the same expectation energy and uncertainty in energy as the initial state, section 7.1.3.

The probabilities of the two states were given at the start of this section. They were:

$$|c_1|^2 = \cos^2(\Delta E\, t/\hbar) \qquad |c_2|^2 = \sin^2(\Delta E\, t/\hbar) \tag{7.32}$$

At time zero, the system is in state ψ_1 for sure, but after a time interval Δt it is in state ψ_2 for sure. The atom has emitted a photon and decayed. An expression for the time that this takes can be found by setting the angle in the sine equal to $\frac{1}{2}\pi$. That gives:

$$\Delta t = \tfrac{1}{2}\pi\hbar/\Delta E$$

But note that there is a problem. According to (7.32), after another time interval Δt the probabilities of the two states will revert back to the initial ones. That means that the low energy atomic state absorbs the photon again and so returns to the excited state!

Effects like that do occur in nuclear magnetic resonance, chapter 13.6, or for atoms in strong laser light and high vacuum, [52, pp. 147-152]. But normally, decayed atoms stay decayed.

To explain that, it must be assumed that the state of the system is "measured" according to the rules of quantum mechanics, chapter 3.4. The macroscopic surroundings "observes" that a photon is released well before the original state can be restored. In the presence of such significant interaction with the macroscopic surroundings, the two-state evolution as described above is no longer valid. In fact, the macroscopic surroundings will have become firmly committed to the fact that the photon has been emitted. Little chance for the atom to get it back under such conditions.

In an improved model of the transition process, section 7.6.1, the need for measurement remains. However, the reasons get more complex.

Interactions with the surroundings are generically called "collisions." For example, a real-life atom in a gas will periodically collide with neighboring atoms and other particles. If a process is fast enough that no interactions with the surroundings occur during the time interval of interest, then the process takes place in the so-called "collisionless regime." Nuclear magnetic resonance and atoms in strong laser light and high vacuum may be in this regime.

However, normal atomic decays take place in the so-called "collision-dominated regime." Here collisions with the surroundings occur almost immediately.

To model that, take the time interval between collisions to be t_c. Assume that the atom evolves as an unperturbed two-state system until time t_c. At that time however, the atom is "measured" by its surroundings and it is either found to be in the initial excited state ψ_1 or in the decayed state with photon ψ_2. According to the rules of quantum mechanics the result is random. However,

they are not completely random. The probability $P_{1\to2}$ for the atom to be found to be decayed is the square magnitude $|c_2|^2$ of the state ψ_2.

That square magnitude was given in (7.32). But it may be approximated to:

$$P_{1\to2} = \frac{|\Delta E|^2}{\hbar^2} t_{\rm c}^2$$

This approximated the sine in (7.32) by its argument, since the time $t_{\rm c}$ is assumed small enough that the argument is small.

Note that the decay process has become probabilistic. You cannot say for sure whether the atom will be decayed or not at time $t_{\rm c}$. You can only give the chances. See chapter 8.6 for a further discussion of that philosophical issue.

However, if you have not just one excited atom, but a large number I of them, then $P_{1\to2}$ above is the relative fraction that will be found to be decayed at time $t_{\rm c}$. The remaining atoms, which are found to be in the excited state, (or rather, have been pushed back into the excited state), start from scratch. Then at time $2t_{\rm c}$, a fraction $P_{1\to2}$ of these will be found to be decayed. And so on. Over time the number I of excited atoms decreases to zero.

As mentioned earlier, the relative fraction of excited atoms that disappears per unit time is called the decay rate λ. That can be found by simply dividing the decay probability $P_{1\to2}$ above by the time $t_{\rm c}$ that the evolution took. So

$$\lambda_{1\to2} = \frac{|H_{21}|^2}{\hbar^2} t_{\rm c} \qquad H_{21} = \Delta E = \langle \psi_2 | H | \psi_1 \rangle.$$

Here the uncertainty in energy ΔE was identified in terms of the Hamiltonian H using the analysis of chapter 5.3.

Physicists call H_{21} the "matrix element." That is well below their usual form, because it really is a matrix element. But before you start seriously doubting the capability of physicists to invariably come up with confusing terms, note that there are lots of different matrices in any advanced physical analysis. So the name does not give its secret away to nonspecialists. To enforce that, many physicists write matrix elements in the form M_{21}, because, hey, the word matrix starts with an m. That hides the fact that it is an element of a Hamiltonian matrix pretty well.

The good news is that the assumption of collisions has solved the problem of decayed atoms undecaying again. Also, the decay process is now probabilistic. And the decay rate $\lambda_{1\to2}$ above is a normal number, not a random one.

Unfortunately, there are a couple of major new problems. One problem is that the state ψ_2 has one more particle than state ψ_1; the emitted photon. That makes it impossible to evaluate the matrix element using nonrelativistic quantum mechanics as covered in this book. Nonrelativistic quantum mechanics does not allow for new particles to be created or old ones to be destroyed. To evaluate the matrix element, you need relativistic quantum mechanics. Section 7.8 will eventually manage to work around that limitation using a dirty trick.

Addendum {A.24} gives the actual relativistic derivation of the matrix element. However, to really understand that addendum, you may have to read a couple of others.

An even bigger problem is that the decay rate above is proportional to the collision time t_c. That makes it completely dependent on the details of the surroundings of the atom. But that is wrong. Atoms have very specific decay rates. These rates are the same under a wide variety of environmental conditions.

The basic problem is that in reality there is not just a single decay process for an excited atom; there are infinitely many. The derivation above assumed that the photon has an energy exactly given by the difference between the atomic states. However, there is uncertainty in energy one way or the other. Decays that produce photons whose frequency is ever so slightly different will occur too. To deal with that complication, asymmetric two-state systems must be considered. That is done in the next section.

Finally, a few words should probably be said about what collisions really are. Darn. Typically, they are pictured as atomic collisions. But that may be in a large part because atomic collisions are quite well understood from classical physics. Atomic collisions do occur, and definitely need to be taken into account, like later in the derivations of {D.41}. But in the above description, collisions take on a second role as doing quantum mechanical "measurements." In that second role, a collision has occurred if the system has been "measured" to be in one state or the other. Following the analysis of chapter 8.6, measurement should be taken to mean that the surroundings has become firmly committed that the system has decayed. In principle, that does not require any actual collision with the atom; the surroundings could simply observe that the photon is present. The bad news is that the entire process of measurement is really not well understood at all. In any case, the bottom line to remember is that collisions do not necessarily represent what you would intuitively call collisions. Their dual role is to represent the typical moment that the surroundings commits itself that a transition has occurred.

Key Points

- The two-state system provides a model for the decay of excited atoms or nuclei.
- Interaction with the surroundings is needed to make the decay permanent. That makes decays probabilistic.
- The [specific] decay rate, λ is the relative fraction of particles that decays per unit time. Its inverse is the mean lifetime τ of the particles. The half-life $\tau_{1/2}$ is the time it takes for half the particles in a big sample to decay. It is shorter than the mean lifetime by a factor $\ln 2$.

Always add decay rates, not lifetimes.

7.6 Asymmetric Two-State Systems

Two-state systems are quantum systems for which just two states ψ_1 and ψ_2 are relevant. If the two states have different expectation energy, or if the Hamiltonian depends on time, the two-state system is asymmetric. Such systems must be considered to fix the problems in the description of spontaneous emission that turned up in the previous section.

The wave function of a two state system is of the form

$$\Psi = c_1\psi_1 + c_2\psi_2 \tag{7.33}$$

where $|c_1|^2$ and $|c_2|^2$ are the probabilities that the system is in state ψ_1, respectively ψ_2.

The coefficients c_1 and c_2 evolve in time according to

$$\boxed{\mathrm{i}\hbar\dot{c}_1 = \langle E_1\rangle c_1 + H_{12}c_2 \qquad \mathrm{i}\hbar\dot{C}_2 = H_{21}c_1 + \langle E_2\rangle c_2} \tag{7.34}$$

where

$$\langle E_1\rangle = \langle\psi_1|H\psi_1\rangle, \quad H_{12} = \langle\psi_1|H\psi_2\rangle, \quad H_{21} = \langle\psi_2|H\psi_1\rangle, \quad \langle E_2\rangle = \langle\psi_2|H\psi_2\rangle$$

with H the Hamiltonian. The Hamiltonian coefficients $\langle E_1\rangle$ and $\langle E_2\rangle$ are the expectation energies of states ψ_1 and ψ_2. The Hamiltonian coefficients H_{12} and H_{21} are complex conjugates. Either one is often referred to as the "matrix element." To derive the above evolution equations, plug the two-state wave function Ψ into the Schrödinger equation and take inner products with $\langle\psi_1|$ and $\langle\psi_2|$, using orthonormality of the states.

It will be assumed that the Hamiltonian is independent of time. In that case the evolution equations can be solved analytically. To do so, the analysis of chapter 5.3 can be used to find the energy eigenstates and then the solution is given by the Schrödinger equation, section 7.1.2. However, the final solution is messy. The discussion here will restrict itself to some general observations about it.

It will be assumed that the solution starts out in the state ψ_1. That means that initially $|c_1|^2 = 1$ and $|c_2|^2 = 0$. Then in the symmetric case discussed in the previous section, the system oscillates between the two states. But that requires that the states have the same expectation energy.

This section addresses the asymmetric case, in which there is a nonzero difference E_{21} between the two expectation energies:

$$\boxed{E_{21} \equiv \langle E_2\rangle - \langle E_1\rangle} \tag{7.35}$$

In the asymmetric case, the system never gets into state ψ_2 completely. There is always some probability for state ψ_1 left. That can be seen from energy conservation: the expectation value of energy must stay the same during the evolution, and it would not if the system went fully into state 2. However, the system will periodically return fully to the state ψ_1. That is all that will be said about the exact solution here.

The remainder of this section will use an approximation called "time-dependent perturbation theory." It assumes that the system stays close to a given state. In particular, it will be assumed that the system starts out in state ψ_1 and stays close to it.

That assumption results in the following probability for the system to be in the state ψ_2, {D.38}:

$$\boxed{|c_2|^2 \approx \left(\frac{|H_{21}|t}{\hbar}\right)^2 \frac{\sin^2(E_{21}t/2\hbar)}{(E_{21}t/2\hbar)^2}} \tag{7.36}$$

For this expression to be a valid approximation, the parenthetical ratio must be small. Note that the final factor shows the effect of the asymmetry of the two state system; E_{21} is the difference in expectation energy between the states. For a symmetric two-state system, the final factor would be 1, (using l'Hôpital).

Key Points

- If the states in a two-state system have different expectation energies, the system is asymmetric.
- If the system is initially in the state ψ_1, it will never fully get into the state ψ_2.
- If the system is initially in the state ψ_1 and remains close to it, then the probability of the state ψ_2 is given by (7.36)

7.6.1 Spontaneous emission revisited

Decay of excited atomic or nuclear states was addressed in the previous section using symmetric two-state systems. But there were some issues. They can now be addressed.

The example is again an excited atomic state that transitions to a lower energy state by emitting a photon. The state ψ_1 is the excited atomic state. The state ψ_2 is the atomic state of lowered energy plus the emitted photon. These states seem states of definite energy, but if they really were, there would not be any decay. Energy states are stationary. There is a slight uncertainty in energy in the states.

Since there is, clearly it does not make much sense to say that the initial and final expectation energies must be the same *exactly*.

In decay processes, a bit of energy slop E_{21} must be allowed between the initial and final expectation values of energy.

In practical terms, that means that the energy of the emitted photon can vary a bit. So its frequency can vary a bit.

Now in infinite space, the possible photon frequencies are infinitely close together. So you are now suddenly dealing with not just one possible decay process, but infinitely many. That would require messy, poorly justified mathematics full of so-called delta functions.

Instead, in this subsection it will be assumed that the atom is not in infinite space, but in a very large periodic box, chapter 6.17. The decay rate in infinite space can then be found by taking the limit that the box size becomes infinite. The advantage of a finite box is that the photon frequencies, and so the corresponding energies, are discrete. So you can sum over them rather than integrate.

Each possible photon state corresponds to a different final state ψ_2, each with its own coefficient c_2. The square magnitude of that coefficient gives the probability that the system can be found in that state ψ_2. And in the approximation of time-dependent perturbation theory, the coefficients c_2 do not interact; the square magnitude of each is given by (7.36). The total probability that the system can be found in *some* decayed state at a time t_c is then

$$P_{1\to\text{all }2} = \sum_{\text{all states }2} \left(\frac{|H_{21}|t_c}{\hbar}\right)^2 \frac{\sin^2(E_{21}t_c/2\hbar)}{(E_{21}t_c/2\hbar)^2}$$

The time t_c will again model the time between "collisions," interactions with the surroundings that "measure" whether the atom has decayed or not. The decay rate, the number of transitions per unit time, is found from dividing by the time:

$$\lambda = \sum_{\text{all states }2} \frac{|H_{21}|^2}{\hbar^2} t_c \frac{\sin^2(E_{21}t_c/2\hbar)}{(E_{21}t_c/2\hbar)^2}$$

The final factor in the sum for the decay rate depends on the energy slop E_{21}. This factor is plotted graphically in figure 7.7. Notice that only a limited range around the point of zero slop contributes much to the decay rate. The spikes in the figure are intended to qualitatively indicate the discrete photon frequencies that are possible in the box that the atom is in. If the box is extremely big, then these spikes will be extremely close together.

Now suppose that you plot the energy slop diagram against the actual photon energy instead of the scaled energy slop $E_{21}t_c/2\hbar$. Then the center of the diagram will be at the nominal energy of the emitted photon and E_{21} will be the deviation from that nominal energy. The spike at the center then represents the transition of atoms where the photon comes out with exactly the nominal energy. And those surrounding spikes whose height is not negligible represent

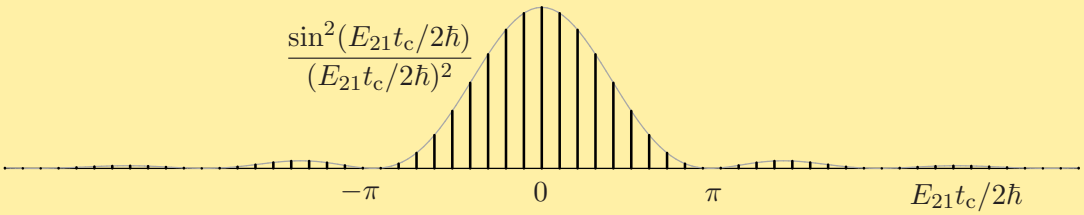


Figure 7.7: Energy slop diagram.

slightly different photon energies that have a reasonable probability of being observed. So the energy slop diagram, plotted against photon energy, graphically represents the uncertainty in energy of the final state that will be observed.

Normally, the observed uncertainty in energy is very small in physical terms. The energy of the emitted photon is almost exactly the nominal one; that allows spectral analysis to identify atoms so well. So the entire diagram figure 7.7 is extremely narrow horizontally when plotted against the photon energy.

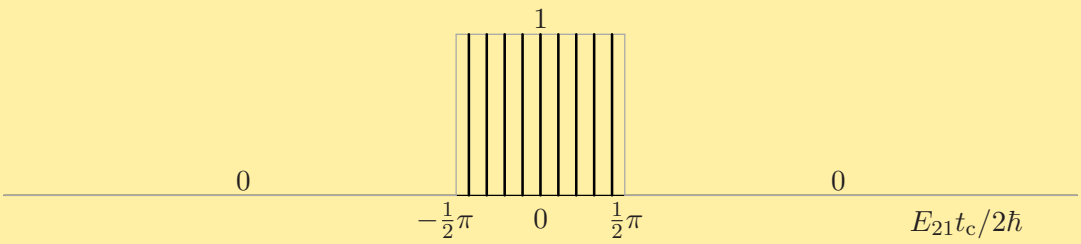


Figure 7.8: Schematized energy slop diagram.

That suggests that you can simplify things by replacing the energy slop diagram by the schematized one of figure 7.8. This diagram is zero if the energy slop is greater than $\pi\hbar/t_c$, and otherwise it is one. And it integrates to the same value as the original function. So, if the spikes are very closely spaced, they still sum to the same value as before. To be sure, if the square matrix element $|H_{21}|^2$ varied nonlinearly over the typical width of the diagram, the transition rate would now sum to something else. But it should not; if the variation in photon energy is negligible, then so should the one in the matrix element be.

Using the schematized energy slop diagram, you only need to sum over the states whose spikes are equal to 1. That are the states 2 whose expectation energy is no more than $\pi\hbar/t_c$ different from the initial expectation energy. And inside this summation range, the final factor can be dropped because it is now 1. That gives:

$$\lambda = \sum_{\substack{\text{all states 2 with} \\ |\langle E_2\rangle - \langle E_1\rangle| < \pi\hbar/t_c}} \frac{|H_{21}|^2}{\hbar^2} t_c \tag{7.37}$$

This can be cleaned up further, assuming that H_{21} is constant and can be taken out of the sum:

$$\boxed{\lambda = 2\pi \frac{|H_{21}|^2}{\hbar} \frac{\mathrm{d}N}{\mathrm{d}\langle E_2 \rangle}} \tag{7.38}$$

This formula is known as "Fermi's golden rule." The final factor is the number of photon states per unit energy range. It is to be evaluated at the nominal photon energy. The formula simply observes that the number of terms in the sum is the number of photon states per unit energy range times the energy range. The equation is considered to originate from Dirac, but Fermi is the one who named it "golden rule number two."

Actually, the original sum (7.37) may be easier to handle in practice since the number of photon states per unit energy range is not needed. But Fermi's rule is important because it shows that the big problem of the previous section with decays has been resolved. The decay rate does no longer depend on the time between collisions t_c. Atoms can have specific values for their decay rates despite the minute details of their surroundings. Shorter collision times do produce less transitions per unit time for a given state. But they also allow more slop in energy, so the number of states that achieve a significant amount of transitions per unit time goes up. The net effect is that the decay rate stays the same, though the uncertainty in energy goes up.

The other problem remains; the evaluation of the matrix element H_{21} requires relativistic quantum mechanics. But it is not hard to guess the general ideas. When the size of the periodic box that holds the system increases, the electromagnetic field of the photons decreases; they have the same energy in a larger volume. That results in smaller values for the matrix element H_{21}. On the other hand, the number of photons per unit energy range $\mathrm{d}N/\mathrm{d}\langle E_2 \rangle$ increases, chapter 6.3. The net result will be that the decay rate remains finite when the box becomes infinite.

That is verified by the relativistic analysis in addendum {A.24}. That addendum completes the analysis in this section by computing the matrix element using relativistic quantum mechanics. Using a description in terms of photon states of definite linear momentum, the matrix element is inversely proportional to the volume of the box, but the density of states is directly proportional to it. (It is somewhat different using a description in terms of photon states of definite angular momentum, {A.25}. But the idea remains the same.)

One problem of section 7.5.3 that has now disappeared is the photon being reabsorbed again. For each individual transition process, the interaction is too weak to produce a finite reversal time. But quantum "measurement" remains required to explain the experiments. The time-dependent perturbation theory used does not apply if the quantum system is allowed to evolve undisturbed over a time long enough for a significant transition probability (to any state) to evolve, {D.38}. That would affect the specific decay rate. If you are merely

interested in the average emission and absorption of a large number of atoms, it is not a big problem. Then you can substitute a classical description in terms of random collisions for the quantum measurement process. That will be done in derivation {D.41}. But to describe what happens to individual atoms one at a time, while still explaining the observed statistics of many of such individual atoms, is another matter.

So far it has been assumed that there is only one atomic initial state of interest and only one final state. However, either state might have a net angular momentum quantum number j that is not zero. In that case, there are $2j + 1$ atomic states that differ only in magnetic quantum number. The magnetic quantum number describes the component of the angular momentum in the chosen z-direction. Now if the atom is in empty space, the direction of the z-axis should not make a difference. Then these $2j + 1$ states will have the same energy. So you cannot include one and not the other. If this happens to the initial atomic state, you will need to average the decay rates over the magnetic states. The physical reason is that if you have a large number I of excited atoms in the given energy state, their magnetic quantum numbers will be randomly distributed. So the average decay rate of the total sample is the average over the initial magnetic quantum numbers. But if it happens to the final state, you have to sum over the final magnetic quantum numbers. Each final magnetic quantum number gives an initial excited atom one more state that it can decay to. The general rule is:

> *Sum over the final atomic states, then average over the initial atomic states.*

The averaging over the initial states is typically trivial. Without a preferred direction, the decay rate will not depend on the initial orientation.

It is interesting to examine the limitations of the analysis in this subsection. First, time-dependent perturbation theory has to be valid. It might seem that the requirement of (7.36) that $H_{21}t_c/\hbar$ is small is automatically satisfied, because the matrix element H_{21} goes to zero for infinite box size. But then the number of states 2 goes to infinity. And if you look a bit closer at the analysis, {D.38}, the requirement is really that there is little probability of *any* transition in time interval t_c. So the time between collisions must be small compared to the lifetime of the state. With typical lifetimes in the range of nanoseconds, atomic collisions are typically a few orders of magnitude more rapid. However, that depends on the relative vacuum.

Second, the energy slop diagram figure 7.7 has to be narrow on the scale of the photon energy. It can be seen that this is true if the time between collisions t_c is large compared to the inverse of the photon frequency. For emission of visible light, that means that the collision time must be large when expressed in femtoseconds. Collisions between atoms will easily meet that requirement.

The width of the energy slop diagram figure 7.7 should give the observed variation E_{21} in the energy of the final state. The diagram shows that roughly

$$E_{21} t_c \sim \pi\hbar$$

Note that this takes the form of the all-powerful energy-time uncertainty equality (7.9). To be sure, the equality above involves the artificial time between collisions, or "measurements," t_c. But you could assume that this time is comparable to the mean lifetime τ of the state. Essentially that supposes that interactions with the surroundings are infrequent enough that the atomic evolution can evolve undisturbed for about the typical decay time. But that nature will definitely commit itself whether or not a decay has occurred as soon as there is a fairly reasonable probability that a photon has been emitted.

That argument then leads to the definition of the typical uncertainty in energy, or "width,"of a state as $\Gamma = \hbar/\tau$, as mentioned in section 7.4.1. In addition, if there are frequent interactions between the atom and its surroundings, the shorter collision time t_c should be expected to increase the uncertainty in energy to more than the width.

Note that the wavy nature of the energy slop diagram figure 7.7 is due to the assumption that the time between "collisions" is always the same. If you start averaging over a more physical random set of collision times, the waves will smooth out. The actual energy slop diagram as usually given is of the form

$$\frac{1}{1 + (E_{21}/\Gamma)^2} \tag{7.39}$$

That is commonly called a [Cauchy] "Lorentz[ian] profile" or distribution or function, or a "Breit-Wigner distribution." Hey, don't blame the messenger. In any case, it still has the same inverse quadratic decay for large energy slop as the diagram figure 7.7. That means that if you start computing the standard deviation in energy, you end up with infinity. That would be a real problem for versions of the energy-time relationship like the one of Mandelshtam and Tamm. Such versions take the uncertainty in energy to be the standard deviation in energy. But it is no problem for the all-powerful energy-time uncertainty equality (7.9), because the standard deviation in energy is not needed.

Key Points

- Some energy slop occurs in decays.
- Taking that into account, meaningful decay rates may be computed following Fermi's golden rule.

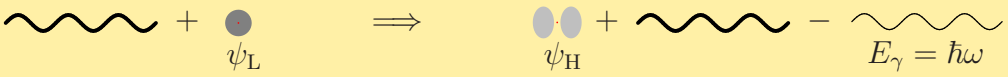
7.7 Absorption and Stimulated Emission

This section will address the basic physic of absorption and emission of radiation by a gas of atoms in an electromagnetic field. The next section will give practical formulae.

(a) Spontaneous emission:



(b) Absorption:



(c) Stimulated emission:



Figure 7.9: Emission and absorption of radiation by an atom.

Figure 7.9 shows the three different processes of interest. The previous sections already discussed the process of spontaneous emission. Here an atom in a state $\psi_{\rm H}$ of high energy emits a photon of electromagnetic radiation and returns to an atomic state $\psi_{\rm L}$ of lower energy. For example, for a hydrogen atom the excited state $\psi_{\rm H}$ might be the ψ_{210} "$2p_z$" state, and the lower energy state $\psi_{\rm L}$ the ψ_{100} "1s" ground state, as defined in chapter 4.3.

To a superb approximation, the photon carries off the difference in energy between the atomic states. In view of the Planck-Einstein relation, that means that its frequency ω is given by

$$\hbar\omega = E_{\rm H} - E_{\rm L}$$

Unfortunately, the discussion of spontaneous emission in the previous sections had to remain incomplete. Nonrelativistic quantum mechanics as covered in this book cannot accommodate the creation of new particles like the photon in this case. The number of particles has to stay the same.

The second process of interest in figure 7.9 is absorption. Here an atom in a low energy state $\psi_{\rm L}$ interacts with an external electromagnetic field. The atom picks up a photon from the field, which allows it to enter an excited energy state $\psi_{\rm H}$. Unlike spontaneous emission, this process can reasonably be described using nonrelativistic quantum mechanics. The trick is to ignore the photon absorbed from the electromagnetic field. In that case, the electromagnetic field can be

approximated as a known one, using classical electromagnetics. After all, if the field has many photons, one more or less is not going to make a difference.

The third process is stimulated emission. In this case an atom in an excited state ψ_H interacts with an electromagnetic field. And now the atom does not do the logical thing; it does not pick up a photon to go to a still more excited state. Instead it uses the presence of the electromagnetic field as an excuse to dump a photon and return to a lower energy state ψ_L.

This process is the operating principle of lasers. Suppose that you bring a large number of atoms into a relatively stable excited state. Then suppose that one of the atoms performs a spontaneous emission. The photon released by that atom can stimulate another excited atom to release a photon too. Then there are two coherent photons, which can go on to stimulate still more excited atoms to release still more photons. And so on in an avalanche effect. It can produce a runaway process of photon release in which a macroscopic amount of monochromatic, coherent light is created.

Masers work on the same principle, but the radiation is of much lower energy than visible light. It is therefore usually referred to as microwaves instead of light. The ammonia molecule is one possible source of such low energy radiation, chapter 5.3.

The analysis in this section will illuminate some of the details of stimulated emission. For example, it turns out that photon absorption by the lower energy atoms, figure 7.9*(b)*, competes on a perfectly equal footing with stimulated emission, figure 7.9*(c)*. If you have a 50/50 mixture of atoms in the excited state ψ_H and the lower energy state ψ_L, just as many photons will be created by stimulated emission as will be absorbed. So no net light will be produced. To get a laser to work, you must initially have a "population inversion;" you must have more excited atoms than lower energy ones.

(Note that the lower energy state is not necessarily the same as the ground state. All else being the same, it obviously helps to have the lower energy state itself decay rapidly to a state of still lower energy. To a considerable extent, you can pick and choose decay rates, because decay rates can vary greatly depending on the amount to which they are forbidden, section 7.4.)

Key Points

- An electromagnetic field can cause atoms to absorb photons.
- However, it can also cause excited atoms to release photons. That is called stimulated emission.
- In lasers and masers, an avalanche effect of stimulated emission produces coherent, monochromatic light.

7.7.1 The Hamiltonian

To describe the effect of an electromagnetic field on an atom using quantum mechanics, as always the Hamiltonian operator is needed.

The atom will be taken to be a hydrogen atom for simplicity. Since the proton is heavy, the electromagnetic field interacts mainly with the electron. The proton will be assumed to be at rest.

It is also necessary to simplify the electromagnetic field. That can be done by decomposing the field into separate "plane waves." The total interaction can usually be obtained by simply summing the effects produced by the separate waves.

A single plane wave has an electric field $\vec{\mathcal{E}}$ and a magnetic field $\vec{\mathcal{B}}$ that can be written in the form, (13.10):

$$\vec{\mathcal{E}} = \hat{k}\mathcal{E}_\mathrm{f} \cos\Big(\omega(t - y/c) - \alpha\Big) \qquad \vec{\mathcal{B}} = \hat{\imath}\frac{1}{c}\mathcal{E}_\mathrm{f} \cos\Big(\omega(t - y/c) - \alpha\Big)$$

For convenience the y-axis was taken in the direction of propagation of the wave. Also the z-axis was taken in the direction of the electric field. Since there is just a single frequency ω, the wave is monochromatic; it is a single color. And because of the direction of the electric field, the wave is said to be polarized in the z-direction. Note that the electric and magnetic fields for plane waves are normal to the direction of propagation and to each other. The constant c is the speed of light, $\mathcal{E}_\mathrm{f}$ the amplitude of the electric field, and α is some unimportant phase angle.

Fortunately, the expression for the wave can be greatly simplified. The electron reacts primarily to the electric field, provided that its kinetic energy is small compared to its rest mass energy. That is certainly true for the electron in a hydrogen atom and for the outer electrons of atoms in general. Therefore the magnetic field can be ignored. (The error made in doing so is described more precisely in {D.39}.) Also, the wave length of the electromagnetic wave is usually much larger than the size of the atom. For example, the Lyman-transition wave lengths are of the order of a thousand Å, while the atom is about one Å. So, as far as the light wave is concerned, the atom is just a tiny speck at the origin. That means that y can be put to zero in the expression for the plane wave. Then the wave simplifies to just:

$$\vec{\mathcal{E}} = \hat{k}\mathcal{E}_\mathrm{f} \cos(\omega t - \alpha) \tag{7.40}$$

This may not be applicable to highly energetic radiation like X-rays.

Now the question is how this field changes the Hamiltonian of the electron. Ignoring the time dependence of the electric field, that is easy. The Hamiltonian is

$$H = H_\mathrm{atom} + e\mathcal{E}_\mathrm{f} \cos(\omega t - \alpha) z \tag{7.41}$$

where $H_{\rm atom}$ is the Hamiltonian of the hydrogen atom without the external electromagnetic field. The expression for $H_{\rm atom}$ was given in chapter 4.3, but it is not of any interest here.

The interesting term is the second one, the perturbation caused by the electromagnetic field. In this term z is the z-position of the electron. It is just like the mgh potential energy of gravity, with the charge e playing the part of the mass m, the electric field strength $\mathcal{E}_{\rm f}\cos(\omega t - \alpha)$ that of the gravity strength g, and z that of the height h.

To be sure, the electric field is time dependent. The above perturbation potential really assumes that "the electron moves so fast that the field seems steady to it." Indeed, if an electron "speed" is ballparked from its kinetic energy, the electron does seem to travel through the atom relatively fast compared to the frequency of the electric field. Of course, it is much better to write the correct unsteady Hamiltonian and then show it works out pretty much the same as the quasi-steady one above. That is done in {D.39}.

Key Points

- An approximate Hamiltonian was written down for the interaction of an atom with an electromagnetic wave.
- By approximation the atom sees a uniform, quasi-steady electric field.

7.7.2 The two-state model

The big question is how the electromagnetic field affects transitions between a typical atomic state $\psi_{\rm L}$ of lower energy and one of higher energy $\psi_{\rm H}$.

The answer depends critically on various Hamiltonian coefficients. In particular, the expectation values of the energies of the two states are needed. They are

$$E_{\rm L} = \langle \psi_{\rm L} | H | \psi_{\rm L} \rangle \qquad E_{\rm H} = \langle \psi_{\rm H} | H | \psi_{\rm H} \rangle$$

Here the Hamiltonian to use is (7.41) of the previous subsection; it includes the electric field. But it can be seen that the energies are unaffected by the electric field. They are the unperturbed atomic energies of the states. That follows from symmetry; if you write out the inner products above using (7.41), the square wave function is the same at any two positions $\vec{r}$ and $-\vec{r}$, but z in the electric field term changes sign. So integration values pairwise cancel each other.

Note however that the two energies are now expectation values of energy; due to the electric field the atomic states develop uncertainty in energy. That is why they are no longer stationary states.

The other key Hamiltonian coefficient is

$$H_{\rm HL} = \langle \psi_{\rm H} | H | \psi_{\rm L} \rangle$$

Plugging in the Hamiltonian (7.41), it is seen that the atomic part $H_{\rm atom}$ does not contribute. The states $\psi_{\rm H}$ and $\psi_{\rm L}$ are orthogonal, and the atomic Hamiltonian just multiplies $\psi_{\rm L}$ by $E_{\rm L}$. But the electric field gives

$$H_{\rm HL} = \mathcal{E}_{\rm f}\langle\psi_{\rm H}|ez|\psi_{\rm L}\rangle\frac{e^{{\rm i}(\omega t-\alpha)}+e^{-{\rm i}(\omega t-\alpha)}}{2}$$

Here the cosine in (7.41) was taken apart into two exponentials using the Euler formula (2.5).

The next question is what these coeffients mean for the transitions between two atomic states $\psi_{\rm L}$ and $\psi_{\rm H}$. First, since the atomic states are complete, the wave function can always be written as

$$\Psi = c_{\rm L}\psi_{\rm L} + c_{\rm H}\psi_{\rm H} + \ldots$$

where the dots stand for other atomic states. The coefficients $c_{\rm L}$ and $c_{\rm H}$ are the key, because their square magnitudes give the probabilities of the states $\psi_{\rm L}$ and $\psi_{\rm H}$. So they determine whether transitions occur between them.

Evolution equations for these coefficients follow from the Schrödinger equation. The way to find them was described in section 7.6, with additional manipulations in derivation {D.38}. The resulting evolution equations are:

$$\boxed{{\rm i}\hbar\dot{\bar{c}}_{\rm L} = \overline{H}_{\rm LH}\bar{c}_{\rm H} + \ldots \qquad {\rm i}\hbar\dot{\bar{c}}_{\rm H} = \overline{H}_{\rm HL}\bar{c}_{\rm L} + \ldots} \tag{7.42}$$

where the dots represent terms involving states other than $\psi_{\rm L}$ and $\psi_{\rm H}$. These equations use the modified coefficients

$$\bar{c}_{\rm L} = c_{\rm L}e^{{\rm i}E_{\rm L}t/\hbar} \qquad \bar{c}_{\rm H} = c_{\rm H}e^{{\rm i}E_{\rm H}t/\hbar} \tag{7.43}$$

The modified coefficients have the same square magnitudes as the original ones and the same values at time zero. That makes them fully equivalent to the original ones. The modified Hamiltonian coefficient in the evolution equations is

$$\overline{H}_{\rm HL} = \overline{H}^*_{\rm LH} = \tfrac{1}{2}\mathcal{E}_{\rm f}\langle\psi_{\rm H}|ez|\psi_{\rm L}\rangle e^{{\rm i}(\omega_0-\omega)t+\alpha} + \tfrac{1}{2}\mathcal{E}_{\rm f}\langle\psi_{\rm H}|ez|\psi_{\rm L}\rangle e^{{\rm i}(\omega_0+\omega)t-\alpha} \tag{7.44}$$

where ω_0 is the frequency of a photon that has the exact energy $E_{\rm H} - E_{\rm L}$.

Note that this modified Hamiltonian coefficient is responsible for the interaction between the states $\psi_{\rm L}$ and $\psi_{\rm H}$. If this Hamiltonian coefficient is zero, the electromagnetic wave cannot cause transitions between the two states. At least not within the approximations made.

Whether this happens depends on whether the inner product $\langle\psi_{\rm H}|ez|\psi_{\rm L}\rangle$ is zero. This inner product is called the "atomic matrix element" because it depends only on the atomic states, not on the strength and frequency of the electric wave.

However, it does depend on the direction of the electric field. The assumed plane wave had its electric field in the z-direction. Different waves can have their electric fields in other directions. Therefore, waves can cause transitions as long as there is at least one nonzero atomic matrix element of the form $\langle \psi_{\rm L} | e r_i | \psi_{\rm H} \rangle$, with r_i equal to x, y, or z. If there is such a nonzero matrix element, the transition is called allowed. Conversely, if all three matrix elements are zero, then transitions between the states $\psi_{\rm L}$ and $\psi_{\rm H}$ are called forbidden.

Note however that so-called forbidden transitions often occur just fine. The derivation in the previous subsection made several approximations, including that the magnetic field can be ignored and that the electric field is independent of position. If these ignored effects are corrected for, many forbidden transitions turn out to be possible after all; they are just much slower.

The approximations made to arrive at the atomic matrix element $\langle \psi_{\rm H} | ez | \psi_{\rm L} \rangle$ are known as the "electric dipole approximation." The corresponding transitions are called "electric dipole transitions." If you want to know where the term comes from, why? Anyway, in that case note first that if the electron charge distribution is symmetric around the proton, the expectation value of ez will be zero by symmetry. Negative z values will cancel positive ones. But the electron charge distribution might get somewhat shifted to the positive z side, say. The total atom is then still electrically neutral, but it behaves a bit like a combination of a negative charge at a positive value of z and an equal and opposite positive charge at a negative value of z. Such a combination of two opposite charges is called a dipole in classical electromagnetics, chapter 13.3. So in quantum mechanics the operator ez gives the dipole strength in the z-direction. And if the above atomic matrix element is nonzero, it can be seen that nontrivial combinations of $\psi_{\rm L}$ and $\psi_{\rm H}$ have a nonzero expectation dipole strength. So the name "electric dipole transitions" is justified, especially since "basic electric transitions" would be understandable by far too many nonexperts.

Allowed and forbidden transitions were discussed earlier in section 7.4. However, that was based on assumed properties of the emitted photon. The allowed atomic matrix elements above, and similar forbidden ones, make it possible to check the various most important results directly from the governing equations. That is done in derivation {D.39}.

There is another requirement to get a decent transition probability. The exponentials in the modified Hamiltonian coefficient (7.44) must not oscillate too rapidly in time. Otherwise opposite values of the exponentials will average away against each other. So no significant transition probability can build up. (This is similar to the cancelation that gives rise to the adiabatic theorem, {D.34}.) Now under real-life conditions, the second exponential in (7.44) will always oscillate rapidly. Normal electromagnetic frequencies are very high. Therefore the second term in (7.44) can normally be ignored.

And in order for the first exponential not too oscillate too rapidly requires a pretty good match between the frequencies ω and ω_0. Recall that ω is the

frequency of the electromagnetic wave, while ω_0 is the frequency of a photon whose energy is the difference between the atomic energies E_H and E_L. If the electric field does not match the frequency of that photon, it will not do much. Using the Planck-Einstein relation, that means that

$$\omega \approx \omega_0 \equiv (E_H - E_L)/\hbar$$

One consequence is that in transitions between two atomic states ψ_L and ψ_H, other states usually do not need to be considered. Unless an other state matches either the energy E_H or E_L, it will give rise to rapidly oscillating exponentials that can be ignored.

In addition, the interest is often in the so-called collision-dominated regime in which the atom evolves for only a short time before being disturbed by "collisions" with its surroundings. In that case, the short evolution time prevents nontrivial interactions between different transition processes to build up. Transition rates for the individual transition processes can be found separately and simply added together.

The obtained evolution equations (7.42) can explain why absorption and stimulated emission compete on an equal footing in the operation of lasers. The reason is that the equations have a remarkable symmetry: for every solution $\bar{c}_L$, $\bar{c}_H$ there is a second solution $\bar{c}_{L,2} = \bar{c}_H^*$, $\bar{c}_{H,2} = -\bar{c}_L^*$ that has the probabilities of the low and high energy states exactly reversed. It means that

> *An electromagnetic field that takes an atom out of the low energy state ψ_L towards the high energy state ψ_H will equally take that atom out of the high energy state ψ_H towards the low energy state ψ_L.*

It is a consequence of the Hermitian nature of the Hamiltonian; it would not apply if $\overline{H}_{LH}$ was not equal to $\overline{H}_{HL}^*$.

Key Points

- The governing evolution equations for the probabilities of two atomic states ψ_L and ψ_H in an electromagnetic wave have been found.
- The equations have a symmetry property that makes electromagnetic waves equally effective for absorption and stimulated emission.
- Normally the electromagnetic field has no significant effect on transitions between the states unless its frequency ω closely matches the frequency ω_0 of a photon with energy $E_H - E_L$.
- The governing equations can explain why some transitions are allowed and others are forbidden. The key are so-called "atomic matrix elements."

7.8 General Interaction with Radiation

Under typical conditions, a collection of atoms is not just subjected to a single electromagnetic wave, as described in the previous section, but to "broadband" incoherent radiation of all frequencies moving in all directions. Also, the interactions of the atoms with their surroundings tend to be rare compared to the frequency of the radiation but frequent compared to the typical life time of the various excited atomic states. In other words, the evolution of the atomic states is collision-dominated. The question in this subsection is what can be said about the emission and absorption of radiation by the atoms under such conditions.

Since both the electromagnetic field and the collisions are random, a statistical rather than a determinate treatment is needed. In it, the probability that a randomly chosen atom can be found in a typical atomic state $\psi_{\rm L}$ of low energy will be called $P_{\rm L}$. Similarly, the probability that an atom can be found in an atomic state $\psi_{\rm H}$ of higher energy will be called $P_{\rm H}$. More simplistic, $P_{\rm L}$ can be called the fraction of atoms in the low energy state and $P_{\rm H}$ the fraction in the high energy state.

The energy of the electromagnetic radiation, per unit volume and per unit frequency range, will be indicated by $\rho(\omega)$. The particular frequency ω_0 that is relevant to transitions between two atomic states $\psi_{\rm L}$ and $\psi_{\rm H}$ is related to the energy difference between the states. In particular,

$$\omega_0 = (E_{\rm H} - E_{\rm L})/\hbar$$

is the nominal frequency of the photon released or absorbed in a transition between the two states.

In those terms, the fractions $P_{\rm L}$ and $P_{\rm H}$ of atoms in the two states evolve in time according to the evolution equations, {D.41},

$$\boxed{\frac{{\rm d}P_{\rm L}}{{\rm d}t} = -B_{{\rm L}\rightarrow{\rm H}}\rho(\omega_0)\,P_{\rm L} + B_{{\rm H}\rightarrow{\rm L}}\rho(\omega_0)\,P_{\rm H} + A_{{\rm H}\rightarrow{\rm L}}\,P_{\rm H} + \ldots} \tag{7.45}$$

$$\boxed{\frac{{\rm d}P_{\rm H}}{{\rm d}t} = +B_{{\rm L}\rightarrow{\rm H}}\rho(\omega_0)\,P_{\rm L} - B_{{\rm H}\rightarrow{\rm L}}\rho(\omega_0)\,P_{\rm H} - A_{{\rm H}\rightarrow{\rm L}}\,P_{\rm H} + \ldots} \tag{7.46}$$

In the first equation, the first term in the right hand side reflects atoms that are excited from the low energy state to the high energy state. That decreases the number of low energy atoms, explaining the minus sign. The effect is of course proportional to the fraction $P_{\rm L}$ of low energy atoms that is available to be excited. It is also proportional to the energy $\rho(\omega_0)$ of the electromagnetic waves that do the actual exciting.

Similarly, the second term in the right hand side of the first equation reflects the fraction of low energy atoms that is created through de-excitation of excited

atoms by the electromagnetic radiation. The final term reflects the low energy atoms created by spontaneous decay of excited atoms. The constant $A_{\mathrm{H}\to\mathrm{L}}$ is the spontaneous emission rate. (It is really the decay rate λ as defined earlier in section 7.5.3, but in the present context the term spontaneous emission rate and symbol A tend to be used.)

The second equation can be understood similarly as the first. If there are transitions with states other than ψ_{L} and ψ_{H}, all their effects should be summed together; that is indicated by the dots in (7.45) and (7.46).

The constants in the equations are collectively referred to as the "Einstein A and B coefficients." Imagine that some big shot in engineering was too lazy to select appropriate symbols for the quantities used in a paper and just called them A and B. Referees and standards committees would be on his/her back, big shot or not. However, in physics they still stick with the stupid symbols almost a century later. At least in this context.

Anyway, the B coefficients are, {D.41},

$$\boxed{B_{\mathrm{L}\to\mathrm{H}} = B_{\mathrm{H}\to\mathrm{L}} = \frac{\pi}{\hbar^2\epsilon_0}\frac{|\langle\psi_{\mathrm{L}}|e\vec{r}|\psi_{\mathrm{H}}\rangle|^2}{3}} \tag{7.47}$$

Here $\epsilon_0 = 8.854\,19\ 10^{-12}$ C^2/J m is the permittivity of space. Note from the appearance of the Planck constant that the emission and absorption of radiation is truly a quantum effect. The second ratio is the average atomic matrix element discussed in the previous section. The fact that $B_{\mathrm{L}\to\mathrm{H}}$ equals $B_{\mathrm{H}\to\mathrm{L}}$ reflects that the electric field is equally effective for absorption as for stimulated emission. It is a consequence of the symmetry property of two-state systems mentioned in the previous section.

The spontaneous emission rate was found by Einstein using a dirty trick, {D.42}. It is

$$\boxed{A_{\mathrm{H}\to\mathrm{L}} = B_{\mathrm{H}\to\mathrm{L}}\rho_{\mathrm{equiv}}(\omega_0) \qquad \rho_{\mathrm{equiv}}(\omega) = \frac{\hbar\omega^3}{\pi^2c^3}} \tag{7.48}$$

One way of thinking of the mechanism of spontaneous emission is that it is an effect of the ground state electromagnetic field. Just like normal particle systems still have nonzero energy left in their ground state, so does the electromagnetic field. You could therefore think of this ground state electromagnetic field as the source of the atomic perturbations that cause the atomic decay. If that picture is right, then the term ρ_{equiv} in the expression above should be the energy of the field in the ground state. In terms of the analysis of chapter 6.8, that would mean that in the ground state, there is exactly one photon left in each radiation mode. Just drop the factor (6.10) from (6.11).

It is a pretty reasonable description, but it is not quite true. In the ground state of the electromagnetic field there is half a photon in each mode, not one.

It is just like a harmonic oscillator, which has half an energy quantum $\hbar\omega$ left in its ground state, chapter 4.1. Also, a ground state energy should not make a difference for the evolution of a system. Instead, because of a twilight effect, the photon that the excited atom interacts with is the one that it will emit, addendum {A.24}.

As a special example of the given evolution equations, consider a closed box whose inside is at absolute zero temperature. Then there is no ambient blackbody radiation, $\rho = 0$. Now assume that initially there is a thin gas of atoms in the box in an excited state $\psi_{\rm H}$. These atoms will decay to whatever are the available atomic states of lower energy. In particular, according to (7.46) the fraction $P_{\rm H}$ of excited atoms left will evolve as

$$\frac{{\rm d}P_{\rm H}}{{\rm d}t} = -\left[A_{{\rm H}\to{\rm L}_1} + A_{{\rm H}\to{\rm L}_2} + A_{{\rm H}\to{\rm L}_3} + \ldots\right] P_{\rm H}$$

where the sum is over all the lower energy states that exist. It describes the effect of all possible spontaneous emission processes that the excited state is subject to. (The above equation is a rewrite of (7.28) of section 7.5.3 in the present notations.)

The above expression assumed that the excited atoms are in a box that is at absolute zero temperature. Atoms in a box that is at room temperature are bathed in thermal blackbody radiation. In principle you would then have to use the full equations (7.45) and (7.46) to figure out what happens to the number of excited atoms. Stimulated emission will add to spontaneous emission and new excited atoms will be created by absorption. However, at room temperature blackbody radiation has negligible energy in the visible light range, chapter 6.8 (6.10). Transitions in this range will not really be affected.

Key Points

- This section described the general evolution equations for a system of atoms in an incoherent ambient electromagnetic field.
- The constants in the equations are called the Einstein A and B coefficients.
- The B coefficients describe the relative response of transitions to incoherent radiation. They are given by (7.47).
- The A coefficients describe the spontaneous emission rate. They are given by (7.48).

7.9 Position and Linear Momentum

The subsequent sections will be looking at the time evolution of various quantum systems, as predicted by the Schrödinger equation. However, before that can

be done, first the eigenfunctions of position and linear momentum must be found. That is something that the book has been studiously avoiding so far. The problem is that the position and linear momentum eigenfunctions have awkward issues with normalizing them.

These normalization problems have consequences for the coefficients of the eigenfunctions. In the orthodox interpretation, the square magnitudes of the coefficients should give the probabilities of getting the corresponding values of position and linear momentum. But this statement will have to be modified a bit.

One good thing is that unlike the Hamiltonian, which is specific to a given system, the position operator

$$\widehat{\vec{r}} = (\widehat{x}, \widehat{y}, \widehat{z})$$

and the linear momentum operator

$$\widehat{\vec{p}} = (\widehat{p}_x, \widehat{p}_y, \widehat{p}_z) = \frac{\hbar}{\mathrm{i}} \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

are the same for all systems. So, you only need to find their eigenfunctions once.

7.9.1 The position eigenfunction

The eigenfunction that corresponds to the particle being at a precise x-position $\underline{x}$, y-position $\underline{y}$, and z-position $\underline{z}$ will be denoted by $R_{\underline{x}\underline{y}\underline{z}}(x, y, z)$. The eigenvalue problem is:

$$\widehat{x} R_{\underline{x}\underline{y}\underline{z}}(x, y, z) = \underline{x} R_{\underline{x}\underline{y}\underline{z}}(x, y, z)$$

$$\widehat{y} R_{\underline{x}\underline{y}\underline{z}}(x, y, z) = \underline{y} R_{\underline{x}\underline{y}\underline{z}}(x, y, z)$$

$$\widehat{z} R_{\underline{x}\underline{y}\underline{z}}(x, y, z) = \underline{z} R_{\underline{x}\underline{y}\underline{z}}(x, y, z)$$

(Note the need in this analysis to use $(\underline{x}, \underline{y}, \underline{z})$ for the measurable particle position, since (x, y, z) are already used for the eigenfunction arguments.)

To solve this eigenvalue problem, try again separation of variables, where it is assumed that $R_{\underline{x}\underline{y}\underline{z}}(x, y, z)$ is of the form $X(x)Y(y)Z(z)$. Substitution gives the partial problem for X as

$$xX(x) = \underline{x} X(x)$$

This equation implies that at all points x not equal to $\underline{x}$, $X(x)$ will have to be zero, otherwise there is no way that the two sides can be equal. So, function $X(x)$ can only be nonzero at the single point $\underline{x}$. At that one point, it can be anything, though.

To resolve the ambiguity, the function $X(x)$ is taken to be the "Dirac delta function,"

$$X(x) = \delta(x - \underline{x})$$

The delta function is, loosely speaking, sufficiently strongly infinite at the single point $x = \underline{x}$ that its integral over that single point is one. More precisely, the delta function is defined as the limiting case of the function shown in the left hand side of figure 7.10.

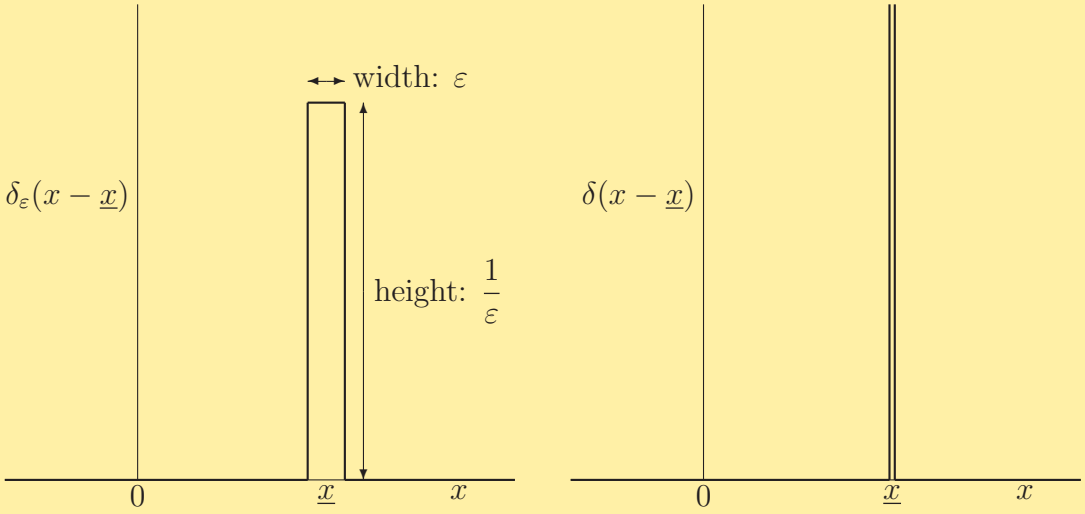


Figure 7.10: Approximate Dirac delta function $\delta_\varepsilon(x - \underline{x})$ is shown left. The true delta function $\delta(x - \underline{x})$ is the limit when ε becomes zero, and is an infinitely high, infinitely thin spike, shown right. It is the eigenfunction corresponding to a position $\underline{x}$.

The fact that the integral is one leads to a very useful mathematical property of delta functions: they are able to pick out one specific value of any arbitrary given function $f(x)$. Just take an inner product of the delta function $\delta(x - \underline{x})$ with $f(x)$. It will produce the value of $f(x)$ at the point $\underline{x}$, in other words, $f(\underline{x})$:

$$\langle \delta(x - \underline{x}) | f(x) \rangle = \int_{x=-\infty}^{\infty} \delta(x - \underline{x}) f(x)\, \mathrm{d}x = \int_{x=-\infty}^{\infty} \delta(x - \underline{x}) f(\underline{x})\, \mathrm{d}x = f(\underline{x}) \tag{7.49}$$

(Since the delta function is zero at all points except $\underline{x}$, it does not make a difference whether $f(x)$ or $f(\underline{x})$ is used in the integral.) This is sometimes called the "filtering property" of the delta function.

The problems for the position eigenfunctions Y and Z are the same as the one for X, and have a similar solution. The complete eigenfunction corresponding to a measured position $(\underline{x}, \underline{y}, \underline{z})$ is therefore:

$$\boxed{R_{\underline{x}\underline{y}\underline{z}}(x, y, z) = \delta(x - \underline{x})\delta(y - \underline{y})\delta(z - \underline{z}) \equiv \delta^3(\vec{r} - \underline{\vec{r}})} \tag{7.50}$$

Here $\delta^3(\vec{r} - \underline{\vec{r}})$ is the three-dimensional delta function, a spike at position $\underline{\vec{r}}$ whose volume integral equals one.

According to the orthodox interpretation, the probability of finding the particle at $(\underline{x}, \underline{y}, \underline{z})$ for a given wave function Ψ should be the square magnitude of the coefficient $c_{\underline{x}\underline{y}\underline{z}}$ of the eigenfunction. This coefficient can be found as an inner product:

$$c_{\underline{x}\underline{y}\underline{z}}(t) = \langle \delta(x - \underline{x})\delta(y - \underline{y})\delta(z - \underline{z}) | \Psi \rangle$$

It can be simplified to

$$c_{\underline{x}\underline{y}\underline{z}}(t) = \Psi(\underline{x}, \underline{y}, \underline{z}; t) \tag{7.51}$$

because of the property of the delta functions to pick out the corresponding function value.

However, the apparent conclusion that $|\Psi(\underline{x}, \underline{y}, \underline{z}; t)|^2$ gives the probability of finding the particle at $(\underline{x}, \underline{y}, \underline{z})$ is wrong. The reason it fails is that eigenfunctions should be normalized; the integral of their square should be one. The integral of the square of a delta function is infinite, not one. That is OK, however; $\vec{r}$ is a continuously varying variable, and the chances of finding the particle at $(\underline{x}, \underline{y}, \underline{z})$ to an *infinite* number of digits accurate would be zero. So, the properly normalized eigenfunctions would have been useless anyway.

Instead, according to Born's statistical interpretation of chapter 3.1, the expression

$$|\Psi(x, y, z; t)|^2 \, \mathrm{d}x\mathrm{d}y\mathrm{d}z$$

gives the probability of finding the particle in an infinitesimal volume $\mathrm{d}x\mathrm{d}y\mathrm{d}z$ around (x, y, z). In other words, $|\Psi(x, y, z; t)|^2$ gives the probability of finding the particle *near* location (x, y, z) *per unit volume.* (The underlines below the position coordinates are no longer needed to avoid ambiguity and have been dropped.)

Besides the normalization issue, another idea that needs to be somewhat modified is a strict collapse of the wave function. Any position measurement that can be done will leave some uncertainty about the precise location of the particle: it will leave $\Psi(x, y, z; t)$ nonzero over a small range of positions, rather than just one position. Moreover, unlike energy eigenstates, position eigenstates are not stationary: after a position measurement, Ψ will again spread out as time increases.

Key Points

- Position eigenfunctions are delta functions.
- They are not properly normalized.
- The coefficient of the position eigenfunction for a position (x, y, z) is the good old wave function $\Psi(x, y, z; t)$.

- Because of the fact that the delta functions are not normalized, the square magnitude of $\Psi(x,y,z;t)$ does not give the probability that the particle is at position (x,y,z).
- Instead the square magnitude of $\Psi(x,y,z;t)$ gives the probability that the particle is near position (x,y,z) per unit volume.
- Position eigenfunctions are not stationary, so localized particle wave functions will spread out over time.

7.9.2 The linear momentum eigenfunction

Turning now to linear momentum, the eigenfunction that corresponds to a precise linear momentum (p_x, p_y, p_z) will be indicated as $P_{p_x p_y p_z}(x,y,z)$. If you again assume that this eigenfunction is of the form $X(x)Y(y)Z(z)$, the partial problem for X is found to be:

$$\frac{\hbar}{\mathrm{i}}\frac{\partial X(x)}{\partial x} = p_x X(x)$$

The solution is a complex exponential:

$$X(x) = Ae^{\mathrm{i}p_x x/\hbar}$$

where A is a constant.

Just like the position eigenfunction earlier, the linear momentum eigenfunction has a normalization problem. In particular, since it does not become small at large $|x|$, the integral of its square is infinite, not one. The solution is to ignore the problem and to just take a nonzero value for A; the choice that works out best is to take:

$$A = \frac{1}{\sqrt{2\pi\hbar}}$$

(However, other books, in particular nonquantum ones, are likely to make a different choice.)

The problems for the y and z linear momenta have similar solutions, so the full eigenfunction for linear momentum takes the form:

$$\boxed{P_{p_x p_y p_z}(x,y,z) = \frac{1}{\sqrt{2\pi\hbar}^3} e^{\mathrm{i}(p_x x + p_y y + p_z z)/\hbar}} \tag{7.52}$$

The coefficient $c_{p_x p_y p_z}(t)$ of the momentum eigenfunction is very important in quantum analysis. It is indicated by the special symbol $\Phi(p_x, p_y, p_z; t)$ and called the "momentum space wave function." Like all coefficients, it can be found by taking an inner product of the eigenfunction with the wave function:

$$\boxed{\Phi(p_x, p_y, p_z; t) = \frac{1}{\sqrt{2\pi\hbar}^3} \langle e^{\mathrm{i}(p_x x + p_y y + p_z z)/\hbar} | \Psi \rangle} \tag{7.53}$$

The momentum space wave function does not quite give the probability for the momentum to be (p_x, p_y, p_z). Instead it turns out that

$$|\Phi(p_x, p_y, p_z; t)|^2 \mathrm{d}p_x \mathrm{d}p_y \mathrm{d}p_z$$

gives the probability of finding the linear momentum within a small momentum range $\mathrm{d}p_x\mathrm{d}p_y\mathrm{d}p_z$ around (p_x, p_y, p_z). In other words, $|\Phi(p_x, p_y, p_z; t)|^2$ gives the probability of finding the particle with a momentum near (p_x, p_y, p_z) per unit "momentum space volume." That is much like the square magnitude $|\Psi(x, y, z; t)|^2$ of the normal wave function gives the probability of finding the particle near location (x, y, z) per unit physical volume. The momentum space wave function Φ is in the momentum space (p_x, p_y, p_z) what the normal wave function Ψ is in the physical space (x, y, z).

There is even an inverse relationship to recover Ψ from Φ, and it is easy to remember:

$$\boxed{\Psi(x, y, z; t) = \frac{1}{\sqrt{2\pi\hbar}^3} \langle e^{-\mathrm{i}(p_x x + p_y y + p_z z)/\hbar} | \Phi \rangle_{\vec{p}}} \tag{7.54}$$

where the subscript on the inner product indicates that the integration is over momentum space rather than physical space.

If this inner product is written out, it reads:

$$\boxed{\Psi(x, y, z; t) = \frac{1}{\sqrt{2\pi\hbar}^3} \iiint_{\text{all } \vec{p}} \Phi(p_x, p_y, p_z; t)\, e^{\mathrm{i}(p_x x + p_y y + p_z z)/\hbar}\, \mathrm{d}p_x \mathrm{d}p_y \mathrm{d}p_z} \tag{7.55}$$

Mathematicians prove this formula under the name "Fourier Inversion Theorem", {A.26}. But it really is just the same sort of idea as writing Ψ as a sum of eigenfunctions ψ_n times their coefficients c_n, as in $\Psi = \sum_n c_n \psi_n$. In this case, the coefficients are given by Φ and the eigenfunctions by the exponential (7.52). The only real difference is that the sum has become an integral since $\vec{p}$ has continuous values, not discrete ones.

Key Points

- The linear momentum eigenfunctions are complex exponentials of the form:
$$\frac{1}{\sqrt{2\pi\hbar}^3} e^{\mathrm{i}(p_x x + p_y y + p_z z)/\hbar}$$
- They are not properly normalized.
- The coefficient of the linear momentum eigenfunction for a momentum (p_x, p_y, p_z) is indicated by $\Phi(p_x, p_y, p_z; t)$. It is called the momentum space wave function.
- Because of the fact that the momentum eigenfunctions are not normalized, the square magnitude of $\Phi(p_x, p_y, p_z; t)$ does not give the probability that the particle has momentum (p_x, p_y, p_z).

- Instead the square magnitude of $\Phi(p_x, p_y, p_z; t)$ gives the probability that the particle has a momentum close to (p_x, p_y, p_z) per unit momentum space volume.
- In writing the complete wave function in terms of the momentum eigenfunctions, you must integrate over the momentum instead of sum.
- The transformation between the physical space wave function Ψ and the momentum space wave function Φ is called the Fourier transform. It is invertible.

7.10 Wave Packets

This section gives a full description of the motion of a particle according to quantum mechanics. It will be assumed that the particle is in free space, so that the potential energy is zero. In addition, to keep the analysis concise and the results easy to graph, it will be assumed that the motion is only in the x-direction. The results may easily be extended to three dimensions by using separation of variables.

One thing that the analysis will show is how limiting the uncertainty in both momentum and position produces the various features of classical Newtonian motion. It may be recalled that in Newtonian motion through free space, the linear momentum p is constant. In addition, since p/m is the velocity v, the classical particle will move at constant speed. So classical Newtonian motion would say:

$$v = \frac{p}{m} = \text{constant} \qquad x = vt + x_0 \qquad \text{for Newtonian motion in free space}$$

(Note that p is used to indicate p_x in this and the following sections.)

7.10.1 Solution of the Schrödinger equation.

As discussed in section 7.1, the unsteady evolution of a quantum system may be determined by finding the eigenfunctions of the Hamiltonian and giving them coefficients that are proportional to $e^{-\mathrm{i}Et/\hbar}$. This will be worked out in this subsection.

For a free particle, there is only kinetic energy, so in one dimension the Hamiltonian eigenvalue problem is:

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \psi}{\partial x^2} = E\psi \tag{7.56}$$

Solutions to this equation take the form of exponentials

$$\psi_E = Ae^{\pm \mathrm{i}\sqrt{2mE}x/\hbar}$$

where A is a constant.

Note that E must be positive: if the square root would be imaginary, the solution would blow up exponentially at large positive or negative x. Since the square magnitude of ψ at a point gives the probability of finding the particle near that position, blow up at infinity would imply that the particle must be at infinity with certainty.

The energy eigenfunction above is really the same as the eigenfunction of the x-momentum operator $\widehat{p}_x$ derived in the previous section:

$$\psi_E = \frac{1}{\sqrt{2\pi\hbar}} e^{\mathrm{i}px/\hbar} \quad \text{with } p = \pm\sqrt{2mE} \tag{7.57}$$

The reason that the momentum eigenfunctions are also energy eigenfunctions is that the energy is all kinetic energy, and the kinetic operator equals $\widehat{T} = \widehat{p}^2/2m$. So eigenfunctions with precise momentum p have precise energy $p^2/2m$.

As shown by (7.55) in the previous section, combinations of momentum eigenfunctions take the form of an integral rather than a sum. In the one-dimensional case that integral is:

$$\Psi(x,t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \Phi(p,t) e^{\mathrm{i}px/\hbar} \,\mathrm{d}p$$

where $\Phi(p,t)$ is called the momentum space wave function.

Whether a sum or an integral, the Schrödinger equation still requires that the coefficient of each energy eigenfunction varies in time proportional to $e^{-\mathrm{i}Et/\hbar}$. The coefficient here is the momentum space wave function Φ, and the energy is $E = p^2/2m$, so the solution of the Schrödinger equation must be:

$$\boxed{\Psi(x,t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \Phi_0(p)\, e^{\mathrm{i}p(x-\frac{p}{2m}t)/\hbar} \,\mathrm{d}p} \tag{7.58}$$

Here $\Phi_0(p) \equiv \Phi(p,0)$ is determined by whatever initial conditions are relevant to the situation that is to be described. The above integral is the final solution for a particle in free space.

Key Points

- In free space, momentum eigenfunctions are also energy eigenfunctions.
- The one-dimensional wave function for a particle in free space is given by (7.58).
- The function Φ_0 is still to be chosen to produce whatever physical situation is to be described.

7.10.2 Component wave solutions

Before trying to interpret the complete obtained solution (7.58) for the wave function of a particle in free space, it is instructive first to have a look at the component solutions, defined by

$$\psi_{\rm w} \equiv e^{{\rm i}p(x-\frac{p}{2m}t)/\hbar} \tag{7.59}$$

These solutions will be called component waves; both their real and imaginary parts are sinusoidal, as can be seen from the Euler formula (2.5).

$$\psi_{\rm w} = \cos\Big(p\Big(x-\frac{p}{2m}t\Big)/\hbar\Big) + {\rm i}\sin\Big(p\Big(x-\frac{p}{2m}t\Big)/\hbar\Big)$$

In figure 7.11, the real part of the wave (in other words, the cosine), is sketched as the red curve; also the magnitude of the wave (which is unity) is shown as the top black line, and minus the magnitude is drawn as the bottom black line. The black lines enclose the real part of the wave, and will be called

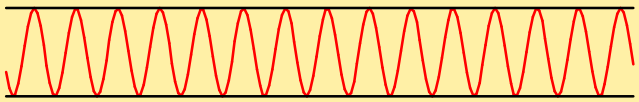


Figure 7.11: The real part (red) and envelope (black) of an example wave.

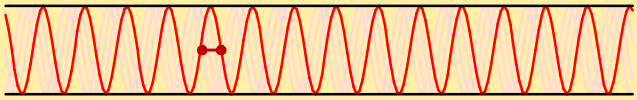
the "envelope." Since their vertical separation is twice the magnitude of the wave function, the vertical separation between the black lines at a point is a measure for the probability of finding the particle near that point.

The constant separation between the black lines shows that there is absolutely no localization of the particle to any particular region. The particle is equally likely to be found at every point in the infinite range. This also graphically demonstrates the normalization problem of the momentum eigenfunctions discussed in the previous section: the total probability of finding the particle just keeps getting bigger and bigger, the larger the range you look in. So there is no way that the total probability of finding the particle can be limited to one as it should be.

The reason for the complete lack of localization is the fact that the component wave solutions have an exact momentum p. With zero uncertainty in momentum, Heisenberg's uncertainty relationship says that there must be infinite uncertainty in position. There is.

There is another funny thing about the component waves: when plotted for different times, it is seen that the real part of the wave moves towards the right with a speed $p/2m = \frac{1}{2}v$, as illustrated in figure 7.12.

This is unexpected, because classically the particle moves with speed v, not $\frac{1}{2}v$. The problem is that the speed with which the wave moves, called the "phase speed," is not meaningful physically. In fact, without anything like a location



Animation: http://www.dommelen.net/quansup/wavemv.gif

Figure 7.12: The wave moves with the phase speed.

for the particle, there is no way to define a physical velocity for a component wave.

Key Points

- Component waves provide no localization of the particle at all.
- Their real part is a moving cosine. Similarly their imaginary part is a moving sine.
- The speed of motion of the cosine or sine is half the speed of a classical particle with that momentum.
- This speed is called the phase speed and is not relevant physically.

7.10.3 Wave packets

As Heisenberg's principle indicates, in order to get some localization of the position of a particle, some uncertainty must be allowed in momentum. That means that you must take the initial momentum space wave function Φ_0 in (7.58) to be nonzero over at least some small *interval* of different momentum values p. Such a combination of component waves is called a "wave packet".

The wave function for a typical wave packet is sketched in figure 7.13. The red line is again the real part of the wave function, and the black lines are the envelope enclosing the wave; they equal plus and minus the magnitude of the wave function.

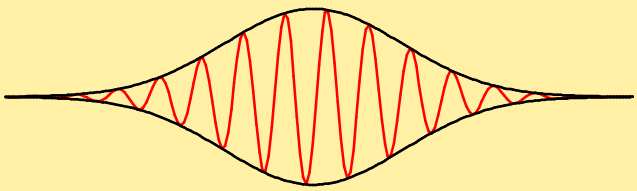


Figure 7.13: The real part (red) and magnitude or envelope (black) of a wave packet. (Schematic).

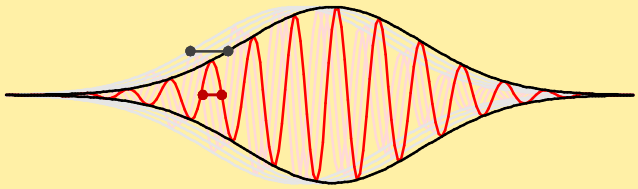
The vertical separation between the black lines is again a measure of the probability of finding the particle near that location. It is seen that the possible

locations of the particle are now restricted to a finite region, the region in which the vertical distance between the black lines is nonzero.

If the envelope changes location with time, and it does, then so does the region where the particle can be found. This then finally is the correct picture of motion in quantum mechanics: the region in which the particle can be found propagates through space.

The limiting case of the motion of a macroscopic Newtonian point mass can now be better understood. As noted in section 7.2.1, for such a particle the uncertainty in position is negligible. The wave packet in which the particle can be found, as sketched in figure 7.13, is so small that it can be considered to be a point. To that approximation the particle then has a point position, which is the normal classical description.

The classical description also requires that the particle moves with velocity $u = p/m$, which is twice the speed $p/2m$ of the wave. So the envelope should move twice as fast as the wave. This is indicated in figure 7.14 by the length of the bars, which show the motion of a point on the envelope and of a point on the wave during a small time interval.



Animation: http://www.dommelen.net/quansup/packetmv.gif

Figure 7.14: The velocities of wave and envelope are not equal.

That the envelope does indeed move at speed p/m can be seen if you define the representative position of the envelope to be the expectation value of position. That position must be somewhere in the middle of the wave packet. The expectation value of position moves according to Ehrenfest's theorem of section 7.2.1 with a speed $\langle p \rangle / m$, where $\langle p \rangle$ is the expectation value of momentum, which must be constant since there is no force. Since the uncertainty in momentum is small for a macroscopic particle, the expectation value of momentum $\langle p \rangle$ can be taken to be "the" momentum p.

Key Points

- A wave packet is a combination of waves with about the same momentum.
- Combining waves into wave packets can provide localization of particles.
- The envelope of the wave packet shows the region where the particle is likely to be found.

0⊢ This region propagates with the classical particle velocity.

7.10.4 Group velocity

As the previous subsection explained, particle motion in classical mechanics is equivalent to the motion of wave packets in quantum mechanics. Motion of a wave packet implies that the region in which the particle can be found changes position.

Motion of wave packets is not just important for understanding where particles in free space end up. It is also critical for the quantum mechanics of for example solids, in which electrons, photons, and phonons (quanta of crystal vibrations) move around in an environment that is cluttered with other particles. And it is also of great importance in classical applications, such as acoustics in solids and fluids, water waves, stability theory of flows, electromagnetodynamics, etcetera. This section explains how wave packets move in such more general systems. Only the one-dimensional case will be considered, but the generalization to three dimensions is straightforward.

The systems of interest have component wave solutions of the general form:

$$\boxed{\text{component wave:} \quad \psi_{\rm w} = e^{{\rm i}(kx-\omega t)}} \tag{7.60}$$

The constant k is called the "wave number," and ω the "angular frequency." The wave number and frequency must be real for the analysis in this section to apply. That means that the magnitude of the component waves must not change with space nor time. Such systems are called nondissipative: although a combination of waves may get dispersed over space, its square magnitude integral will be conserved. (This is true on account of Parseval's relation, {A.26}.)

For a particle in free space according to the previous subsection:

$$k = \frac{p}{\hbar} \qquad \omega = \frac{p^2}{2m\hbar}$$

Therefore, for a particle in free space the wave number k is just a rescaled linear momentum, and the frequency ω is just a rescaled kinetic energy. This will be different for a particle in a nontrivial surroundings.

Regardless of what kind of system it is, the relationship between the frequency and the wave number is called the

$$\boxed{\text{dispersion relation:} \quad \omega = \omega(k)} \tag{7.61}$$

It really defines the physics of the wave propagation.

Since the waves are of the form $e^{{\rm i}k(x-\frac{\omega}{k}t)}$, the wave is constant if $x = (\omega/k)t$ plus any constant. Such points move with the

$$\boxed{\text{phase velocity:} \quad v_{\rm p} \equiv \frac{\omega}{k}} \tag{7.62}$$

In free space, the phase velocity is half the classical velocity.

However, as noted in the previous subsection, wave packets do not normally move with the phase velocity. The velocity that they do move with is called the "group velocity." For a particle in free space, you can infer that the group velocity is the same as the classical velocity from Ehrenfest's theorem, but that does not work for more general systems. The approach will therefore be to simply define the group velocity as

$$\boxed{\text{group velocity:} \quad v_{\rm g} \equiv \frac{{\rm d}\omega}{{\rm d}k}} \tag{7.63}$$

and then to explore how the so-defined group velocity relates to the motion of wave packets.

Wave packets are combinations of component waves, and the most general combination of waves takes the form

$$\boxed{\Psi(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{\Phi}_0(k) e^{{\rm i}(kx-\omega t)} \, {\rm d}k} \tag{7.64}$$

Here $\overline{\Phi}_0$ is the complex amplitude of the waves. The combination $\overline{\Phi}_0 e^{-{\rm i}\omega t}$ is called the "Fourier transform" of Ψ. The factor $\sqrt{2\pi}$ is just a normalization factor that might be chosen differently in another book. Wave packets correspond to combinations in which the complex amplitude $\overline{\Phi}_0(k)$ is only nonzero in a small range of wave numbers k. More general combinations of waves may of course always be split up into such wave packets.

To describe the motion of wave packets is not quite as straightforward as it may seem: the envelope of a wave packet extends over a finite region, and different points on it actually move at somewhat different speeds. So what do you take as the point that defines the motion if you want to be precise? There is a trick here: consider very long times. For large times, the propagation distance is so large that it dwarfs the ambiguity about what point to take as the position of the envelope.

Finding the wave function Ψ for large time is a messy exercise banned to derivation {D.44}. But the conclusions are fairly straightforward. Assume that the range of waves in the packet is restricted to some small interval $k_1 < k < k_2$. In particular, assume that the variation in group velocity is relatively small and monotonous. In that case, for large times the wave function will be negligibly small except in the region

$$v_{\rm g1} t < x < v_{\rm g2} t$$

(In case $v_{\rm g1} > v_{\rm g2}$, invert these inequalities.) Since the variation in group velocity is small for the packet, it therefore definitely does move with "the" group velocity.

It is not just possible to say where the wave function is nonzero at large times. It is also possible to write a complete approximate wave function for large times:

$$\Psi(x,t) \sim \frac{e^{\mp \mathrm{i}\pi/4}}{\sqrt{|v'_{\mathrm{g}0}|t}} \overline{\Phi}_0(k_0) e^{\mathrm{i}(k_0 x - \omega_0 t)} \qquad v_{\mathrm{g}0} = \frac{x}{t}$$

Here k_0 is the wave number at which the group speed is exactly equal to x/t, ω_0 is the corresponding frequency, $v'_{\mathrm{g}0}$ is the derivative of the group speed at that point, and $\mp$ stands for the sign of $-v'_{\mathrm{g}0}$.

While this precise expression may not be that important, it is interesting to note that Ψ decreases in magnitude proportional to $1/\sqrt{t}$. That can be understood from conservation of the probability to find the particle. The wave packet spreads out proportional to time because of the small but nonzero variation in group velocity. Therefore Ψ must be proportional to $1/\sqrt{t}$ if its square integral is to remain unchanged.

One other interesting feature may be deduced from the above expression for Ψ. If you examine the wave function on the scale of a few oscillations, it looks as if it was a single component wave of wave number k_0 and frequency ω_0. Only if you look on a bigger scale do you see that it really is a wave packet. To understand why, just look at the differential

$$\mathrm{d}(k_0 x - \omega_0 t) = k_0 \mathrm{d}x - \omega_0 \mathrm{d}t + x\mathrm{d}k_0 - t\mathrm{d}\omega_0$$

and observe that the final two terms cancel because $\mathrm{d}\omega_0/\mathrm{d}k_0$ is the group velocity, which equals x/t. Therefore changes in k_0 and ω_0 do not show up on a small scale.

For the particle in free space, the result for the large time wave function can be written out further to give

$$\Psi(x,t) \sim e^{-\mathrm{i}\pi/4} \sqrt{\frac{m}{t}} \Phi_0\left(\frac{mx}{t}\right) e^{\mathrm{i}mx^2/2\hbar t}$$

Since the group speed p/m in this case is monotonously increasing, the wave packets have negligible overlap, and this is in fact the large time solution for any combination of waves, not just narrow wave packets.

In a typical true quantum mechanics case, Φ_0 will extend over a range of wave numbers that is not small, and may include both positive and negative values of the momentum p. So, there is no longer a meaningful velocity for the wave function: the wave function spreads out in all directions at velocities ranging from negative to positive. For example, if the momentum space wave function Φ_0 consists of *two* narrow nonzero regions, one at a positive value of p and one at a negative value, then the wave function in normal space splits into two separate wave packets. One packet moves with constant speed towards the left, the other with constant speed towards the right. The same particle is now

going in two completely different directions at the same time. That would be unheard of in classical Newtonian mechanics.

Key Points

- Component waves have the generic form $e^{\mathrm{i}(kx-\omega t)}$.
- The constant k is the wave number.
- The constant ω is the angular frequency.
- The relation between ω and k is called the dispersion relation.
- The phase velocity is ω/k. It describes how fast the wave moves.
- The group velocity is $\mathrm{d}\omega/\mathrm{d}k$. It describes how fast wave packets move.
- Relatively simple expressions exist for the wave function of wave packets at large times.

7.10.5 Electron motion through crystals

One important application of group velocity is the motion of conduction electrons through crystalline solids. This subsection discusses it.

Conduction electrons in solids must move around the atoms that make up the solid. You cannot just forget about these atoms in discussing the motion of the conduction electrons. Even semi-classically speaking, the electrons in a solid move in a roller-coaster ride around the atoms. Any external force on the electrons is *on top* of the large forces that the crystal already exerts. So it is simply wrong to say that the external force gives mass times acceleration of the electrons. Only the total force would do that.

Typically, on a microscopic scale the solid is crystalline; in other words, the atoms are arranged in a periodic pattern. That means that the forces on the electrons have a periodic nature. As usual, any direct interactions between particles will be ignored as too complex to analyze. Therefore, it will be assumed that the potential energy seen by an electron is a given periodic function of position.

It will also again be assumed that the motion is one-dimensional. In that case the energy eigenfunctions are determined from a one-dimensional Hamiltonian eigenvalue problem of the form

$$-\frac{\hbar^2}{2m_{\mathrm{e}}}\frac{\partial^2\psi}{\partial x^2} + V(x)\psi = E\psi \tag{7.65}$$

Here $V(x)$ is a periodic potential energy, with some given atomic-scale period d.

Three-dimensional energy eigenfunctions may be found as products of one-dimensional ones; compare chapter 3.5.8. Unfortunately however, that only works here if the three-dimensional potential is some sum of one-dimensional ones, as in

$$V(x,y,z) = V_x(x) + V_y(y) + V_z(z)$$

That is really quite limiting. The general conclusions that will be reached in this subsection continue to apply for any periodic potential, not just a sum of one-dimensional ones.

The energy eigenfunction solutions to (7.65) take the form of "Bloch waves:"

$$\psi^{\rm p}_k(x) = \psi^{\rm p}_{{\rm p},k}(x) e^{{\rm i}kx} \tag{7.66}$$

where $\psi^{\rm p}_{{\rm p},k}$ is a periodic function of period d like the potential.

The reason that the energy eigenfunctions take the form of Bloch waves is not that difficult to understand. It is a consequence of the fact that commuting operators have common eigenfunctions, chapter 4.5.1. Consider the "translation operator" $\mathcal{T}_d$ that shifts wave functions over one atomic period d. Since the potential is exactly the same after a wave function is shifted over an atomic period, the Hamiltonian commutes with the translation operator. It makes no difference whether you apply the Hamiltonian before or after you shift a wave function over an atomic period. Therefore, the energy eigenfunctions can be taken to be also eigenfunctions of the translation operator. The translation eigenvalue must have magnitude one, since the magnitude of a wave function does not change when you merely shift it. Therefore the eigenvalue can always be written as $e^{{\rm i}kd}$ for *some* real value k. And that means that if you write the eigenfunction in the Bloch form (7.66), then the exponential will produce the eigenvalue during a shift. So the part $\psi^{\rm p}_{{\rm p},k}$ must be the same after the shift. Which means that it is periodic of period d. (Note that you can always write any wave function in Bloch form; the nontrivial part is that $\psi^{\rm p}_{{\rm p},k}$ is periodic for actual Bloch waves.)

If the crystal is infinite in size, the wave number k can take any value. (For a crystal in a finite-size periodic box as studied in chapter 6.22, the values of k are discrete. However, this subsection will assume an infinite crystal.)

To understand what the Bloch form means for the electron motion, first consider the case that the periodic factor $\psi^{\rm p}_{{\rm p},k}$ is just a trivial constant. In that case the Bloch waves are eigenfunctions of linear momentum. The linear momentum p is then $\hbar k$. That case applies if the crystal potential is just a trivial constant. In particular, it is true if the electron is in free space.

Even if there is a nontrivial crystal potential, the so-called "crystal momentum" is still defined as:

$$\boxed{p_{\rm cm} = \hbar k} \tag{7.67}$$

(In three dimensions, substitute the vectors $\vec{p}$ and $\vec{k}$). But crystal momentum is not normal momentum. In particular, for an electron in a crystal you can no

longer get the propagation velocity by dividing the crystal momentum by the mass.

Instead you can get the propagation velocity by differentiating the energy with respect to the crystal momentum, {D.45}:

$$\boxed{v = \frac{\mathrm{d}E^{\mathrm{p}}}{\mathrm{d}p_{\mathrm{cm}}} \qquad p_{\mathrm{cm}} = \hbar k} \tag{7.68}$$

(In three dimensions, replace the p-derivative by $1/\hbar$ times the gradient with respect to $\vec{k}$.) In free space, $E^{\mathrm{p}} = \hbar\omega$ and $p_{\mathrm{cm}} = \hbar k$, so the above expression for the electron velocity is just the expression for the group velocity.

One conclusion that can be drawn is that electrons in an ideal crystal keep moving with the same speed for all times like they do in free space. They do not get scattered at all. The reason is that energy eigenfunctions are stationary. Each eigenfunction corresponds to a single value of k and so to a corresponding single value of the propagation speed v above. An electron wave packet will involve a small range of energy eigenfunctions, and a corresponding small range of velocities. But since the range of energy eigenfunctions does not change with time, neither does the range of velocities. Scattering, which implies a change in velocity, does not occur.

This perfectly organized motion of electrons through crystals is quite surprising. If you make up a classical picture of an electron moving through a crystal, you would expect that the electron would pretty much bounce off every atom it encountered. It would then perform a drunkard's walk from atom to atom. That would really slow down electrical conduction. But it does not happen. And indeed, experimentally electrons in metals may move past many thousands of atoms without getting scattered. In very pure copper at very low cryogenic temperatures electrons may even move past many millions of atoms before getting scattered.

Note that a total lack of scattering only applies to truly ideal crystals. Electrons can still get scattered by impurities or other crystal defects. More importantly, at normal temperatures the atoms in the crystal are not exactly in their right positions due to thermal motion. That too can scatter electrons. In quantum terms, the electrons then collide with the phonons of the crystal vibrations. The details are too complex to be treated here, but it explains why metals conduct much better still at cryogenic temperatures than at room temperature.

The next question is how does the propagation velocity of the electron change if an external force F_{ext} is applied? It turns out that Newton's second law, in terms of momentum, still works if you substitute the crystal momentum $\hbar k$ for the normal momentum, {D.45}:

$$\boxed{\frac{\mathrm{d}p_{\mathrm{cm}}}{\mathrm{d}t} = F_{\mathrm{ext}} \qquad p_{\mathrm{cm}} = \hbar k} \tag{7.69}$$

However, since the velocity is not just the crystal momentum divided by the mass, you cannot convert the left hand side to the usual mass times acceleration. The acceleration is instead, using the chain rule of differentiation,

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}^2 E^{\mathrm{p}}}{\mathrm{d}p_{\mathrm{cm}}^2}\frac{\mathrm{d}p_{\mathrm{cm}}}{\mathrm{d}t} = \frac{\mathrm{d}^2 E^{\mathrm{p}}}{\mathrm{d}p_{\mathrm{cm}}^2} F_{\mathrm{ext}}$$

For mass times acceleration to be the force, the factor multiplying the force in the final expression would have to be the reciprocal of the electron mass. It clearly is not; in general it is not even a constant.

But physicists still like to think of the effect of force as mass times acceleration of the electrons. So they cheat. They ignore the true mass of the electron. Instead they simply define a new "effective mass" for the electron so that the external force equals that effective mass times the acceleration:

$$\boxed{m_{\mathrm{eff}} \equiv 1 \Big/ \frac{\mathrm{d}^2 E^{\mathrm{p}}}{\mathrm{d}p_{\mathrm{cm}}^2} \qquad p_{\mathrm{cm}} = \hbar k} \tag{7.70}$$

Unfortunately, the effective mass is often a completely different number than the true mass of the electron. Indeed, it is quite possible for this "mass" to become negative for some range of wave numbers. Physically that means that if you put a force on the electron that pushes it one way, it will accelerate in the opposite direction! That can really happen. It is a consequence of the wave nature of quantum mechanics. Waves in crystals can be reflected just like electromagnetic waves can, and a force on the electron may move it towards stronger reflection.

For electrons near the bottom of the conduction band, the effective mass idea may be a bit more intuitive. At the bottom of the conduction band, the energy has a minimum. From calculus, if the energy E^{p} has a minimum at some wave number vector, then in a suitably oriented axis system it can be written as the Taylor series

$$E^{\mathrm{p}} = E^{\mathrm{p}}_{\mathrm{min}} + \tfrac{1}{2}\frac{\partial^2 E^{\mathrm{p}}}{\partial k_x^2}k_x^2 + \tfrac{1}{2}\frac{\partial^2 E^{\mathrm{p}}}{\partial k_y^2}k_y^2 + \tfrac{1}{2}\frac{\partial^2 E^{\mathrm{p}}}{\partial k_z^2}k_z^2 + \ldots$$

Here the wave number values are measured from the position of the minimum. This can be rewritten in terms of the crystal momenta and effective masses in each direction as

$$E^{\mathrm{p}} = E^{\mathrm{p}}_{\mathrm{min}} + \tfrac{1}{2}\frac{1}{m_{\mathrm{eff},x}}p_{\mathrm{cm},x}^2 + \tfrac{1}{2}\frac{1}{m_{\mathrm{eff},y}}p_{\mathrm{cm},y}^2 + \tfrac{1}{2}\frac{1}{m_{\mathrm{eff},z}}p_{\mathrm{cm},z}^2 + \ldots \tag{7.71}$$

In this case the effective masses are indeed positive, since second derivatives must be positive near a minimum. These electrons act much like classical particles. They move in the right direction if you put a force on them. Unfortunately,

the effective masses are not necessarily similar to the true electron mass, or even the same in each direction.

For the effective mass of the holes at the top of a valence band things get much messier still. For typical semiconductors, the energy no longer behaves as an analytic function, even though the energy in a specific direction continues to vary quadratically with the magnitude of the wave number. So the Taylor series is no longer valid. You then end up with such animals as "heavy holes," "light holes," and "split-off holes." Such effects will be ignored in this book.

Key Points

- The energy eigenfunctions for periodic potentials take the form of Bloch waves, involving a wave number k.
- The crystal momentum is defined as $\hbar k$.
- The first derivative of the electron energy with respect to the crystal momentum gives the propagation velocity.
- The second derivative of the electron energy with respect to the crystal momentum gives the reciprocal of the effective mass of the electron.

7.11 Almost Classical Motion

This section examines the motion of a particle in the presence of a single external force. Just like in the previous section, it will be assumed that the initial position and momentum are narrowed down sufficiently that the particle is restricted to a relatively small, coherent, region. Solutions of this type are called "wave packets."

In addition, for the examples in this section the forces vary slowly enough that they are approximately constant over the spatial extent of the wave packet. Hence, according to Ehrenfest's theorem, section 7.2.1, the wave packet should move according to the classical Newtonian equations.

The examples in this section were obtained on a computer, and should be numerically exact. Details about how they were computed can be found in addendum {A.27}, if you want to understand them better, or create some yourself.

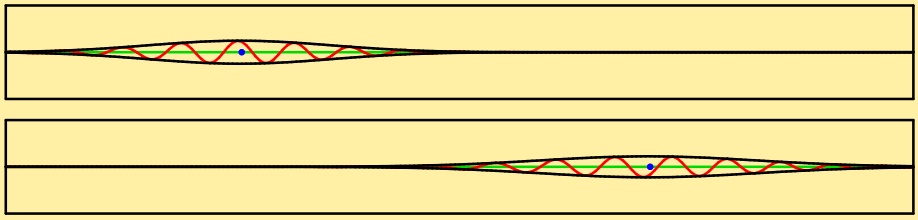
There is an easy general way to find approximate energy eigenfunctions and eigenvalues applicable under the conditions used in this section. It is called the WKB method. Addendum {A.28} has a description.

7.11.1 Motion through free space

First consider the trivial case that there are no forces; a particle in free space. This will provide the basis against which the motion with forces in the next

subsections can be compared to.

Classically, a particle in free space moves at a constant velocity. In quantum mechanics, the wave packet does too; figure 7.15 shows it at two different times.



Animation: http://www.dommelen.net/quansup/free.gif

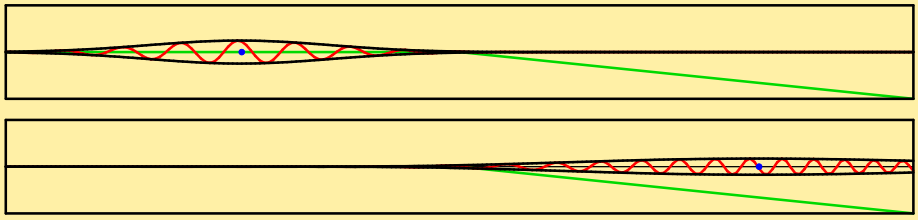
Hi-res: http://www.dommelen.net/quansup/freehi.html

Figure 7.15: A particle in free space.

If you step back far enough that the wave packet in the figures begins to resemble just a dot, you have classical motion. The blue point indicates the position of maximum wave function magnitude, as a visual anchor. It provides a reasonable approximation to the expectation value of position whenever the wave packet contour is more or less symmetric. A closer examination shows that the wave packet is actually changing a bit in size in addition to translating.

7.11.2 Accelerated motion

Figure 7.16 shows the motion when the potential energy (shown in green) ramps down starting from the middle of the plotted range. Physically this corresponds to a constant accelerating force beyond that point. A classical point particle would move at constant speed until it encounters the ramp, after which it would start accelerating at a constant rate. The quantum mechanical solution shows a corresponding acceleration of the wave packet, but in addition the wave packet stretches a lot.



Animation: http://www.dommelen.net/quansup/acc.gif

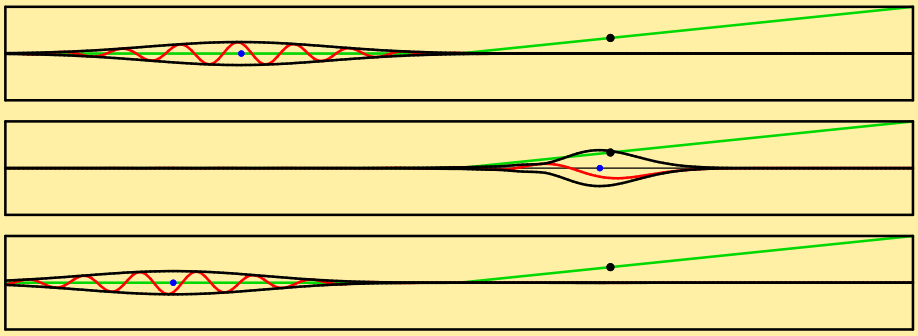
Hi-res: http://www.dommelen.net/quansup/acchi.html

Figure 7.16: An accelerating particle.

7.11.3 Decelerated motion

Figure 7.17 shows the motion when the potential energy (shown in green) ramps up starting from the center of the plotting range. Physically this corresponds to a constant decelerating force beyond that point. A classical point particle would move at constant speed until it encounters the ramp, after which it would start decelerating until it runs out of kinetic energy; then it would be turned back, returning to where it came from.

The quantum mechanical solution shows a corresponding reflection of the wave packet back to where it came from. The black dot on the potential energy line shows the "turning point" where the potential energy becomes equal to the nominal energy of the wave packet. That is the point where classically the particle runs out of kinetic energy and is turned back.



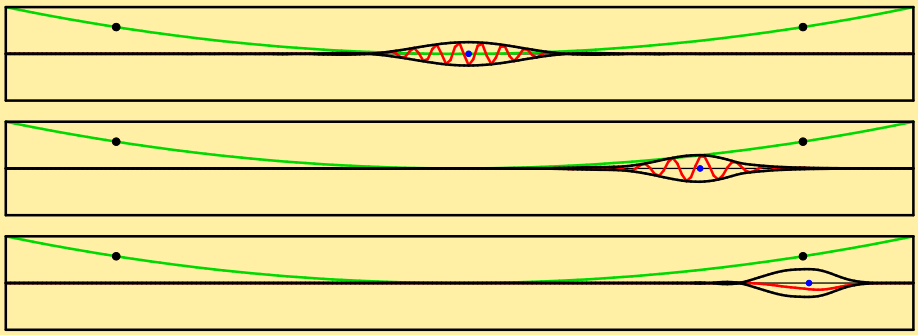
Animation: `http://www.dommelen.net/quansup/bounce.gif`
Hi-res: `http://www.dommelen.net/quansup/bouncehi.html`

Figure 7.17: A decelerating particle.

7.11.4 The harmonic oscillator

The harmonic oscillator describes a particle caught in a force field that prevents it from escaping in either direction. In all three previous examples the particle could at least escape towards the far left. The harmonic oscillator was the first real quantum system that was solved, in chapter 4.1, but only now, near the end of part I, can the classical picture of a particle oscillating back and forward actually be created.

There are some mathematical differences from the previous cases, because the energy levels of the harmonic oscillator are discrete, unlike those of the particles that are able to escape. But if the energy levels are far enough above the ground state, localized wave packets similar to the ones in free space may be formed, {A.27}. The animation in figure 7.18 gives the motion of a wave packet whose nominal energy is hundred times the ground state energy.



Animation: http://www.dommelen.net/quansup/harmmv.gif
Hi-res: http://www.dommelen.net/quansup/harmhi.html

Figure 7.18: Unsteady solution for the harmonic oscillator. The third picture shows the maximum distance from the nominal position that the wave packet reaches.

The wave packet performs a periodic oscillation back and forth just like a classical point particle would. In addition, it oscillates at the correct classical frequency ω. Finally, the point of maximum wave function, shown in blue, fairly closely obeys the classical limits of motion, shown as black dots.

Curiously, the wave function does *not* return to the same values after one period: it has changed sign after one period and it takes two periods for the wave function to return to the same values. It is because the sign of the wave function cannot be observed physically that classically the particle oscillates at frequency ω, and not at $\frac{1}{2}\omega$ like the wave function does.

Key Points

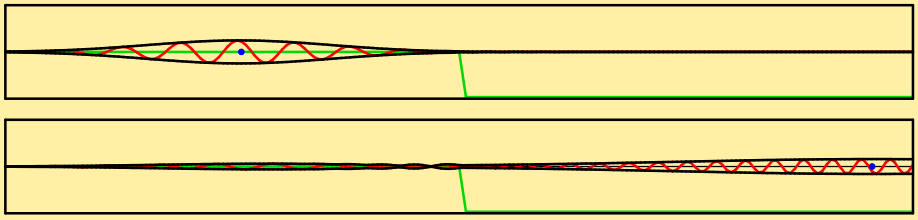
- When the forces change slowly enough on quantum scales, wave packets move just like classical particles do.
- Examined in detail, wave packets may also change shape over time.

7.12 Scattering

The motion of the wave packets in section 7.11 approximated that of classical Newtonian particles. However, if the potential starts varying nontrivially over distances short enough to be comparable to a quantum wave length, much more interesting behavior results, for which there is no classical equivalent. This section gives a couple of important examples.

7.12.1 Partial reflection

A classical particle entering a region of changing potential will keep going as long as its total energy exceeds the potential energy. Consider the potential shown in green in figure 7.19; it drops off to a lower level and then stays there. A classical particle would accelerate to a higher speed in the region of drop off and maintain that higher speed from there on.



Animation: http://www.dommelen.net/quansup/drop.gif

Hi-res: http://www.dommelen.net/quansup/drophi.html

Figure 7.19: A partial reflection.

However, the potential in this example varies so rapidly on quantum scales that the classical Newtonian picture is completely wrong. What actually happens is that the wave packet splits into two, as shown in the bottom figure. One part returns to where the packet came from, the other keeps on going.

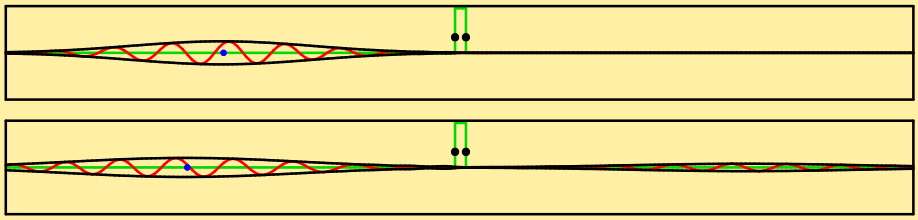
One hypothetical example used in chapter 3.1 was that of sending a single particle both to Venus and to Mars. As this example shows, a scattering setup gives a very real way of sending a single particle in two different directions at the same time.

Partial reflections are the norm for potentials that vary nontrivially on quantum scales, but this example adds a second twist. Classically, a *decelerating* force is needed to turn a particle back, but here the force is everywhere accelerating only! As an actual physical example of this weird behavior, neutrons trying to enter nuclei experience attractive forces that come on so quickly that they may be repelled by them.

7.12.2 Tunneling

A classical particle will never be able to progress past a point at which the potential energy exceeds its total energy. It will be turned back. However, the quantum mechanical truth is, if the region in which the potential energy exceeds the particle's energy is narrow enough on a quantum scale, the particle can go right through it. This effect is called "tunneling."

As an example, figure 7.20 shows part of the wave packet of a particle passing right through a region where the peak potential exceeds the particle's expectation energy by a factor three.

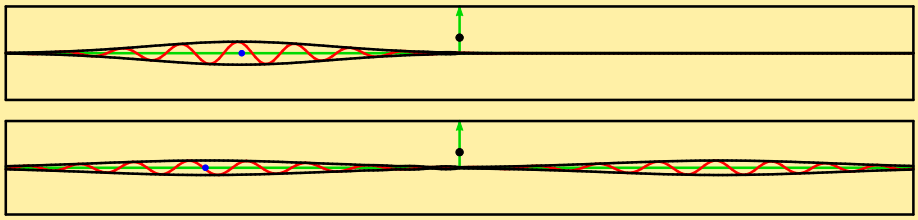


Animation: http://www.dommelen.net/quansup/tunnel.gif
Hi-res: http://www.dommelen.net/quansup/tunnelhi.html

Figure 7.20: An tunneling particle.

Of course, the energy values have some uncertainty, but it is small. The reason the particle can pass through is not because it has a chance of having three times its nominal energy. It absolutely does not; the simulation set the probability of having more than twice the nominal energy to zero exactly. The particle has a chance of passing through because its motion is governed by the Schrödinger equation, instead of the equations of classical physics.

And if that is not convincing enough, consider the case of a delta function barrier in figure 7.21; the limit of an infinitely high, infinitely narrow barrier. Being infinitely high, classically *nothing* can get past it. But since it is also infinitely narrow, a quantum particle will hardly notice a weak-enough delta function barrier. In figure 7.21, the strength of the delta function was chosen just big enough to split the wave function into equal reflected and transmitted parts. If you look for the particle afterwards, you have a 50/50 chance of finding it at either side of this "impenetrable" barrier.



Animation: http://www.dommelen.net/quansup/del.gif
Hi-res: http://www.dommelen.net/quansup/delhi.html

Figure 7.21: Penetration of an infinitely high potential energy barrier.

Curiously enough, a delta function well, (with the potential going down instead of up), reflects the same amount as the barrier version.

Tunneling has consequences for the mathematics of bound energy states. Classically, you can confine a particle by sticking it in between, say two delta function potentials, or between two other potentials that have a maximum potential energy V that exceeds the particle's energy E. But such a particle trap does not work in quantum mechanics, because given time, the particle would

tunnel through a local potential barrier. In quantum mechanics, a particle is bound only if its energy is less than the potential energy at infinite distance. Local potential barriers only work if they have infinite potential energy, and that over a larger range than a delta function.

Note however that in many cases, the probability of a particle tunneling out is so infinitesimally small that it can be ignored. For example, since the electron in a hydrogen atom has a binding energy of 13.6 eV, a 110 or 220 V ordinary household voltage should in principle be enough for the electron to tunnel out of a hydrogen atom. But don't wait for it; it is likely to take much more than the total life time of the universe. You would have to achieve such a voltage drop within an atom-scale distance to get some action.

One major practical application of tunneling is the scanning tunneling microscope. Tunneling can also explain alpha decay of nuclei, and it is a critical part of much advanced electronics, including current leakage problems in VLSI devices.

Key Points

- If the potential varies nontrivially on quantum scales, wave packets do not move like classical particles.
- A wave packet may split into separate parts that move in different ways.
- A wave packet may be reflected by an accelerating force.
- A wave packet may tunnel through regions that a classical particle could not enter.

7.13 Reflection and Transmission Coefficients

Scattering and tunneling can be described in terms of so-called "reflection and transmission coefficients." This section explains the underlying ideas.

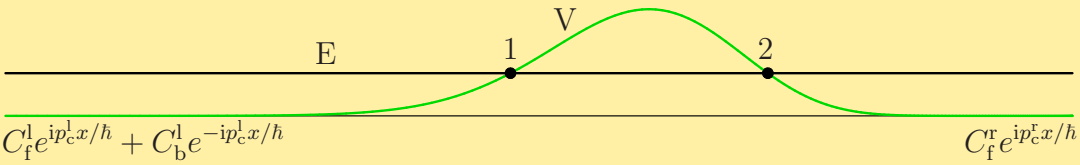


Figure 7.22: Schematic of a scattering potential and the asymptotic behavior of an example energy eigenfunction for a wave packet coming in from the far left.

Consider an arbitrary scattering potential like the one in figure 7.22. To the far left and right, it is assumed that the potential assumes a constant value. In

such regions the energy eigenfunctions take the form

$$\psi_E = C_{\rm f} e^{{\rm i}p_{\rm c}x/\hbar} + C_{\rm b} e^{-{\rm i}p_{\rm c}x/\hbar}$$

where $p_{\rm c} = \sqrt{2m(E-V)}$ is the classical momentum and $C_{\rm f}$ and $C_{\rm b}$ are constants. When eigenfunctions of slightly different energies are combined together, the terms $C_{\rm f} e^{{\rm i}p_{\rm c}x/\hbar}$ produce wave packets that move forwards in x, graphically from left to right, and the terms $C_{\rm b} e^{-{\rm i}p_{\rm c}x/\hbar}$ produce packets that move backwards. So the subscripts indicate the direction of motion.

This section is concerned with a single wave packet that comes in from the far left and is scattered by the nontrivial potential in the center region. To describe this, the coefficient $C_{\rm b}$ must be zero in the far-right region. If it was nonzero, it would produce a second wave packet coming in from the far right.

In the far-left region, the coefficient $C_{\rm b}$ is normally not zero. In fact, the term $C_{\rm b}^{\rm l} e^{-{\rm i}p_{\rm c}x/\hbar}$ produces the part of the incoming wave packet that is reflected back towards the far left. The relative amount of the incoming wave packet that is reflected back is called the "reflection coefficient" R. It gives the probability that the particle can be found to the left of the scattering region after the interaction with the scattering potential. It can be computed from the coefficients of the energy eigenfunction in the left region as, {A.32},

$$\boxed{R = \frac{|C_{\rm b}^{\rm l}|^2}{|C_{\rm f}^{\rm l}|^2}} \tag{7.72}$$

Similarly, the relative fraction of the wave packet that passes through the scattering region is called the "transmission coefficient" T. It gives the probability that the particle can be found at the other side of the scattering region afterwards. It is most simply computed as $T = 1 - R$: whatever is not reflected must pass through. Alternatively, it can be computed as

$$\boxed{T = \frac{p_{\rm c}^{\rm r}|C_{\rm f}^{\rm r}|^2}{p_{\rm c}^{\rm l}|C_{\rm f}^{\rm l}|^2} \qquad p_{\rm c}^{\rm l} = \sqrt{2m(E-V_{\rm l})} \quad p_{\rm c}^{\rm r} = \sqrt{2m(E-V_{\rm r})}} \tag{7.73}$$

where $p_{\rm c}^{\rm l}$ respectively $p_{\rm c}^{\rm r}$ are the values of the classical momentum in the far left and right regions.

Note that a coherent wave packet requires a small amount of uncertainty in energy. Using the eigenfunction at the nominal value of energy in the above expressions for the reflection and transmission coefficients will involve a small error. It can be made to go to zero by reducing the uncertainty in energy, but then the size of the wave packet will expand correspondingly.

In the case of tunneling through a high and wide barrier, the WKB approximation may be used to derive a simplified expression for the transmission coefficient, {A.29}. It is

$$\boxed{T \approx e^{-2\gamma_{12}} \qquad \gamma_{12} = \frac{1}{\hbar}\int_{x_1}^{x_2} |p_{\rm c}|\,{\rm d}x \quad |p_{\rm c}| = \sqrt{2m(V-E)}} \tag{7.74}$$

where x_1 and x_2 are the "turning points" in figure 7.22, in between which the potential energy exceeds the total energy of the particle.

Therefore in the WKB approximation, it is just a matter of doing a simple integral to estimate what is the probability for a wave packet to pass through a barrier. One famous application of that result is for the alpha decay of atomic nuclei. In such decay a so-called alpha particle tunnels out of the nucleus.

For similar considerations in three-dimensional scattering, see addendum {A.30}.

Key Points

- A transmission coefficient gives the probability for a particle to pass through an obstacle. A reflection coefficient gives the probability for it to be reflected.
- A very simple expression for these coefficients can be obtained in the WKB approximation.

Chapter 8

The Meaning of Quantum Mechanics

Engineers tend to be fairly matter-of-fact about the physics they use. Many use entropy on a daily basis as a computational tool without worrying much about its vague, abstract mathematical definition. Such a practical approach is even more important for quantum mechanics.

Famous quantum mechanics pioneer Niels Bohr had this to say about it:

> "For those who are not shocked when they first come across quantum theory cannot possibly have understood it." [Niels Bohr, quoted in W. Heisenberg (1971) *Physics and Beyond*. Harper and Row.]

Feynman was a Caltech quantum physicist who received a Nobel Prize for the creation of quantum electrodynamics with Schwinger and Tomonaga. He also pioneered nanotechnology with his famous talk "There's Plenty of Room at the Bottom." About quantum mechanics, he wrote:

> "There was a time when the newspapers said that only twelve men understood the theory of relativity. I do not believe there ever was such a time. There might have been a time when only one man did, because he was the only guy who caught on, before he wrote his paper. But after people read the paper, a lot of people understood the theory of relativity in some way or other, certainly more than twelve. On the other hand, I think I can safely say that nobody understands quantum mechanics." [Richard P. Feynman (1965) *The Character of Physical Law* 129. BBC/Penguin.]

Still, saying that quantum mechanics is ununderstandable raises the obvious question: "If we cannot understand it, does it at least seem plausible?" That is the question to be addressed in this chapter. When you read this chapter, you will see that the answer is simple and clear. Quantum mechanics is the most implausible theory ever formulated. Nobody would ever formulate a theory like

quantum mechanics in jest, because none would believe it. Physics ended up with quantum mechanics not because it seemed the most logical explanation, but because countless observations made it unavoidable.

8.1 Schrödinger's Cat

Schrödinger, apparently not an animal lover, came up with an example illustrating what the conceptual difficulties of quantum mechanics really mean in everyday terms. This section describes the example.

A cat is placed in a closed box. Also in the box is a Geiger counter and a tiny amount of radioactive material that will cause the Geiger counter to go off in a typical time of an hour. The Geiger counter has been rigged so that if it goes off, it releases a poison that kills the cat.

Now the decay of the radioactive material is a quantum-mechanical process; the different times for it to trigger the Geiger counter each have their own probability. According to the orthodox interpretation, "measurement" is needed to fix a single trigger time. If the box is left closed to prevent measurement, then at any given time, there is only a *probability* of the Geiger counter having been triggered. The cat is then alive, and also dead, each with a nonzero probability.

Of course no reasonable person is going to believe that she is looking at a box with a cat in it that is both dead and alive. The problem is obviously with what is to be called a "measurement" or "observation." The countless trillions of air molecules are hardly going to miss "observing" that they no longer enter the cat's nose. The biological machinery in the cat is not going to miss "observing" that the blood is no longer circulating. More directly, the Geiger counter is not going to miss "observing" that a decay has occurred; it is releasing the poison, isn't it?

If you postulate that the Geiger counter is in this case doing the "measurement" that the orthodox interpretation so deviously leaves undefined, it agrees with our common sense. But of course, this Deus ex Machina only *rephrases* our common sense; it provides no explanation *why* the Geiger counter would cause quantum mechanics to apparently terminate its normal evolution, no proof or plausible reason that the Geiger counter is *able* to fundamentally change the normal evolution of the wave function, and not even a shred of hard evidence *that* it terminates the evolution, if the box is truly closed.

There is a strange conclusion to this story. The entire point Schrödinger was trying to make was that no sane person is going to believe that a cat can be both dead and kicking around alive at the same time. But when the equations of quantum mechanics are examined more closely, it is found that they require exactly that. The wave function evolves into describing a series of different *realities*. In our own reality, the cat dies at a specific, apparently random time, just as common sense tells us. Regardless whether the box is open or not.

But, as discussed further in section 8.6, the mathematics of quantum mechanics extends beyond our reality. Other realities develop, which we humans are utterly unable to observe, and in each of those other realities, the cat dies at a different time.

8.2 Instantaneous Interactions

Special relativity has shown that we humans cannot transmit information at more than the speed of light. However, according to the orthodox interpretation, nature does not limit itself to the same silly restrictions that it puts on us. This section discusses why not.

Consider again the H_2^+-ion, with the single electron equally shared by the two protons. If you pull the protons apart, maintaining the symmetry, you get a wave function that looks like figure 8.1. You might send one proton off to your

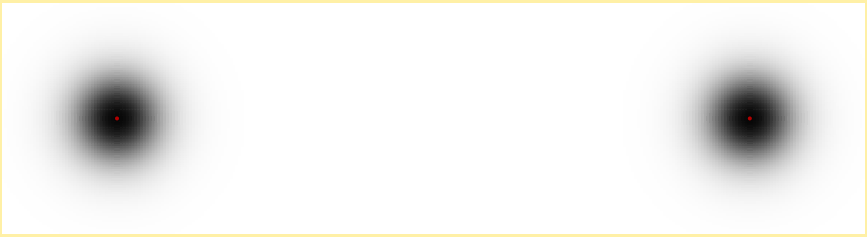


Figure 8.1: Separating the hydrogen ion.

observer on Mars, the other to your observer on Venus. Where is the *electron*, on Mars or on Venus?

According to the orthodox interpretation, the answer is: *neither*. A position for the electron *does not exist*. The electron is not on Mars. It is not on Venus. Only when either observer makes a measurement to see whether the electron is there, nature throws its dice, and based on the result, might put the electron on Venus and zero the wave function on Mars. But regardless of the distance, it could just as well have put the electron on Mars, if the dice would have come up differently.

You might think that nature cheats, that when you take the protons apart, nature already decides where the electron is going to be. That the Venus proton secretly hides the electron "in its sleeve", ready to make it appear if an observation is made. John Bell devised a clever test to force nature to reveal whether it has something hidden in its sleeve during a similar sort of trick.

The test case Bell used was a generalization of an experiment proposed by Bohm. It involves spin measurements on an electron/positron pair, created by the decay of an π-meson. Their combined spins are in the singlet state because the meson has no net spin. In particular, if you measure the spins of the electron and positron in any given direction, there is a 50/50% chance for each that it

turns out to be positive or negative. However, if one is positive, the other must be negative. So there are only two different possibilities:

1. electron positive and positron negative,
2. electron negative and positron positive.

Now suppose Earth happens to be almost the same distance from Mars and Venus, and you shoot the positron out to Venus, and the electron to Mars, as shown at the left in the figure below:



Figure 8.2: The Bohm experiment before the Venus measurement (left), and immediately after it (right).

You have observers on both planets waiting for the particles. According to quantum mechanics, the traveling electron and positron are both in an indeterminate state.

The positron reaches Venus a fraction of a second earlier, and the observer there measures its spin in the direction up from the ecliptic plane. According to the orthodox interpretation, nature now makes a random selection between the two possibilities, and assume it selects the positive spin value for the positron, corresponding to a spin that is up from the ecliptic plane, as shown in figure 8.2. Immediately, then, the spin state of the electron on Mars must also have collapsed; the observer on Mars is guaranteed to now measure negative spin, or spin down, for the electron.

The funny thing is, if you believe the orthodox interpretation, the information about the measurement of the positron has to reach the electron instantaneously, much faster than light can travel. This apparent problem in the orthodox interpretation was discovered by Einstein, Podolski, and Rosen. They doubted it could be true, and argued that it indicated that something must be missing in quantum mechanics.

In fact, instead of superluminal effects, it seems much more reasonable to assume that earlier on earth, when the particles were sent on their way, nature attached a secret little "note" of some kind to the positron, saying the equivalent of "If your spin up is measured, give the positive value", and that it attached a little note to the electron "If your spin up is measured, give the negative value." The results of the measurements are still the same, and the little notes travel along with the particles, well below the speed of light, so all seems now fine. Of course, these would not be true notes, but some kind of additional information beyond the normal quantum mechanics. Such postulated additional information sources are called "hidden variables."

Bell saw that there was a fundamental flaw in this idea if you do a large number of such measurements and you allow the observers to select from more

than one measurement direction at random. He derived a neat little general formula, but the discussion here will just show the contradiction in a single case. In particular, the observers on Venus and Mars will be allowed to select randomly one of three measurement directions $\vec{a}$, $\vec{b}$, and $\vec{c}$ separated by 120 degrees:

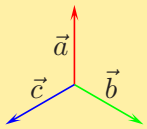


Figure 8.3: Spin measurement directions.

Let's see what the little notes attached to the electrons might say. They might say, for example, "Give the + value if $\vec{a}$ is measured, give the − value if $\vec{b}$ is measured, give the + value if $\vec{c}$ is measured." The relative fractions of the various possible notes generated for the electrons will be called $f_1, f_2, \ldots$. There are 8 different possible notes:

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\vec{a}$	+	+	+	+	−	−	−	−
$\vec{b}$	+	+	−	−	+	+	−	−
$\vec{c}$	+	−	+	−	+	−	+	−

The sum of the fractions f_1 through f_8 must be one. In fact, because of symmetry, each note will probably on average be generated for $\frac{1}{8}$ of the electrons sent, but this will not be needed.

Of course, each note attached to the positron must always be just the opposite of the one attached to the electron, since the positron must measure + in a direction when the electron measures − in that direction and vice-versa.

Now consider those measurements in which the Venus observer measures direction $\vec{a}$ and the Mars observer measures direction $\vec{b}$. In particular, the question is in what fraction of such measurements the Venus observer measures the opposite sign from the Mars observer; call it $f_{ab,\text{opposite}}$. This is not that hard to figure out. First consider the case that Venus measures − and Mars +. If the Venus observer measures the − value for the positron, then the note attached *to the electron* must say "measure + for $\vec{a}$"; further, if the Mars observer measures the + value for $\vec{b}$, that one should say "measure +" too. So, looking at the table, the relative fraction where Venus measures −and Mars measures + is where the electron's note has a + for both $\vec{a}$ and $\vec{b}$: $f_1 + f_2$.

Similarly, the fraction of cases where Venus finds + and Mars −is $f_7 + f_8$, and you get in total:

$$f_{ab,\text{opposite}} = f_1 + f_2 + f_7 + f_8 = 0.25$$

The value 0.25 is what quantum mechanics predicts; the derivation will be skipped here, but it has been verified in the experiments done after Bell's work. Those experiments also made sure that nature did not get the chance to do *sub*luminal communication. The same way you get

$$f_{ac,\text{opposite}} = f_1 + f_3 + f_6 + f_8 = 0.25$$

and

$$f_{bc,\text{opposite}} = f_1 + f_4 + f_5 + f_8 = 0.25$$

Now there is a problem, because the numbers add up to 0.75, but the fractions add up to at least 1: the sum of f_1 through f_8 is one.

A seemingly perfectly logical and plausible explanation by great minds is tripped up by some numbers that just do not want to match up. They only leave the alternative nobody really wanted to believe.

Attaching notes does not work. Information on what the observer on Venus decided to measure, the one thing that could not be put in the notes, must have been communicated *instantly* to the electron on Mars regardless of the distance.

It can also safely be concluded that we humans will never be able to see inside the actual machinery of quantum mechanics. For, suppose the observer on Mars could see the wave function of the electron collapse. Then the observer on Venus could send her Morse signals faster than the speed of light by either measuring or not measuring the spin of the positron. Special relativity would then allow signals to be sent into the past, and that leads to logical contradictions such as the Venus observer preventing her mother from having her.

While the results of the spin measurements can be observed, they do not allow superluminal communication. While the observer on Venus affects the results of the measurements of the observer on Mars, they will look completely random to that observer. Only when the observer on Venus sends over the results of her measurements, at a speed less than the speed of light, and the two sets of results are *compared*, do meaningful patterns how up.

The Bell experiments are often used to argue that Nature must really make the collapse decision using a true random number generator, but that is of course crap. The experiments indicate that Nature instantaneously transmits the collapse decision on Venus to Mars, but say nothing about how that decision was reached.

Superluminal effects still cause paradoxes, of course. The left of figure 8.4 shows how a Bohm experiment appears to an observer on earth. The spins



Figure 8.4: Earth's view of events (left), and that of a moving observer (right).

remain undecided until the measurement by the Venus observer causes both the positron and the electron spins to collapse.

However, for a moving observer, things would look very different. Assuming that the observer and the particles are all moving at speeds comparable to the speed of light, the same situation may look like the right of figure 8.4, chapter 1.1.4. In this case, the observer on *Mars* causes the wave function to collapse at a time that the positron has only just started moving towards Venus!

So the orthodox interpretation is not quite accurate. It should really have said that the measurement on Venus causes a *convergence* of the wave function, not an absolute collapse. What the observer of Venus really achieves in the orthodox interpretation is that after her measurement, *all* observers agree that the positron wave function is collapsed. Before that time, some observers are perfectly correct in saying that the wave function is already collapsed, and that the Mars observer did it.

It should be noted that when the equations of quantum mechanics are correctly applied, the collapse and superluminal effects disappear. That is explained in section 8.6. But, due to the fact that there are limits to our observational capabilities, as far as our own human experiences are concerned, the paradoxes remain real.

To be perfectly honest, it should be noted that the example above is not quite the one of Bell. Bell really used the inequality:

$$|2(f_3 + f_4 + f_5 + f_6) - 2(f_2 + f_4 + f_5 + f_7)| \leqslant 2(f_2 + f_3 + f_6 + f_7)$$

So the discussion cheated. And Bell allowed general directions of measurement not just 120 degree ones. See [25, pp. 423-426]. The above discussion seems a lot less messy, even though not historically accurate.

8.3 Global Symmetrization

When computing, say a hydrogen molecule, it is all nice and well to say that the wave function must be antisymmetric with respect to exchange of the two electrons 1 and 2, so the spin state of the molecule must be the singlet one. But what about, say, electron 3 in figure 8.1, which can with 50% chance be found on Mars and otherwise on Venus? Should not the wave function also be antisymmetric, for example, with respect to exchange of this electron 3 in one of two places in space with electron 1 on the hydrogen molecule on Earth? And would this not locate electron 3 in space also in part on the hydrogen molecule, and electron 1 also partly in space?

The answer is: absolutely. Nature treats *all* electrons as one big connected bunch. The given solution for the hydrogen molecule is not correct; it should have included *every* electron in the universe, not just two of them. Every

electron in the universe is just as much present on this single hydrogen molecule as the assumed two.

From the difficulty in describing the 33 electrons of the arsenic atom, imagine having to describe all electrons in the universe at the same time! If the universe is truly flat, this number would not even be finite. Fortunately, it turns out that the observed quantities can be correctly predicted pretending there are only two electrons involved. Antisymmetrization with far-away electrons does not change the properties of the local solution.

If you are thinking that more advanced quantum theories will eventually do away with the preposterous notion that all electrons are present everywhere, do not be too confident. As mentioned in addendum {A.15.1}, the idea has become a fundamental tenet in quantum field theory.

8.4 A story by Wheeler

Consider a simple question. Why are all electrons so absolutely equal? Would it not be a lot less boring if they had a range of masses and charges? As in "I found a really big electron this morning, with an unbelievable charge!" It does not happen.

And it is in fact far, far, worse than that. In quantum mechanics electrons are absolutely identical. If you *really* write the correct (classical) wave function for an hydrogen atom following the rules of quantum mechanics, then in principle you must include every electron in the universe as being present, in part, on the atom. Electrons are so equal that one cannot be present on a hydrogen atom unless every electron in the universe is.

There is a simple explanation that the famous physicist Wheeler gave to his talented graduate student Richard Feynman. In Feynman's words:

> "As a by-product of this same view, I received a telephone call one day at the graduate college at Princeton from Professor Wheeler, in which he said, 'Feynman, I know why all electrons have the same charge and the same mass' 'Why?' 'Because, they are all the same electron!' And, then he explained on the telephone, ..." [Richard P. Feynman (1965) Nobel prize lecture. ⟦5⟧]

What Professor Wheeler explained on the phone is sketched in the space-time diagram figure 8.5. The "world-line" of the only electron there is is constantly traveling back and forwards between the past and the future. At any given time, like today, this single electron can be observed at countless different locations. At the locations where the electron is traveling to the future it behaves like a normal electron. And Wheeler recognized that where the electron is traveling towards the past, it behaves like a positively charged electron, called a positron. The mystery of all those countless identical electrons was explained.

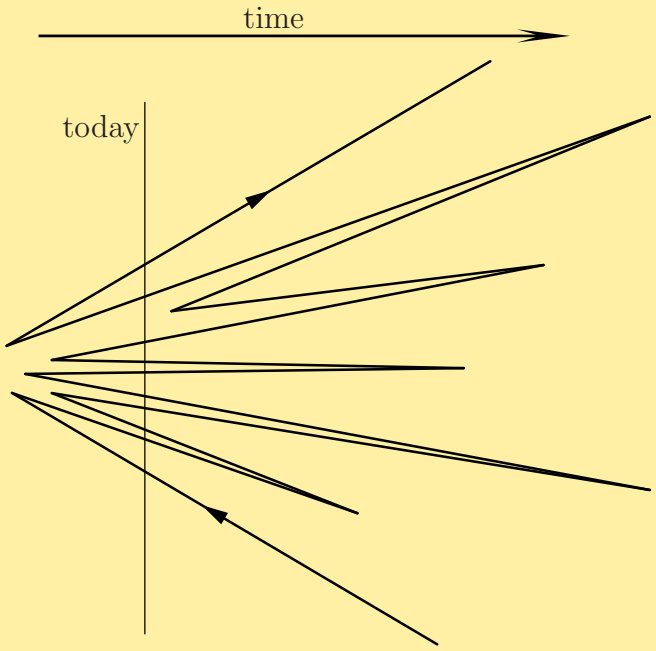


Figure 8.5: The space-time diagram of Wheeler's single electron.

What had Feynman to say about that!? Again in his words:

> " 'But, Professor,' I said, 'there aren't as many positrons as electrons.' 'Well, maybe they are hidden in the protons or something,' he said. I did not take the idea that all the electrons were the same one from him as seriously as I took the observation that positrons could simply be represented as electrons going from the future to the past in a back section of their world lines. That, I stole!" [Richard P. Feynman (1965) Nobel prize lecture. [[5]]]

And there are other problems, like that electrons can be created or destroyed in weak interactions.

But without doubt, if this was art instead of quantum mechanics, Wheeler's proposal would be considered one of the greatest works of all time. It is stunning in both its utter simplicity and its inconceivable scope.

There is a place for esthetics in quantum mechanics, as the Dirac equation illustrates. Therefore this section will take a very biased look at whether the idea is really so truly inconceivable as it might appear. To do so, only positrons and electrons will be considered, with their attendant photons. Shape-shifting electrons are a major additional complication. And recall classical mechanics. Some of the most esthetical results of classical mechanics are the laws of conservation of energy and momentum. Relativity and then quantum mechanics eventually found that classical mechanics is fundamentally completely wrong. But did conservation of energy and momentum disappear? Quite the contrary.

They took on an even deeper and more esthetically gratifying role in those theories.

With other particles shoved out of the way, the obvious question is the one of Feynman. Where are all the positrons? One idea is that they ended up in some other part of space. But that seems to be hard to reconcile with the fact that space seems quite similar in all directions. The positrons will still have to be around us. So why do we not see them? Recall that the model considered here has no protons for positrons to hide in.

Obviously, if the positrons have nowhere to hide, they must be in plain view. That seems theoretically possible if it is assumed that the positron quantum wave functions are delocalized on a gigantic scale. Note that astronomy is short of a large amount of mass in the universe one way or the other. Delocalized antimatter to the tune of the visible matter would be just a drop in the bucket.

A bit of mathematical trickery called the Cauchy-Schwartz inequality can be used to illustrate the idea. Consider a "universe" of volume $\mathcal{V}$. For simplicity, assume that there is just one electron and one positron in this universe. More does not seem to make a fundamental difference, at least not in a simplistic model. The electron has wave function ψ_1 and the positron ψ_2. The Cauchy-Schwartz inequality says that:

$$\left| \int_{\mathcal{V}} \psi_1^* \psi_2 \, \mathrm{d}^2\vec{r} \right|^2 \leqslant \int_{\mathcal{V}} |\psi_1|^2 \, \mathrm{d}^2\vec{r} \int_{\mathcal{V}} |\psi_2|^2 \, \mathrm{d}^2\vec{r} = 1$$

Take the left hand side as representative for the interaction rate between electrons and positrons. Then if the wave functions of both electrons and positrons are completely delocalized, the interaction rate is 1. However, if only the positrons are completely delocalized, it is much smaller. Suppose the electron is localized within a volume $\varepsilon\mathcal{V}$ with ε a very small number. Then the interaction rate is reduced from 1 to ε. If both electrons and positrons are localized within volumes of size $\varepsilon\mathcal{V}$ it gets messier. If the electron and positron move completely randomly and quickly through the volume, the average interaction rate would still be ε. But electrons and positrons attract each other through their electric charges, and on a large scale also through gravity. That could increase the interaction rate greatly.

The obvious next question is then, how come that positrons are delocalized and electrons are not? The simple answer to that is: because electrons come to us from the compact Big Bang stages of the universe. The positrons come to us from the final stages of the evolution of the universe where it has expanded beyond limit.

Unfortunately, that answer, while simple, is not satisfactory. Motion in quantum mechanics is essentially time reversible. And that means that you should be able to explain the evolution of both electrons and positrons coming out of the initial Big Bang universe. Going forward in time.

A more reasonable idea is that the other options do not produce stable situations. Consider a localized positron in an early universe that by random chance happens to have more localized electrons than positrons. Because of attraction effects, such a positron is likely to find a localized electron to annihilate with. That is one less localized positron out of an already reduced population. A delocalized positron could interact similarly with a delocalized electron, but there are less of these. The reverse situation holds for electrons. So you could imagine a runaway process where the positron population evolves to delocalized states and the electrons to localized ones.

Another way to look at it is to consider how wave functions get localized in the first place. The wave function of a localized isolated particle wants to disperse out over time. Cosmic expansion would only add to that. In the orthodox view, particles get localized because they are "measured." The basics of this process, as described by another graduate student of Wheeler, Everett III, are in section 8.6. Unfortunately, the process remains poorly understood. But suppose, say, that matter localizes matter but delocalizes antimatter, and vice-versa. In that case a slight dominance of matter over antimatter could conceivably lead to a run-away situation where the matter gets localized and the antimatter delocalized.

Among all the exotic sources that have been proposed for the "dark matter" in the universe, delocalized antimatter does not seem to get mentioned. So probably someone has already solidly shown that it is impossible.

But that does not invalidate Wheeler's basic idea, of course. As Wheeler himself suggested, the positrons could in fact be hiding inside the protons through the weak-force mechanism. Then of course, you need to explain how the positrons came to be hiding inside the protons. Why not the electrons inside the antiprotons? That would be messier, but it does not mean it could not be true. In fact, it is one of the surprises of advanced particle physics that the entire lepton-quark family seems to be one inseparable multi-component particle, [27, p. 210]. It seems only fair to say that Wheeler's idea *predicted* this. For clearly, the electron could not maintain its unmutable identity if repeatedly changed into particles with a separate and independent identity. So Wheeler's idea may not be so crazy after all, looking at the facts. It provides a real explanation why identical particles are so perfectly identical. And it predicted something that would only be observed well into the future.

Still, the bottom line remains the beauty of the idea. As the mathematician Weyl noted, unfazed after Einstein shot down an idea of his:

> "When there is a conflict between beauty and truth, I choose beauty."

8.5 Failure of the Schrödinger Equation?

Section {8.2} mentioned sending half of the wave function of an electron to Venus, and half to Mars. A scattering setup as described in chapter 7.12 provides a practical means for actually doing this, (at least, for taking the wave function apart in two separate parts.) The obvious question is now: can the Schrödinger equation also describe the physically observed "collapse of the wave function", where the electron changes from being on both Venus and Mars with a 50/50 probability to, say, being on Mars with absolute certainty?

The answer obtained in this and the next subsection will be most curious: no, the Schrödinger equation flatly *contradicts* that the wave function collapses, but yes, it *requires* that measurement leads to the experimentally observed collapse. The analysis will take us to a mind-boggling but really unavoidable conclusion about the very nature of our universe.

This subsection will examine the problem the Schrödinger equation has with describing a collapse. First of all, the solutions of the linear Schrödinger equation do not allow a mathematically exact collapse like some nonlinear equations do. But that does not necessarily imply that solutions would not be able to collapse physically. It would be conceivable that the solution could evolve to a state where the electron is on Mars with such high probability that it can be taken to be certainty. In fact, a common notion is that, somehow, interaction with a macroscopic "measurement" apparatus could lead to such an end result.

Of course, the constituent particles that make up such a macroscopic measurement apparatus still need to satisfy the laws of physics. So let's make up a reasonable model for such a complete macroscopic system, and see what can then be said about the possibility for the wave function to evolve towards the electron being on Mars.

The model will ignore the existence of anything beyond the Venus, Earth, Mars system. It will be assumed that the three planets consist of a humongous, but finite, number of conserved classical particles $1, 2, 3, 4, 5, \ldots$, with a supercolossal wave function:

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \vec{r}_3, S_{z3}, \vec{r}_4, S_{z4}, \vec{r}_5, S_{z5}, \ldots)$$

Particle 1 will be taken to be the scattered electron. It will be assumed that the wave function satisfies the Schrödinger equation:

$$\mathrm{i}\hbar \frac{\partial \Psi}{\partial t} = -\sum_i \sum_{j=1}^{3} \frac{\hbar^2}{2m_i} \frac{\partial^2 \Psi}{\partial r_{i,j}^2} + V(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \vec{r}_3, S_{z3}, \vec{r}_4, S_{z4}, \ldots)\Psi \qquad (8.1)$$

Trying to write the solution to this problem would of course be prohibitive, but the evolution of the probability of the electron to be on Venus can still be extracted from it with some fairly standard manipulations. First, taking the combination of the Schrödinger equation times Ψ^* minus the complex conjugate

of the Schrödinger equation times Ψ produces after some further manipulation an equation for the time derivative of the probability:

$$\mathrm{i}\hbar\frac{\partial\Psi^*\Psi}{\partial t} = -\sum_i \sum_{j=1}^{3} \frac{\hbar^2}{2m_i}\frac{\partial}{\partial r_{i,j}}\left(\Psi^*\frac{\partial\Psi}{\partial r_{i,j}} - \Psi\frac{\partial\Psi^*}{\partial r_{i,j}}\right) \tag{8.2}$$

The question is the probability for the electron to be on Venus, and you can get that by integrating the probability equation above over all possible positions and spins of the particles *except* for particle 1, for which you have to restrict the spatial integration to Venus and its immediate surroundings. If you do that, the left hand side becomes the rate of change of the probability for the electron to be on Venus, regardless of the position and spin of all the other particles.

Interestingly, assuming times at which the Venus part of the scattered electron wave is definitely at Venus, the right hand side integrates to zero: the wave function is supposed to disappear at large distances from this isolated system, and whenever particle 1 would be at the border of the surroundings of Venus.

It follows that the probability for the electron to be at Venus cannot change from 50%. A true collapse of the wave function of the electron as postulated in the orthodox interpretation, where the probability to find the electron at Venus changes to 100% or 0% cannot occur.

Of course, the model was simple; you might therefore conjecture that a true collapse could occur if additional physics is included, such as nonconserved particles like photons, or other relativistic effects. But that would obviously be a moving target. The analysis made a good-faith effort to examine whether including macroscopic effects may cause the observed collapse of the wave function, and the answer was no. Having a scientifically open mind requires you to at least follow the model to its logical end; nature might be telling you something here.

Is it really true that the results disagree with the observed physics? You need to be careful. There is no reasonable doubt that if a measurement is performed about the presence of the electron on Venus, the wave function will be observed to collapse. But all you established above is that the wave function does not collapse; you did not establish whether or not it will be *observed* to collapse. To answer the question whether a collapse will be *observed*, you will need to include the observers in your reasoning.

The problem is with the innocuous looking phrase *regardless of the position and spin of all the other particles* in the arguments above. Even while the total probability for the electron to be at Venus must stay at 50% in this example system, it is still perfectly possible for the probability to become 100% for one state of the particles that make up the observer and her tools, and to be 0% for another state of the observer and her tools.

It is perfectly possible to have a state of the observer with brain particles, ink-on-paper particles, tape recorder particles, that all say that the electron is

on Venus, combined with 100% probability that the electron is on Venus, and a second state of the observer with brain particles, ink-on-paper particles, tape recorder particles, that all say the electron must be on Mars, combined with 0% probability for the electron to be on Venus. Such a scenario is called a "relative state interpretation;" the states of the observer and the measured object become entangled with each other.

The state of the electron does not change to a single state of presence or absence; instead two states of the macroscopic universe develop, one with the electron absent, the other with it present. As explained in the next subsection, the Schrödinger equation does not just *allow* this to occur, it *requires* this to occur. So, far from being in conflict with the observed collapse, the model above requires it. The model produces the right physics: observed collapse is a consequence of the Schrödinger equation, not of something else.

But all this ends up with the rather disturbing thought that there are now two states of the universe, and the two are different in what they think about the electron. This conclusion was unexpected; it comes as the unavoidable consequence of the mathematical equations that quantum mechanics abstracted for the way nature operates.

8.6 The Many-Worlds Interpretation

The Schrödinger equation has been enormously successful, but it describes the wave function as always smoothly evolving in time, in apparent contradiction to its postulated collapse in the orthodox interpretation. So, it would seem to be extremely interesting to examine the solution of the Schrödinger equation for measurement processes more closely, to see whether and how a collapse might occur.

Of course, if a true solution for a single arsenic atom already presents an unsurmountable problem, it may seem insane to try to analyze an entire macroscopic system such as a measurement apparatus. But in a brilliant Ph.D. thesis with Wheeler at Princeton, Hugh Everett, III did exactly that. He showed that the wave function does *not* collapse. However it *seems* to us humans that it does, so we *are* correct in applying the rules of the orthodox interpretation anyway. This subsection explains briefly how this works.

Let's return to the experiment of section 8.2, where a positron is sent to Venus and an entangled electron to Mars, as in figure 8.6. The spin states are



Figure 8.6: Bohm's version of the Einstein, Podolski, Rosen Paradox.

uncertain when the two are sent from Earth, but when Venus measures the spin of the positron, it miraculously causes the spin state of the electron on Mars to collapse too. For example, if the Venus positron collapses to the spin-up state in the measurement, the Mars electron *must* collapse to the spin-down state. The problem, however, is that there is nothing in the Schrödinger equation to describe such a collapse, nor the superluminal communication between Venus and Mars it implies.

The reason that the collapse and superluminal communication are needed is that the two particles are entangled in the singlet spin state of chapter 5.5.6. This is a 50% / 50% probability state of (electron up and positron down) / (electron down and positron up).

It would be easy if the positron would just be spin up and the electron spin down, as in figure 8.7. You would still not want to write down the supercolossal



Figure 8.7: Nonentangled positron and electron spins; up and down.

wave function of *everything,* the particles along with the observers and their equipment for this case. But there is no doubt what it describes. It will simply describe that the observer on Venus measures spin up, and the one on Mars, spin down. There is no ambiguity.

The same way, there is no question about the opposite case, figure 8.8. It



Figure 8.8: Nonentangled positron and electron spins; down and up.

will produce a wave function of everything describing that the observer on Venus measures spin down, and the one on Mars, spin up.

Everett, III recognized that the solution for the entangled case is blindingly simple. Since the Schrödinger equation is *linear*, the wave function for the entangled case must simply be the sum of the two nonentangled ones above, as shown in figure 8.9. If the wave function in each nonentangled case describes a universe in which a particular state is solidly established for the spins, then the conclusion is undeniable: the wave function in the entangled case describes *two* universes, each of which solidly establishes states for the spins, but which end up with opposite results.

This explains the result of the orthodox interpretation that only eigenvalues are measurable. The linearity of the Schrödinger equation leaves no other option:

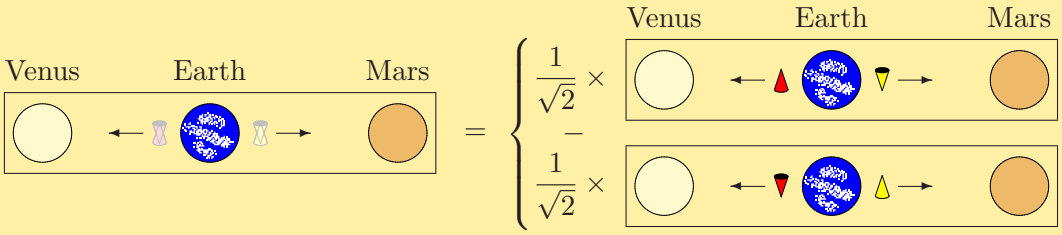


Figure 8.9: The wave functions of two universes combined

> *Assume that any measurement device at all is constructed that for a spin-up positron results in a universe that has absolutely no doubt that the spin is up, and for a spin-down positron results in a universe that has absolutely no doubt that the spin is down. In that case a combination of spin up and spin down states must unavoidably result in a combination of two universes, one in which there is absolutely no doubt that the spin is up, and one in which there is absolutely no doubt that it is down.*

Note that this observation does not depend on the details of the Schrödinger equation, just on its linearity. For that reason it stays true even including relativity.

The two universes are completely unaware of each other. It is the very nature of linearity that if two solutions are combined, they do not affect each other at all: neither universe would change in the least whether the other universe is there or not. For each universe, the other universe "exists" only in the sense that the Schrödinger equation must have created it given the initial entangled state.

Nonlinearity would be needed to allow the solutions of the two universes to couple together to produce a single universe with a combination of the two eigenvalues, and there is none. A universe measuring a combination of eigenvalues is made impossible by linearity.

While the wave function has not collapsed, what has changed is *the most meaningful way to describe it.* The wave function still by its very nature assigns a value to every possible configuration of the universe, in other words, to every possible universe. That has never been a matter of much controversy. And after the measurement it is still perfectly *correct* to say that the Venus observer has marked down in her notebook that the positron was up and down, and has transmitted a message to earth that the positron was up and down, and earth has marked on in its computer disks and in the brains of the assistants that the positron was found to be up and down, etcetera.

But it is *much more precise* to say that after the measurement there are two universes, one in which the Venus observer has observed the positron to be up, has transmitted to earth that the positron was up, and in which earth has

marked down on its computer disks and in the brains of the assistants that the positron was up, etcetera; and a second universe in which the same happened, but with the positron everywhere down instead of up. This description is much more precise since it notes that up always goes with up, and down with down. As noted before, this more precise way of describing what happens is called the "relative state formulation."

Note that in each universe, it *appears* that the wave function has collapsed. Both universes agree on the fact that the decay of the π-meson creates an electron/positron pair in a singlet state, but after the measurement, the notebook, radio waves, computer disks, brains in one universe all say that the positron is up, and in the other, all down. Only the unobservable full wave function "knows" that the positron is still both up and down.

And there is no longer a spooky superluminal action: in the first universe, the electron was already down when sent from earth. In the other universe, it was sent out as up. Similarly, for the case of the last subsection, where half the wave function of an electron was sent to Venus, the Schrödinger equation does not fail. There is still half a chance of the electron to be on Venus; it just gets decomposed into one universe with one electron, and a second one with zero electron. In the first universe, earth sent the electron to Venus, in the second to Mars. The contradictions of quantum mechanics disappear when the *complete* solution of the Schrödinger equation is examined.

Next, let's examine why the results would seem to be covered by rules of chance, even though the Schrödinger equation is fully deterministic. To do so, assume earth keeps on sending entangled positron and electron pairs. When the third pair is on its way, the situation looks as shown in the third column of figure 8.10. The wave function now describes 8 universes. Note that in *most* universes the observer starts seeing an apparently random sequence of up and down spins. When repeated enough times, the sequences appear random in practically speaking every universe. Unable to see the other universes, the observer in each universe has no choice but to call her results random. Only the full wave function knows better.

Everett, III also derived that the statistics of the apparently random sequences are proportional to the absolute squares of the eigenfunction expansion coefficients, as the orthodox interpretation says.

How about the uncertainty relationship? For spins, the relevant uncertainty relationship states that it is impossible for the spin in the up/down directions and in the front/back directions to be certain at the same time. Measuring the spin in the front/back direction will make the up/down spin uncertain. But if the spin was always up, how can it change?

This is a bit more tricky. Let's have the Mars observer do a couple of additional experiments on one of her electrons, first one front/back, and then another again up/down, to see what happens. To be more precise, let's also ask her to write the result of each measurement on a blackboard, so that there is a

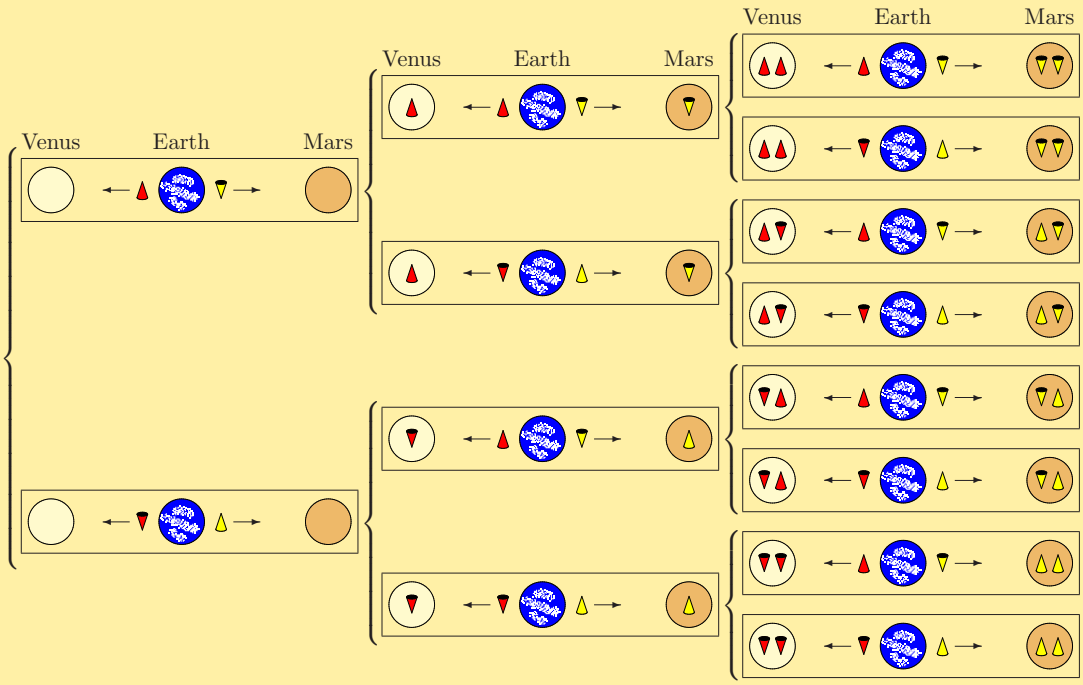


Figure 8.10: The Bohm experiment repeated.

good record of what was found. Figure 8.11 shows what happens.

When the electron is sent from Earth, two universes can be distinguished, one in which the electron is up, and another in which it is down. In the first one, the Mars observer measures the spin to be up and marks so on the blackboard. In the second, she measures and marks the spin to be down.

Next the observer in each of the two universes measures the spin front/back. Now it can be shown that the spin-up state in the first universe is a linear combination of equal amounts of spin-front and spin-back. So the second measurement splits the wave function describing the first universe into two, one with spin-front and one with spin-back.

Similarly, the spin-down state in the second universe is equivalent to equal amounts of spin-front and spin-back, but in this case with opposite sign. Either way, the wave function of the second universe still splits into a universe with spin front and one with spin back.

Now the observer in each universe does her third measurement. The front electron consists of equal amounts of spin up and spin down electrons, and so does the back electron, just with different sign. So, as the last column in figure 8.11 shows, in the third measurement, as much as half the eight universes measure the vertical spin to be the opposite of the one they got in the first measurement!

The full wave function knows that if the first four of the final eight universes are summed together, the net spin is still down (the two down spins have equal

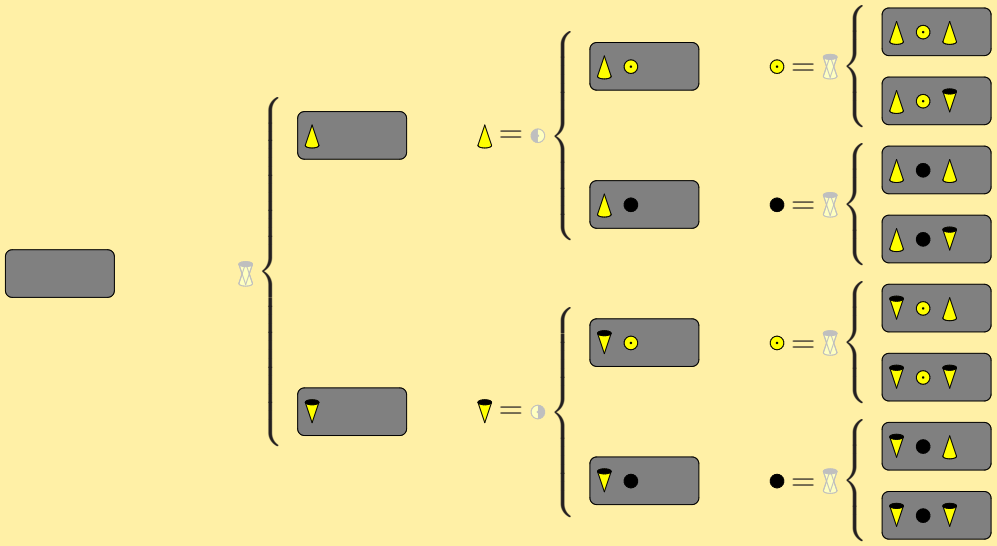


Figure 8.11: Repeated experiments on the same electron.

and opposite amplitude). But the observers have only their blackboard (and what is recorded in their brains, etcetera) to guide them. And that information seems to tell them unambiguously that the front-back measurement "destroyed" the vertical spin of the electron. (The four observers that measured the spin to be unchanged can repeat the experiment a few more times and are sure to eventually find that the vertical spin does change.)

The unavoidable conclusion is that the Schrödinger equation does *not* fail. It describes the observations exactly, in full agreement with the orthodox interpretation, without any collapse. The *appearance* of a collapse is actually just a limitation of our human observational capabilities.

Of course, in other cases than the spin example above, there are more than just two symmetric states, and it becomes much less self-evident what the proper partial solutions are. However, it does not seem hard to make some conjectures. For Schrödinger's cat, you might model the radioactive decay that gives rise to the Geiger counter going off as due to a nucleus with a neutron wave packet rattling around in it, trying to escape. As chapter 7.12.1 showed, in quantum mechanics each rattle will fall apart into a transmitted and a reflected wave. The transmitted wave would describe the formation of a universe where the neutron escapes at that time to set off the Geiger counter which kills the cat, and the reflected wave a universe where the neutron is still contained.

For the standard quantum mechanics example of an excited atom emitting a photon, a model would be that the initial excited atom is perturbed by the ambient electromagnetic field. The perturbations will turn the atom into a linear combination of the excited state with a bit of a lower energy state thrown in, surrounded by a perturbed electromagnetic field. Presumably this situation

can be taken apart in a universe with the atom still in the excited state, and the energy in the electromagnetic field still the same, and another universe with the atom in the lower energy state with a photon escaping in addition to the energy in the original electromagnetic field. Of course, the process would repeat for the first universe, producing an eventual series of universes in almost all of which the atom has emitted a photon and thus transitioned to a lower energy state.

So this is where we end up. The equations of quantum mechanics describe the physics that we observe perfectly well. Yet they have forced us to the uncomfortable conclusion that, mathematically speaking, we are not at all unique. Beyond our universe, the mathematics of quantum mechanics requires an infinity of unobservable other universes that are nontrivially different from us.

Note that the existence of an infinity of universes is not the issue. They are already required by the very formulation of quantum mechanics. The wave function of say an arsenic atom already assigns a nonzero probability to every possible configuration of the positions of the electrons. Similarly, a wave function of the universe will assign a nonzero probability to every possible configuration of the universe, in other words, to every possible universe. The existence of an infinity of universes is therefore not something that should be ascribed to Everett, III {N.15}.

However, when quantum mechanics was first formulated, people quite obviously believed that, practically speaking, there would be just one universe, the one we observe. No serious physicist would deny that the monitor on which you may be reading this has uncertainty in its position, yet the uncertainty you are dealing with here is so astronomically small that it can be ignored. Similarly it might appear that all the other substantially different universes should have such small probabilities that they can be ignored. The actual contribution of Everett, III was to show that this idea is not tenable. Nontrivial universes *must* develop that are substantially different.

Formulated in 1957 and then largely ignored, Everett's work represents without doubt one of the human race's greatest accomplishments; a stunning discovery of what we are and what is our place in the universe.

8.7 The Arrow of Time

This section has some further musings on the many worlds interpretation. One question is why it matters. What is wrong with postulating a fairy-tale collapse mechanism that makes people feel unique? The alternate realities are fundamentally unobservable, so in normal terms they do truly not exist. For all practical purposes, the wave function really does collapse.

The main reason is of course because people are curious. We would also want to understand what nature is really all about, even if we may not like the

answer very much.

But there is also a more practical side. An understanding of nature can help guess what is likely to happen under circumstances that are not well known. And clearly, there is a difference in thinking. The Everett model is a universe following the established equations of physics in which observers only observe a very narrow and evolving part of a much larger reality. The Copenhagen model is a single universe run by gnomes that allow microscopic deviations from a unique reality following the equations of physics, but kindly eliminate anything bigger.

One major difference is what is considered to be *real*. In Everett's theory, for an observer reality is not the complete wave function but a small selection of it. That becomes a philosophical point when considering "vacuum energy." According to quantum field theory, even empty space still contains half a photon of electromagnetic energy at each frequency, {A.23.4}. That is much like a harmonic oscillator still has half a quantum of kinetic and potential energy left in its ground state. The electric and magnetic fields have quantum uncertainty. If you "measure" the electric or magnetic field in vacuum, you will get a nonzero value. The same applies to other fields of particles. Unfortunately, if you sum these energies over all frequencies, you get infinity. Even if the frequencies are assumed to be limited to scales about which there is solid knowledge, there is still an enormous amount of energy here. Its gravitational effect should be gigantic, it should dwarf anything else.

Somehow that does not happen. Now, in Everett's interpretation a particle only becomes real for a universe when a state is established in which there is no doubt that the particle exists. That obviously greatly limits the vacuum energy that affects that universe. The existence of other particles might be firmly established in other universes, but these will then affect those other universes. In the Copenhagen interpretation, however, there are no other universes, and therefore no good reason to exclude any vacuum energy from affecting the gravity of the only universe there is.

Then there is the arrow of time. It is observed that time has directionality. So why does time only go one way, from early to late? You might argue that "early" and "late" are just words. But they are not. They are given meaning by the second law of thermodynamics. This law says that a measurable definition of disorder in the observed universe, called entropy, always increases with time. The law applies to macroscopic systems. However, macroscopic systems consist of particles that satisfy microscopic mechanics. And the Schrödinger equation has no particular preference for the time t above the backward time $-t$. So what happened to the processes that run according to $-t$, backwards to what we would consider forward in time? Why do we not observe such processes? And why are we composed of matter, not antimatter? And why does nature not look the same when viewed in the mirror? What is so different about a mirror image of the universe that we observe?

The conventional view postulates ad-hoc asymmetries that "just happened" to be that way. Why would that happen and why would it be in the same direction everywhere in an infinite space-time and an infinity of possible universes therein?

Then the conventional view adds evolution equations that magnify that asymmetry using small perturbation theory. That sounds reasonable until you examine those evolution equations more closely, chapter 11.10. The mechanism that provides the increasing asymmetry is, you guessed it, exactly that poorly defined collapse mechanism. Collapse is simply stated to apply for times greater than the "measurement" time. Obviously that produces asymmetry in time. But why could the collapse not apply for times less than the collapse time instead?

Now stand back from the details and take a look at the larger philosophical question. The well established equations of nature have no particular preference for either direction of time. True, the direction of time is correlated with matter versus antimatter, and with mirror symmetry. But that still does not make either direction of time any better than the other. According to the laws of physics that have been solidly established, there does not seem to be any big reason for nature to prefer one direction of time above the other.

According to Everett's theory, there is no reason to assume that it does. The many-worlds interpretation allows the wave function to describe both universes that are observed to evolve towards one direction of time and universes that are observed to evolve in the other direction.

That is not a trivial observation. The problem of the observed time asymmetry for a symmetric physics has now been removed. It has been replaced by the question why forward evolving systems appear to correlate with forward evolving systems, and backward evolving systems with backward evolving ones. While that is not a trivial question either, it is not implausible.

Perhaps, if we spend more time on listening to what nature is really telling us, rather than make up stories for what we want to believe, we would now understand those processes a lot more clearly.

Part III

Gateway Topics

Chapter 9

Numerical Procedures

Since analytical solutions in quantum mechanics are extremely limited, numerical solution is essential. This chapter outlines some of the most important ideas. The most glaring omission at this time is the DFT (Density Functional Theory.) A writer needs a sabbatical.

9.1 The Variational Method

Solving the equations of quantum mechanics is typically difficult, so approximations must usually be made. One very effective tool for finding approximate solutions is the variational principle. This section gives some of the basic ideas, including ways to apply it best.

9.1.1 Basic variational statement

Finding the state of a physical system in quantum mechanics means finding the wave function Ψ that describes it. For example, at sufficiently low temperatures, physical systems will be described by the ground state wave function. The problem is that if there are more than a couple of particles in the system, the wave function is a very high-dimensional function. It is far too complex to be crunched out using brute force on any current computer.

However, the expectation value of energy is just a simple single number for any given wave function. It is defined as

$$\langle E \rangle = \langle \Psi | H \Psi \rangle$$

where H is the Hamiltonian of the system. The key observation on which the variational method is based is that the ground state is the state among all allowable wave functions that has the lowest expectation value of energy:

$$\boxed{\langle E \rangle \text{ is minimal for the ground state wave function.}} \tag{9.1}$$

That means that if you would find $\langle E \rangle$ for all possible system wave functions, you would be able to pick out the ground state simply as the state that has the lowest value.

Of course, finding the expectation value of the energy for all possible wave functions is still an impossible task. But you may be able to guess a generic type of wave function that you would expect to be able to approximate the ground state well, under suitable conditions. Normally, "suitable conditions" means that the approximation will be good only if various parameters appearing in the approximate wave function are well chosen.

That then leaves you with the much smaller task of finding good values for this limited set of parameters. Here the key idea is:

$$\boxed{\langle E \rangle \text{ is lowest for the best approximation to the ground state.}} \tag{9.2}$$

Following that idea, what you do is adjust the parameters values so that you get the lowest possible value of the expectation energy for your type of approximate wave function. The true ground state wave function always has the lowest possible energy, so the lower you make your approximate energy, the closer that energy is to the exact value.

So this procedure gives you the best possible approximation to the true energy, and energy is usually the key quantity in quantum mechanics. In addition you know for sure that the true energy must be lower than your approximation, which is also often very useful information.

The variational method as described above has already been used earlier in this book to find an approximate ground state for the hydrogen molecular ion, chapter 4.6, and for the hydrogen molecule, chapter 5.2. It will also be used to find an approximate ground state for the helium atom, {A.38.2}. The method works quite well even for the crude approximate wave functions used in those examples.

To be sure, it is not at all obvious that getting the best energy will also produce the best wave function. After all, "best" is a somewhat tricky term for a complex object like a wave function. To take an example from another field, surely you would not argue that the best *sprinter* in the world must also be the best *person* in the world.

But in this case, your wave function will in fact be close to the exact wave function if you manage to get close enough to the exact energy. More precisely, assuming that the ground state is unique, the closer your energy gets to the exact energy, the closer your wave function gets to the exact wave function. One way of thinking about it is to note that your approximate wave function is always a combination of the desired exact ground state plus polluting amounts of higher energy states. By minimizing the energy, in some sense you minimize the amount of these polluting higher energy states. The mathematics of that idea is explored in more detail in addendum {A.7}.

And there are other benefits to specifically getting the energy as accurate as possible. One problem is often to figure out whether a system is bound. For example, can you add another electron to a hydrogen atom and have that electron at least weakly bound? The answer is not obvious. But if using a suitable approximate solution, you manage to show that the *approximate* energy of the bound system is less than that of having the additional electron at infinity, then you have *proved* that the bound state exist. Despite the fact that your solution has errors. The reason is that, by definition, the ground state must have lower energy than your approximate wave function. So the ground state is even more tightly bound together than your approximate wave function says.

Another reason to specifically getting the energy as accurate as possible is that energy values are directly related to how fast systems evolve in time when not in the ground state, chapter 7.

For the above reasons, it is also great that the errors in energy turn out to be unexpectedly small in a variational procedure, when compared to the errors in the guessed wave function, {A.7}.

To get the second lowest energy state, you could search for the lowest energy among all wave functions orthogonal to the ground state. But since you would not know the exact ground state, you would need to use your approximate one instead. That would involve some error, and it is no longer sure that the true second-lowest energy level is no higher than what you compute, but anyway. The suprising accuracy in energy will still apply.

If you want to get truly accurate results in a variational method, in general you will need to increase the number of parameters. The molecular example solutions were based on the atomic ground states, and you could consider adding some excited states to the mix. In general, a procedure using appropriate guessed functions is called a Rayleigh-Ritz method. Alternatively, you could just chop space up into little pieces, or "elements," and use a simple polynomial within each piece. That is called a finite-element method. In either case, you end up with a finite, but relatively large number of unknowns; the parameters and coefficients of the functions, or the coefficients of the polynomials.

9.1.2 Differential form of the statement

You might by now wonder about the wisdom of trying to find the minimum energy by searching through the countless possible combinations of a lot of parameters. Brute-force search worked fine for the hydrogen molecule examples since they really only depended nontrivially on the distance between the nuclei. But if you add some more parameters for better accuracy, you quickly get into trouble. Semi-analytical approaches like Hartree-Fock even leave whole functions unspecified. In that case, simply put, every single function value is an unknown parameter, and a function has infinitely many of them. You would be searching in an infinite-dimensional space, and might search forever.

Usually it is a much better idea to write some equations for the minimum energy first. From calculus, you know that if you want to find the minimum of a function, the sophisticated way to do it is to note that the derivatives of the function must be zero at the minimum. Less rigorously, but a lot more intuitive, at the minimum of a function the changes in the function due to *small* changes in the variables that it depends on must be zero. Mathematicians may not like that, since the word "small" has no rigorous meaning. But unless you misuse your small quantities, you can always convert your results using them to rigorous mathematics after the fact.

In the simplest possible example of a function $f(x)$ of one variable x, a rigorous mathematician would say that at a minimum, the derivative $f'(x)$ must be zero. But a physicists may not like that, for if you say derivative, you must say with respect to what variable; you must say what x is as well as what f is. There is often more than one possible choice for x, with none preferred under all circumstances. So a typical physicist would say that the change $\mathrm{d}f$ in f due to a small change in whatever variable it depends on must be zero. It is the same thing, since for a small enough change $\mathrm{d}x$ in the variable, $\mathrm{d}f = f'\mathrm{d}x$, so that if f' is zero, then so is $\mathrm{d}f$. (Mathematically more accurately, if $\mathrm{d}x$ becomes small enough, $\mathrm{d}f$ becomes zero *compared to* $\mathrm{d}x$.) If there is more than one independent variable that the function depends on, then the derivatives become partial derivatives, $\mathrm{d}f$ becomes ∂f, and specifying the precise derivatives would become much messier still.

In variational procedures, it is common to use δf instead of $\mathrm{d}f$ or ∂f for the small change in f. This book will do so too.

So in quantum mechanics, the fact that the expectation energy must be minimal in the ground state can be written as:

$$\boxed{\delta \langle E \rangle = 0 \text{ for all acceptable small changes in wave function}} \tag{9.3}$$

The changes must be acceptable; you cannot allow that the changed wave function is no longer normalized. Also, if there are boundary conditions, the changed wave function should still satisfy them. (There may be exceptions permitted to the latter under some conditions, but these will be ignored here.) So, in general you have "constrained minimization;" you cannot make your changes completely arbitrary.

9.1.3 Using Lagrangian multipliers

As an example of how the variational formulation of the previous subsection can be applied analytically, and how it can also describe eigenstates of higher energy, this subsection will work out a very basic example. The idea is to figure out what you get if you truly zero the changes in the expectation value of energy $\langle E \rangle = \langle \psi | H | \psi \rangle$ over *all* acceptable wave functions ψ. (Instead of just over all

possible versions of a numerical approximation, say.) It will illustrate how the "Lagrangian multiplier" method can deal with the constraints.

The differential statement is:

$$\delta\langle\psi|H|\psi\rangle = 0 \text{ for all acceptable changes } \delta\psi \text{ in } \psi$$

But "acceptable" is not a mathematical concept. What does it mean? Well, if it is assumed that there are no boundary conditions, (like the harmonic oscillator, but unlike the particle in a pipe,) then acceptable just means that the wave function must remain normalized under the change. So the change in $\langle\psi|\psi\rangle$ must be zero, and you can write more specifically:

$$\delta\langle\psi|H|\psi\rangle = 0 \text{ whenever } \delta\langle\psi|\psi\rangle = 0.$$

But how do you crunch a statement like that down mathematically? Well, there is a very important mathematical trick to simplify this. Instead of rigorously trying to enforce that the changed wave function is still normalized, just allow *any* change in wave function. But add "penalty points" to the change in expectation energy if the change in wave function goes out of allowed bounds:

$$\delta\langle\psi|H|\psi\rangle - \epsilon\delta\langle\psi|\psi\rangle = 0$$

Here ϵ is the penalty factor. Such penalty factors are called "Lagrangian multipliers" after a famous mathematician who probably watched a lot of soccer. For a change in wave function that does not go out of bounds, the second term is zero, so nothing changes. And if the change does go out of bounds, the second term will cancel any resulting erroneous gain or decrease in expectation energy, {D.48}, assuming that the penalty factor is carefully tuned. Note that the penalty factor ϵ must be real because the other two quantities in the equation above are changes in real functions.

You do not, however, have to explicitly tune the penalty factor yourself. All you need to know is that a proper one exists. In actual application, all you do in addition to ensuring that the penalized change in expectation energy is zero is ensure that at least the *un*changed wave function is normalized. It is really a matter of counting equations versus unknowns. Compared to simply setting the change in expectation energy to zero with no constraints on the wave function, one additional unknown has been added, the penalty factor. And quite generally, if you add one more unknown to a system of equations, you need one more equation to still have a unique solution. As the one-more equation, use the normalization condition. With enough equations to solve, you will get the correct solution, which means that the implied value of the penalty factor should be OK too.

So what does this variational statement now produce? Writing out the differences explicitly, you must have

$$\Big(\langle\psi+\delta\psi|H|\psi+\delta\psi\rangle - \langle\psi|H|\psi\rangle\Big) - \epsilon\Big(\langle\psi+\delta\psi|\psi+\delta\psi\rangle - \langle\psi|\psi\rangle\Big) = 0$$

Multiplying out, canceling equal terms and ignoring terms that are quadratically small in $\delta\psi$, you get

$$\langle\delta\psi|H|\psi\rangle + \langle\psi|H|\delta\psi\rangle - \epsilon\Big(\langle\delta\psi|\psi\rangle + \langle\psi|\delta\psi\rangle\Big) = 0$$

Remarkably, you can throw away the second of each pair of inner products in the expression above. To see why, remember that you can allow any change $\delta\psi$ you want, including the $\delta\psi$ you are now looking at times $-$i. If you plug that into the above equation and divide the entire thing by i to get rid of the added factors i again, you get

$$\langle\delta\psi|H|\psi\rangle - \langle\psi|H|\delta\psi\rangle - \epsilon\Big(\langle\delta\psi|\psi\rangle - \langle\psi|\delta\psi\rangle\Big) = 0$$

The two additional minus signs arise because an $-$i comes out of the *left* side of an inner product as i, but out of the right side as $-$i. Averaging this equation with the original above it has the effect of throwing away the second of each pair of inner products in the original equation.

You can now combine the remaining two terms into one inner product with $\delta\psi$ on the left:

$$\langle\delta\psi|H\psi - \epsilon\psi\rangle = 0$$

If this is to be zero for *any* change $\delta\psi$, then the right hand side of the inner product must unavoidably be zero. For example, just take $\delta\psi$ equal to a small number ε times the right hand side, you will get ε times the square norm of the right hand side, and that can only be zero if the right hand side is. So $H\psi - \epsilon\psi = 0$, or

$$H\psi = \epsilon\psi.$$

So you see that you have recovered the Hamiltonian eigenvalue problem from the requirement that the variation of the expectation energy is zero. Unavoidably then, ϵ will have to be an energy eigenvalue E. It often happens that Lagrangian multipliers have a physical meaning beyond being merely penalty factors. But note that there is no requirement for this to be the ground state. Any energy eigenstate would satisfy the equation; the variational principle works for them all.

Indeed, you may remember from calculus that the derivatives of a function may be zero at more than one point. For example, a function might also have a maximum, or local minima and maxima, or stationary points where the function is neither a maximum nor a minimum, but the derivatives are zero anyway. This sort of thing happens here too: the ground state is the state of lowest possible energy, but there will be other states for which $\delta\langle E\rangle$ is zero, and these will correspond to energy eigenstates of higher energy, {D.49}.

9.2 The Born-Oppenheimer Approximation

Exact solutions in quantum mechanics are hard to come by. In almost all cases, approximation is needed. The Born-Oppenheimer approximation in particular is a key part of real-life quantum analysis of atoms and molecules and the like. The basic idea is that the uncertainty in the nuclear positions is too small to worry about when you are trying to find the wave function for the electrons. That was already assumed in the earlier approximate solutions for the hydrogen molecule and molecular ion. This section discusses the approximation, and how it can be used, in more depth.

9.2.1 The Hamiltonian

The general problem to be discussed in this section is that of a number of electrons around a number of nuclei. You first need to know what is the true problem to be solved, and for that you need the Hamiltonian.

This discussion will be restricted to the strictly nonrelativistic case. Corrections for relativistic effects on energy, including those involving spin, can in principle be added later, though that is well beyond the scope of this book. The physical problem to be addressed is that there are a finite number I of electrons around a finite number J of nuclei in otherwise empty space. That describes basic systems of atoms and molecules, but modifications would have to be made for ambient electric and magnetic fields and electromagnetic waves, or for the infinite systems of electrons and nuclei used to describe solids.

The electrons will be numbered using an index i, and whenever there is a second electron involved, its index will be called $\underline{i}$. Similarly, the nuclei will be numbered with an index j, or $\underline{j}$ where needed. The nuclear charge of nucleus number j, i.e. the number of protons in that nucleus, will be indicated by Z_j, and the mass of the nucleus by $m_j^{\rm n}$. Roughly speaking, the mass $m_j^{\rm n}$ will be the sum of the masses of the protons and neutrons in the nucleus; however, internal nuclear energies are big enough that there are noticeable relativistic deviations in total nuclear rest mass from what you would think. All the electrons have the same mass $m_{\rm e}$ since relativistic mass changes due to motion are ignored.

Under the stated assumptions, the Hamiltonian of the system consists of a number of contributions that will be looked at one by one. First there is the kinetic energy of the electrons, the sum of the kinetic energy operators of the individual electrons:

$$\widehat{T}^{\rm E} = -\sum_{i=1}^{I} \frac{\hbar^2}{2m_{\rm e}} \nabla_i^2 = -\sum_{i=1}^{I} \frac{\hbar^2}{2m_{\rm e}} \left(\frac{\partial^2}{\partial {r_{1i}}^2} + \frac{\partial^2}{\partial {r_{2i}}^2} + \frac{\partial^2}{\partial {r_{3i}}^2} \right). \tag{9.4}$$

where $\vec{r}_i = (r_{1i}, r_{2i}, r_{3i})$ is the position of electron number i. Note the use of (r_1, r_2, r_3) as the notation for the components of position, rather than (x, y, z).

For more elaborate mathematics, the index notation (r_1, r_2, r_3) is often more convenient, since you can indicate any generic component by the single expression r_α, (with the understanding that $\alpha = 1$, 2, or 3,) instead of writing them out all three separately.

Similarly, there is the kinetic energy of the nuclei,

$$\widehat{T}^{\rm N} = -\sum_{j=1}^{J} \frac{\hbar^2}{2m_j^{\rm n}} \nabla_j^{\rm n2} = -\sum_{j=1}^{J} \frac{\hbar^2}{2m_j^{\rm n}} \left(\frac{\partial^2}{\partial {r_{1j}^{\rm n}}^2} + \frac{\partial^2}{\partial {r_{2j}^{\rm n}}^2} + \frac{\partial^2}{\partial {r_{3j}^{\rm n}}^2} \right). \tag{9.5}$$

where $\vec{r}_j^{\,\rm n} = (r_{1j}^{\rm n}, r_{2j}^{\rm n}, r_{3j}^{\rm n})$ is the position of nucleus number j.

Next there is the potential energy due to the attraction of the I electrons by the J nuclei. That potential energy is, summing over all electrons and over all nuclei:

$$V^{\rm NE} = -\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{Z_j e^2}{4\pi\epsilon_0} \frac{1}{r_{ij}} \tag{9.6}$$

where $r_{ij} \equiv |\vec{r}_i - \vec{r}_j^{\,\rm n}|$ is the distance between electron number i and nucleus number j, and $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is the permittivity of space.

Next there is the potential energy due to the electron-electron repulsions:

$$V^{\rm EE} = \tfrac{1}{2} \sum_{i=1}^{I} \sum_{\substack{\underline{i}=1 \\ \underline{i}\neq i}}^{I} \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{i\underline{i}}} \tag{9.7}$$

where $r_{i\underline{i}} \equiv |\vec{r}_i - \vec{r}_{\underline{i}}|$ is the distance between electron number i and electron number $\underline{i}$. Half of this repulsion energy will be attributed to electron i and half to electron $\underline{i}$, accounting for the factor $\frac{1}{2}$.

Finally, there is the potential energy due to the nucleus-nucleus repulsions,

$$V^{\rm NN} = \tfrac{1}{2} \sum_{j=1}^{J} \sum_{\substack{\underline{j}=1 \\ \underline{j}\neq j}}^{J} \frac{Z_j Z_{\underline{j}} e^2}{4\pi\epsilon_0} \frac{1}{r_{j\underline{j}}}, \tag{9.8}$$

where $r_{j\underline{j}} \equiv |\vec{r}_j^{\,\rm n} - \vec{r}_{\underline{j}}^{\,\rm n}|$ is the distance between nucleus number j and nucleus number $\underline{j}$.

Solving the full quantum problem for this system of electrons and nuclei exactly would involve finding the eigenfunctions ψ to the Hamiltonian eigenvalue problem

$$\left[\widehat{T}^{\rm E} + \widehat{T}^{\rm N} + V^{\rm NE} + V^{\rm EE} + V^{\rm NN} \right] \psi = E\psi \tag{9.9}$$

Here ψ is a function of the position and spin coordinates of all the electrons and all the nuclei, in other words:

$$\psi = \psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_I, S_{zI},\ \vec{r}_1^{\,\rm n}, S_{z1}^{\rm n}, \vec{r}_2^{\,\rm n}, S_{z2}^{\rm n}, \ldots, \vec{r}_J^{\,\rm n}, S_{zJ}^{\rm n}) \tag{9.10}$$

You might guess solving this problem is a tall order, and you would be perfectly right. It can only be done analytically for the very simplest case of one electron and one nucleus. That is the hydrogen atom solution, using an effective electron mass to include the nuclear motion. For any decent size system, an accurate numerical solution is a formidable task too.

9.2.2 Basic Born-Oppenheimer approximation

The general idea of the Born-Oppenheimer approximation is simple. First note that the nuclei are thousands of times heavier than the electrons. A proton is almost two thousand times heavier than an electron, and that does not even count any neutrons in the nuclei.

So, if you take a look at the kinetic energy operators of the two,

$$\widehat{T}^{\rm E} = -\sum_{i=1}^{I} \frac{\hbar^2}{2m_{\rm e}} \left(\frac{\partial^2}{\partial {r_{1i}}^2} + \frac{\partial^2}{\partial {r_{2i}}^2} + \frac{\partial^2}{\partial {r_{3i}}^2} \right)$$

$$\widehat{T}^{\rm N} = -\sum_{j=1}^{J} \frac{\hbar^2}{2m_j^{\rm n}} \left(\frac{\partial^2}{\partial {r_{1j}^{\rm n}}^2} + \frac{\partial^2}{\partial {r_{2j}^{\rm n}}^2} + \frac{\partial^2}{\partial {r_{3j}^{\rm n}}^2} \right)$$

then what would seem more reasonable than to ignore the kinetic energy $\widehat{T}^{\rm N}$ of the nuclei? It has those heavy masses in the bottom.

An alternative, and better, way of phrasing the assumption that $\widehat{T}^{\rm N}$ can be ignored is to say that you ignore the uncertainty in the positions of the nuclei. For example, visualize the hydrogen molecule, figure 5.2. The two protons, the nuclei, have pretty well defined positions in the molecule, while the electron wave function extends over the entire region like a big blob of possible measurable positions. So how important could the uncertainty in position of the nuclei really be?

Assuming that the nuclei do not suffer from quantum uncertainty in position is really equivalent to putting $\hbar$ to zero in their kinetic energy operator above, making the operator disappear, because $\hbar$ is nature's measure of uncertainty. And without a kinetic energy term for the nuclei, there is nothing left in the mathematics to force them to have uncertain positions. Indeed, you can now just guess numerical *values* for the positions of the nuclei, and solve the approximated eigenvalue problem $H\psi = E\psi$ for those assumed values.

That thought is the Born-Oppenheimer approximation in a nutshell. Just do the electrons, assuming suitable positions for the nuclei a priori. The solutions that you get doing so will be called $\psi^{\rm E}$ to distinguish them from the true solutions ψ that do not use the Born-Oppenheimer approximation. Mathematically $\psi^{\rm E}$ will still be a function of the electron and nuclear positions:

$$\psi^{\rm E} = \psi^{\rm E}(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_I, S_{zI};\ \vec{r}_1^{\,\rm n}, S_{z1}^{\rm n}, \vec{r}_2^{\,\rm n}, S_{z2}^{\rm n}, \ldots, \vec{r}_J^{\,\rm n}, S_{zJ}^{\rm n}). \qquad (9.11)$$

But physically it will be a quite different thing: it describes the probability of finding the electrons, *given* the positions of the nuclei. That is why there is a semi-colon between the electron positions and the nuclear positions. The nuclear positions are here *assumed* positions, while the electron positions are *potential* positions, for which the square magnitude of the wave function $\psi^{\rm E}$ gives the probability. This is an electron wave function only.

In application, it is usually most convenient to write the Hamiltonian eigenvalue problem for the electron wave function as

$$\left[\widehat{T}^{\rm E} + V^{\rm NE} + V^{\rm EE} + V^{\rm NN}\right] \psi^{\rm E} = (E^{\rm E} + V^{\rm NN})\psi^{\rm E},$$

which just means that the eigenvalue is called $E^{\rm E} + V^{\rm NN}$ instead of simply $E^{\rm E}$. The reason is that you can then get rid of $V^{\rm NN}$, and obtain the electron wave function eigenvalue problem in the more concise form

$$\boxed{\left[\widehat{T}^{\rm E} + V^{\rm NE} + V^{\rm EE}\right] \psi^{\rm E} = E^{\rm E}\psi^{\rm E}} \tag{9.12}$$

After all, for given nuclear coordinates, $V^{\rm NN}$ is just a bothersome constant in the solution of the electron wave function that you may just as well get rid of.

Of course, after you compute your electron eigenfunctions, you want to get something out of the results. Maybe you are looking for the ground state of a molecule, like was done earlier for the hydrogen molecule and molecular ion. In that case, the simplest approach is to try out various nuclear positions and for each likely set of nuclear positions compute the electronic ground state energy $E^{\rm E}_{\rm gs}$, the lowest eigenvalue of the electronic problem (9.12) above.

For different assumed nuclear positions, you will get different values for the electronic ground state energy, and the nuclear positions corresponding to the actual ground state of the molecule will be the ones for which the total energy is least:

$$\boxed{\text{nominal ground state condition: } E^{\rm E}_{\rm gs} + V^{\rm NN} \text{ is minimal}} \tag{9.13}$$

This is what was used to solve the hydrogen molecule cases discussed in earlier chapters; a computer program was written to print out the energy $E^{\rm E}_{\rm gs} + V^{\rm NN}$ for a lot of different spacings between the nuclei, allowing the spacing that had the lowest total energy to be found by skimming down the print-out. That identified the ground state. The biggest error in those cases was not in using the Born-Oppenheimer approximation or the nominal ground state condition above, but in the crude way in which the electron wave function for given nuclear positions was approximated.

For more accurate work, the nominal ground state condition (9.13) above does have big limitations, so the next subsection discusses a more advanced approach.

9.2.3 Going one better

Solving the wave function for electrons only, given positions of the nuclei is definitely a big simplification. But identifying the ground state as the position of the nuclei for which the electron energy plus nuclear repulsion energy is minimal is much less than ideal.

Such a procedure ignores the motion of the nuclei, so it is no use for figuring out any molecular dynamics beyond the ground state. And even for the ground state, it is really wrong to say that the nuclei are at the position of minimum energy, because the uncertainty principle does not allow precise positions for the nuclei.

Instead, the nuclei behave much like the particle in a harmonic oscillator. They are stuck in an electron blob that wants to push them to their nominal positions. But uncertainty does not allow that, and the wave function of the nuclei spreads out a bit around the nominal positions, adding both kinetic and potential energy to the molecule. One example effect of this "zero point energy" is to lower the required dissociation energy a bit from what you would expect otherwise.

It is not a big effect, maybe on the order of tenths of electron volts, compared to typical electron energies described in terms of multiple electron volts (and much more for the inner electrons in all but the lightest atoms.) But it is not as small as might be guessed based on the fact that the nuclei are at least thousands of times heavier than the electrons.

Moreover, though relatively small in energy, the motion of the nuclei may actually be the one that is physically the important one. One reason is that the electrons tend to get stuck in single energy states. That may be because the differences between electron energy levels tend to be so large compared to a typical unit $\frac{1}{2}kT$ of thermal energy, about one hundredth of an electron volt, or otherwise because they tend to get stuck in states for which the next higher energy levels are already filled with other electrons. The interesting physical effects then become due to the seemingly minor nuclear motion.

For example, the heat capacity of typical diatomic gases, like the hydrogen molecule or air under normal conditions, is not in any direct sense due to the electrons; it is kinetic energy of translation of the molecules plus a comparable energy due to angular momentum of the molecule; read, angular motion of the nuclei around their mutual center of gravity. The heat capacity of solids too is largely due to nuclear motion, as is the heat conduction of non metals.

For all those reasons, you would really, really, like to actually compute the motion of the nuclei, rather than just claim they are at fixed points. Does that mean that you need to go back and solve the combined wave function for the complete system of electrons plus nuclei anyway? Throw away the Born-Oppenheimer approximation results?

Fortunately, the answer is mostly no. It turns out that nature is quite coop-

erative here, for a change. After you have done the electronic structure computations for all relevant positions of the nuclei, you can proceed with computing the motion of nuclei as a separate problem. For example, if you are interested in the ground state nuclear motion, it is governed by the Hamiltonian eigenvalue problem

$$\left[\widehat{T}^{\rm N} + V^{\rm NN} + E_1^{\rm E}\right]\psi_1^{\rm N} = E\psi_1^{\rm N}$$

where $\psi_1^{\rm N}$ is a wave function involving the nuclear coordinates only, *not* any electronic ones. The trick is in the potential energy to use in such a computation; it is not just the potential energy of nucleus to nucleus repulsions, but you must include an additional energy $E_1^{\rm E}$.

So, what is this $E_1^{\rm E}$? Easy, it is the electronic ground state energy $E_{\rm gs}^{\rm E}$ that you computed for assumed positions of the nuclei. So it will depend on where the nuclei are, but it does *not* depend on where the electrons are. You can just computed $E_1^{\rm E}$ for a sufficient number of relevant nuclear positions, tabulate the results somehow, and interpolate them as needed. $E_1^{\rm E}$ is then a known function function of the nuclear positions and so is $V^{\rm NN}$. Proceed to solve for the wave function for the nuclei $\psi_1^{\rm N}$ as a problem not directly involving any electrons.

And it does not necessarily have to be just to compute the ground state. You might want to study thermal motion or whatever. As long as the electrons are not kicked strongly enough to raise them to the next energy level, you can assume that they are in their ground state, even if the nuclei are not. The usual way to explain this is to say something like that the electrons "move so fast compared to the slow nuclei that they have all the time in the world to adjust themselves to whatever the electronic ground state is for the current nuclear positions."

You might even decide to use classical molecular dynamics based on the potential $V^{\rm NN} + E_1^{\rm E}$ instead of quantum mechanics. It would be much faster and easier, and the results are often good enough.

So what if you are interested in what your molecule is doing when the electrons are at an elevated energy level, instead of in their ground state? Can you still do it? Sure. If the electrons are in an elevated energy level $E_n^{\rm E}$, (for simplicity, it will be assumed that the electron energy levels are numbered with a single index n,) just solve

$$\boxed{\left[\widehat{T}^{\rm N} + V^{\rm NN} + E_n^{\rm E}\right]\psi_n^{\rm N} = E\psi_n^{\rm N}} \tag{9.14}$$

or equivalent.

Note that for a different value of n, this is truly a different motion problem for the nuclei, since the potential energy will be different. If you are a visual sort of person, you might vaguely visualize the potential energy for a given value of n plotted as a surface in some high-dimensional space, and the state of the nuclei moving like a roller-coaster along that potential energy surface, speeding

up when the surface goes down, slowing down if it goes up. There is one such surface for each value of n. Anyway. The bottom line is that people refer to these different potential energies as "potential energy surfaces." They are also called "adiabatic surfaces" because "adiabatic" normally means processes sufficiently fast that heat transfer can be ignored. So, some quantum physicists figured that it would be a good idea to use the same term for quantum processes that are so slow that quasi-equilibrium conditions persist throughout, and that have nothing to do with heat transfer.

Of course, any approximation can fail. It is possible to get into trouble solving your problem for the nuclei as explained above. The difficulties arise if two electron energy levels, call them $E^{\rm E}_n$ and $E^{\rm E}_{\underline{n}}$, become almost equal, and in particular when they cross. In simple terms, the difficulty is that if energy levels are equal, the energy eigenfunctions are not unique, and the slightest thing can throw you from one eigenfunction to the completely different one.

You might now get alarmed, because for example the hydrogen molecular ion *does* have two different ground state solutions with the same energy. Its single electron can be in either the spin-up state or the spin down state, and it does not make any difference for the energy because the assumed Hamiltonian does not involve spin. In fact, all systems with an odd number of electrons will have a second solution with all spins reversed and the same energy {D.50}. There is no need to worry, though; these reversed-spin solutions go their own way and do not affect the validity of (9.14). It is spatial, rather than spin nonuniqueness that is a concern.

There is a derivation of the nuclear eigenvalue problem (9.14) in derivation {D.51}, showing what the ignored terms are and why they can usually be ignored.

9.3 The Hartree-Fock Approximation

Many of the most important problems that you want to solve in quantum mechanics are all about atoms and/or molecules. These problems involve a number of electrons around a number of atomic nuclei. Unfortunately, a full quantum solution of such a system of any nontrivial size is very difficult. However, approximations can be made, and as section 9.2 explained, the real skill you need to master is solving the wave function for the electrons given the positions of the nuclei.

But even given the positions of the nuclei, a brute-force solution for any nontrivial number of electrons turns out to be prohibitively laborious. The Hartree-Fock approximation is one of the most important ways to tackle that problem, and has been so since the early days of quantum mechanics. This section explains some of the ideas.

9.3.1 Wave function approximation

The key to the basic Hartree-Fock method is the assumptions it makes about the form of the electron wave function. It will be assumed that there are a total of I electrons in orbit around a number of nuclei. The wave function describing the set of electrons then has the general form:

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \ldots, \vec{r}_i, S_{zi}, \ldots \vec{r}_I, S_{zI})$$

where $\vec{r}_i$ is the position of the electron numbered i, and S_{zi} its spin (i.e. internal angular momentum) in a chosen z-direction. Recall that while the position of an electron can be anywhere in three-dimensional space, its spin component S_z can have only two measurable values: $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$. Because of the factor $\frac{1}{2}$, an electron is a "particle of spin one-half". Such a particle is also called a "spin doublet" because of the two possible spin values.

The square magnitude of the wave function above gives the probability for the electrons $i = 1, 2, \ldots, I$ to be near the position $\vec{r}_i$, per unit volume, with spin component S_{zi}.

Of course, what the wave function is will also depend on where the nuclei are. However, in this section, the nuclei are supposed to be at given positions. Therefore to reduce the clutter, the dependence of the electron wave function on the nuclear positions will not be shown explicitly.

Hartree-Fock approximates the wave function above in terms of *single-electron* wave functions. Each single-electron wave function takes the form of a product of a spatial function $\psi^{\rm s}$ of the electron position $\vec{r}$, times a function of the electron spin component S_z. The spin function is either taken to be $\uparrow$ or $\downarrow$; by definition, function $\uparrow(S_z)$ equals 1 if the spin S_z is $\frac{1}{2}\hbar$, and 0 if it is $-\frac{1}{2}\hbar$. Conversely, function $\downarrow(S_z)$ equals 0 if S_z is $\frac{1}{2}\hbar$ and 1 if it is $-\frac{1}{2}\hbar$. Function $\uparrow$ is called "spin-up" and $\downarrow$ "spin-down."

A complete single-electron wave function is then of the form

$$\psi^{\rm s}(\vec{r})\updownarrow(S_z)$$

where $\updownarrow$ is either $\uparrow$ or $\downarrow$. Such a single-electron wave function is called an "orbital" or more accurately a "spin orbital." The reason is that people tend to think of the single-electron wave function as describing a single electron being in a particular orbit around the nuclei with a particular spin. Wrong, of course: the electrons do not have well-defined positions on these scales, so you cannot talk about "orbits" But people do tend to think of the "spatial orbitals" $\psi^{\rm s}(\vec{r})$ that way anyway.

For simplicity, it will be assumed that the spin orbitals are taken to be "normalized;" if you integrate the square magnitude of $\psi^{\rm s}(\vec{r})\updownarrow(S_z)$ over all possible positions $\vec{r}$ of the electron and sum over the two possible values of its spin S_z, you get 1. Physically that merely expresses that the electron must be at *some*

position and have *some* spin for certain (probability 1). The integral plus sum combination can be expressed using the concise bra[c]ket notation from chapter 2;

$$\langle \psi^{\rm s}\updownarrow | \psi^{\rm s}\updownarrow \rangle = 1$$

Such a bracket, or "inner product," is equivalent to a dot product for functions.

If there is more than one electron, as will be assumed in this section, a single spin orbital $\psi^{\rm s}\updownarrow$ is not enough to create a valid wave function for the complete system. In fact, the "Pauli exclusion principle" says that each of the I electrons must go into a different spin orbital, chapter 5.7. So a series of orbitals is needed,

$$\psi^{\rm s}_1\updownarrow_1, \psi^{\rm s}_2\updownarrow_2, \ldots, \psi^{\rm s}_n\updownarrow_n, \ldots, \psi^{\rm s}_N\updownarrow_N$$

where the number of orbitals N must be at least as big as the number of electrons I.

It will be assumed that any two different spin orbitals $\psi^{\rm s}_n\updownarrow_n$ and $\psi^{\rm s}_{\underline{n}}\updownarrow_{\underline{n}}$ are taken to be "orthogonal;" by definition this means that their bracket is zero:

$$\left\langle \psi^{\rm s}_n\updownarrow_n \middle| \psi^{\rm s}_{\underline{n}}\updownarrow_{\underline{n}} \right\rangle = 0 \quad \text{if} \quad n \neq \underline{n}$$

In short, it is assumed that the set of spin orbitals is orthonormal; mutually orthogonal and normalized.

Note that the bracket above can be written as a product of a spatial bracket and a spin one:

$$\left\langle \psi^{\rm s}_n\updownarrow_n \middle| \psi^{\rm s}_{\underline{n}}\updownarrow_{\underline{n}} \right\rangle \equiv \left\langle \psi^{\rm s}_n \middle| \psi^{\rm s}_{\underline{n}} \right\rangle \times \left\langle \updownarrow_n \middle| \updownarrow_{\underline{n}} \right\rangle$$

So for different spin orbitals to be orthogonal, either the spatial states or the spin states must orthogonal; they do not both need to be orthogonal. (To verify the expression above, just write the first bracket out in terms of a spatial integral over $\vec{r}$ and a sum over the two values of S_z and reorder terms.)

Note also that the spin states $\uparrow$ and $\downarrow$ are an orthonormal set:

$$\langle \uparrow | \uparrow \rangle = \langle \downarrow | \downarrow \rangle = 1 \qquad \langle \uparrow | \downarrow \rangle = \langle \downarrow | \uparrow \rangle = 0$$

So if the spin states are opposite, the spatial states do not need to be orthogonal. In fact, the spatial states can then be the same.

The base Hartree-Fock method uses the absolute minimum number of orbitals $N = I$. In that case, the simplest you could do to create a system wave function is to put electron 1 in orbital 1, electron 2 in orbital 2, etcetera. That would give the system wave function

$$\psi^{\rm s}_1(\vec{r}_1)\updownarrow_1(S_{z1})\psi^{\rm s}_2(\vec{r}_2)\updownarrow_2(S_{z2})\psi^{\rm s}_3(\vec{r}_3)\updownarrow_3(S_{z3})\ldots\psi^{\rm s}_I(\vec{r}_I)\updownarrow_I(S_{zI})$$

A product of single-electron wave functions like this is called a "Hartree product."

But a single Hartree product like the one above is physically not acceptable as a wave function. The Pauli exclusion principle is only part of what is needed, chapter 5.7. The full requirement is that a system wave function must be "antisymmetric under electron exchange:" the wave function must simply change sign when any two electrons are swapped. But if, say, electrons 1 and 2 are swapped in the Hartree product above, it produces the new Hartree product

$$\psi^{\rm s}_1(\vec r_2)\updownarrow_1(S_{z2})\psi^{\rm s}_2(\vec r_1)\updownarrow_2(S_{z2})\psi^{\rm s}_3(\vec r_3)\updownarrow_3(S_{z3})\ldots\psi^{\rm s}_I(\vec r_I)\updownarrow_I(S_{zI})$$

That is a fundamentally different wave function, not just minus the first Hartree product; orbitals $\psi^{\rm s}_1\updownarrow_1$ and $\psi^{\rm s}_2\updownarrow_2$ are not allowed to be equivalent.

To get a wave function that does simply change sign when electrons 1 and 2 are swapped, you can take the first Hartree product minus the second one. That solves that problem. But it is not enough: the wave function must *also* simply change sign if electrons 1 and 3 are swapped. Or if 2 and 3 are swapped, etcetera.

So you must add more Hartree products with swapped electrons to the mix. A lot more in fact. There are $I!$ ways to order I electrons, and each ordering adds one Hartree product to the mix. (The Hartree product gets a plus sign or a minus sign in the mix depending on whether the number of swaps to get there from the first one is even or odd). So for, say, a single carbon atom with $I = 6$ electrons, writing down the full Hartree-Fock wave function would mean writing down $6! = 720$ Hartree products. Roughly a thousand of them, in short. Of course, writing all that out would be insane. Fortunately, there is a more concise way to write the complete wave function; it uses a so-called "Slater determinant,"

$$\frac{1}{\sqrt{I!}}\begin{vmatrix}
\psi^{\rm s}_1(\vec r_1)\updownarrow_1(S_{z1}) & \psi^{\rm s}_2(\vec r_1)\updownarrow_2(S_{z1}) & \ldots & \psi^{\rm s}_n(\vec r_1)\updownarrow_n(S_{z1}) & \ldots & \psi^{\rm s}_I(\vec r_1)\updownarrow_I(S_{z1}) \\
\psi^{\rm s}_1(\vec r_2)\updownarrow_1(S_{z2}) & \psi^{\rm s}_2(\vec r_2)\updownarrow_2(S_{z2}) & \ldots & \psi^{\rm s}_n(\vec r_2)\updownarrow_n(S_{z2}) & \ldots & \psi^{\rm s}_I(\vec r_2)\updownarrow_I(S_{z2}) \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\psi^{\rm s}_1(\vec r_i)\updownarrow_1(S_{zi}) & \psi^{\rm s}_2(\vec r_i)\updownarrow_2(S_{zi}) & \ldots & \psi^{\rm s}_n(\vec r_i)\updownarrow_n(S_{zi}) & \ldots & \psi^{\rm s}_I(\vec r_i)\updownarrow_I(S_{zi}) \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\psi^{\rm s}_1(\vec r_I)\updownarrow_1(S_{zI}) & \psi^{\rm s}_2(\vec r_I)\updownarrow_2(S_{zI}) & \ldots & \psi^{\rm s}_n(\vec r_I)\updownarrow_n(S_{zI}) & \ldots & \psi^{\rm s}_I(\vec r_I)\updownarrow_I(S_{zI})
\end{vmatrix} \tag{9.15}$$

The determinant multiplies out to the $I!$ individual Hartree products. (See chapter 5.7 and the notations section for more on determinants.) The factor $1//\sqrt{I!}$ is there to ensure that the wave function remains normalized after summing the $I!$ Hartree products together.

The most general system wave function Ψ using only $N = I$ orbitals is any coefficient a of magnitude 1 times the above Slater determinant. However, displaying the Slater determinant fully as above is still a lot to write and read. Therefore, from now on the Slater determinant will be abbreviated as in

$$\Psi = a|\det \psi^{\rm s}_1\updownarrow_1, \psi^{\rm s}_2\updownarrow_2, \ldots, \psi^{\rm s}_n\updownarrow_n, \ldots, \psi^{\rm s}_I\updownarrow_I\rangle \tag{9.16}$$

where $|\det \ldots\rangle$ is the Slater determinant.

It is important to realize that using the minimum number of single-electron functions will unavoidably produce an error that is mathematically speaking not small {N.16}. To get a vanishingly small error, you would need a large number of different Slater determinants, not just one. Still, the results you get with the basic Hartree-Fock approach may be good enough to satisfy your needs. Or you may be able to improve upon them enough with "post-Hartree-Fock methods."

But none of that would be likely if you just selected the single-electron functions $\psi^{\rm s}_1\uparrow_1$, $\psi^{\rm s}_2\uparrow_2$, ... at random. The cleverness in the Hartree-Fock approach will be in writing down equations for these single-electron wave functions that produce the *best* approximation possible with a single Slater determinant.

Recall the approximate solutions that were written down for the electrons in atoms in chapter 5.9. These solutions were really single Slater determinants. To improve on these results, you might think of trying to find more accurate ways to average out the effects of the neighboring electrons than just putting them in the nucleus as that chapter essentially did. You could smear them out over some optimal area, say. But even if you did that, the Hartree-Fock solution will still be better, because it gives the best possible approximation obtainable with *any* single determinant.

That assumes of course that the spins are taken the same way. Consider that problem for a second. Typically, a nonrelativistic approach is used, in which spin effects on the energy are ignored. Then spin only affects the antisymmetrization requirements.

Things are straightforward if you try to solve, say, a helium atom. The correct ground state takes the form

$$\Psi_{\rm He}(\vec{r}_1,\vec{r}_2) \times \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}},$$

The factor $\Psi_{\rm He}(\vec{r}_1,\vec{r}_2)$ is the spatial wave function that has the absolutely lowest energy, regardless of any antisymmetrization concerns. This wave function must be symmetric (unchanged) under electron exchange since the two electrons are identical and the ground state is unique. The antisymmetrization requirement is met because the spins combine together as shown in the second factor above; this factor changes sign when the electrons are exchanged. So the spatial state does not have to change sign.

The combined spin state shown in the second factor above is called the "singlet state," chapter 5.5.6. In the singlet state the two spins cancel each other *completely*: the net electron spin is zero. If you measure the net spin component in any direction, not just the chosen z-direction, you get zero.

Based on the exact helium ground state wave function above, you would take the Hartree-Fock approximation to be of the form

$$|\det \psi^{\rm s}_1\uparrow, \psi^{\rm s}_2\downarrow\rangle$$

and then you would make things easier for yourself by postulating a priori that the spatial orbitals are the same, $\psi_1^{\rm s} = \psi_2^{\rm s}$. Lo and behold, when you multiply out the Slater determinant,

$$\frac{1}{\sqrt{2}} \begin{vmatrix} \psi_1^{\rm s}(\vec{r}_1)\uparrow(S_{z1}) & \psi_1^{\rm s}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_1^{\rm s}(\vec{r}_2)\uparrow(S_{z2}) & \psi_1^{\rm s}(\vec{r}_2)\downarrow(S_{z2}) \end{vmatrix}$$

you get

$$\psi_1^{\rm s}(\vec{r}_1)\psi_1^{\rm s}(\vec{r}_2) \times \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

This automagically reproduces the correct singlet spin state! (The approximation comes in because the exact *spatial* ground state, $\Psi_{\rm He}(\vec{r}_1, \vec{r}_2)$ is not just the product of two single-electron functions as in Hartree-Fock.) And you only need to find one spatial orbital instead of two.

As discussed in chapter 5.9, a beryllium atom has two electrons with opposite spins in the "1s" shell like helium, and two more in the "2s" shell. An appropriate Hartree-Fock wave function would be

$$|\det \psi_1^{\rm s}\uparrow, \psi_1^{\rm s}\downarrow, \psi_3^{\rm s}\uparrow, \psi_3^{\rm s}\downarrow\rangle$$

in other words, two pairs of orbitals with the same spatial states and opposite spins. Similarly, Neon has an additional 6 paired electrons in a closed "2p" shell, and you could use 3 more pairs of orbitals with the same spatial states and opposite spins. The number of spatial orbitals that must be found in such solutions is only half the number of electrons. This procedure is called the "closed shell Restricted Hartree-Fock (RHF)" method. It restricts the form of the spatial states to be pair-wise equal.

But now look at lithium. Lithium has two paired 1s electrons like helium, and an unpaired 2s electron. For the third orbital in the Hartree-Fock determinant, you will now have to make a choice: whether to take it of the form $\psi_3^{\rm s}\uparrow$ or $\psi_3^{\rm s}\downarrow$. Lets assume you take $\psi_3^{\rm s}\uparrow$, so the wave function is

$$|\det \psi_1^{\rm s}\uparrow, \psi_2^{\rm s}\downarrow, \psi_3^{\rm s}\uparrow\rangle$$

You have introduced a bias in the determinant: there is now a real difference between the spatial orbitals $\psi_1^{\rm s}$ and $\psi_2^{\rm s}$: $\psi_1^{\rm s}\uparrow$ has the same spin as the third spin orbital, but $\psi_2^{\rm s}\downarrow$ the opposite.

If you find the best approximation to the energy among *all* possible spatial orbitals $\psi_1^{\rm s}$, $\psi_2^{\rm s}$, and $\psi_3^{\rm s}$, you will end up with orbitals $\psi_1^{\rm s}$ and $\psi_2^{\rm s}$ that are not the same. Allowing for them to be different is called the "Unrestricted Hartree-Fock (UHF)" method. In general, you no longer require that equivalent spatial orbitals are the same in their spin-up and spin down versions. For a bigger system, you will end up with one set of orthonormal spatial orbitals for the spin-up orbitals and a different set of orthonormal spatial orbitals for the spin-down ones. These two sets of orthonormal spatial orbitals are *not* mutually

orthogonal; the only reason the complete *spin* orbitals are still orthonormal is because the two spins are orthogonal, $\langle\uparrow|\downarrow\rangle = 0$.

If instead of using unrestricted Hartree-Fock, you insist on demanding that the spatial orbitals for spin up and down do form a single set of orthonormal functions, it is called "open shell Restricted Hartree-Fock (RHF)." In the case of lithium, you would then demand that $\psi^{\rm s}_2$ equals $\psi^{\rm s}_1$. Since the best (in terms of energy) solution has them different, your solution is then no longer the best possible. You pay a price, but you now only need to find two spatial orbitals rather than three. The spin orbital $\psi^{\rm s}_3\uparrow$ without a matching opposite-spin orbital counts as an open shell. For nitrogen, you might want to use three open shells to represent the three different spatial states $2{\rm p}_x$, $2{\rm p}_y$, and $2{\rm p}_z$ with an unpaired electron in it.

If you use unrestricted Hartree-Fock instead, you will need to compute more spatial functions, and you pay another price, spin. Since all spin effects in the Hamiltonian are ignored, it commutes with the spin operators. So, the exact energy eigenfunctions are also, or can be taken to be also, spin eigenfunctions. Restricted Hartree-Fock has the capability of producing approximate energy eigenstates with well defined spin. Indeed, as you saw for helium, in restricted Hartree-Fock all the paired spin-up and spin-down states combine into zero-spin singlet states. If any additional unpaired states are all spin up, say, you get an energy eigenstate with a net spin equal to the sum of the spins of the unpaired states. This allows you to deal with typical atoms, including lithium and nitrogen, very nicely.

But a true unrestricted Hartree-Fock solution does not have correct, definite, spin. For two electrons to produce states of definite combined spin, the coefficients of spin-up and spin-down must come in specific ratios. As a simple example, an unrestricted Slater determinant of $\psi^{\rm s}_1\uparrow$ and $\psi^{\rm s}_2\downarrow$ with unequal spatial orbitals multiplies out to

$$|\det \psi^{\rm s}_1\uparrow, \psi^{\rm s}_2\downarrow\rangle = \frac{\psi^{\rm s}_1(\vec{r}_1)\psi^{\rm s}_2(\vec{r}_2)\uparrow(S_{z1})\downarrow(S_{z2}) - \psi^{\rm s}_2(\vec{r}_1)\psi^{\rm s}_1(\vec{r}_2)\downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

or, writing the spin combinations in terms of singlets (which change sign under electron exchange) and triplets (which do not),

$$\frac{\psi^{\rm s}_1(\vec{r}_1)\psi^{\rm s}_2(\vec{r}_2) + \psi^{\rm s}_2(\vec{r}_1)\psi^{\rm s}_1(\vec{r}_2)}{2} \times \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}} +$$

$$\frac{\psi^{\rm s}_1(\vec{r}_1)\psi^{\rm s}_2(\vec{r}_2) - \psi^{\rm s}_2(\vec{r}_1)\psi^{\rm s}_1(\vec{r}_2)}{2} \times \frac{\uparrow(S_{z1})\downarrow(S_{z2}) + \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

So the spin will be some combination of the singlet state, the first term, and a triplet state, the second. And the precise combination will depend on the spatial locations of the electrons to boot. Now while the singlet state has net spin 0, triplet states have net spin 1. So the net spin is uncertain, either 0 or

1, even though it should not be. (Spin 1 implies that the measured component of the spin in any direction must be one of the triplet of values $\hbar$, 0, or $-\hbar$. For the particular triplet state shown above, the component of spin in the z-direction happens to be zero. But the net spin is not; in a direction normal to the z-direction, the spin will be measured to be either $\hbar$ or $-\hbar$.) However, despite the spin problem, it may be noted that unrestricted wave functions are commonly used as first approximations of doublet (spin $\frac{1}{2}$) and triplet (spin 1) states anyway [46, p. 105].

To show that all this can make a real difference, take the example of the hydrogen molecule, chapter 5.2, when the two nuclei are far apart. The correct electronic ground state is

$$\frac{\psi_{\rm L}(\vec{r}_1)\psi_{\rm R}(\vec{r}_2) + \psi_{\rm R}(\vec{r}_1)\psi_{\rm L}(\vec{r}_2)}{\sqrt{2}} \times \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

where $\psi_{\rm L}(\vec{r}_1)\psi_{\rm R}(\vec{r}_2)$ is the state in which electron 1 is around the left proton and electron 2 around the right one, and $\psi_{\rm R}(\vec{r}_1)\psi_{\rm L}(\vec{r}_2)$ is the same state but with the electrons reversed. Note that, like for the helium atom, the spatial state is symmetric under electron exchange. However, it is not just a product of two single-electron functions but a sum of two of such products. Note also that the correct spin state is the singlet one with zero net spin, just like for the helium atom. It takes care of the antisymmetrization requirement.

Now try to approximate this solution with a restricted closed shell Hartree-Fock wave function of the form

$$|\det \psi_1^{\rm s}\uparrow, \psi_1^{\rm s}\downarrow\rangle$$

Multiplying out the determinant gives

$$\psi_1^{\rm s}(\vec{r}_1)\psi_1^{\rm s}(\vec{r}_2) \times \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

Note that you do get the correct singlet spin state. But $\psi_1^{\rm s}$ will be something like $(\psi_{\rm L} + \psi_{\rm R})/\sqrt{2}$; the energy of either electron is lowest when it is near one of the nuclei. If you multiply out the resulting spatial wave function, the terms include $\psi_{\rm L}\psi_{\rm L}$ and $\psi_{\rm R}\psi_{\rm R}$, in addition to the correct $\psi_{\rm L}\psi_{\rm R}$ and $\psi_{\rm R}\psi_{\rm L}$. That produces a 50/50 chance that the two electrons are found around the *same* nucleus. That is all wrong, since the electrons repel each other: if one electron is around the left nucleus, the other electron should be around the right nucleus. The computed energy, which should be that of two neutral hydrogen atoms far apart, will be much too high due to electron-electron repulsion.

(Fortunately, at the nuclear separation distance corresponding to the ground state of the complete molecule, the errors are much less, [46, p. 166]. Note that if you put the two nuclei completely on top of each other, you get a helium atom, for which Hartree-Fock gives a much more reasonable electron energy.

Only when you are "breaking the bond," dissociating the molecule, i.e. taking the nuclei far apart, do you get into major trouble.)

If instead you would use unrestricted Hartree-Fock, say

$$|\det \psi_1^{\rm s}\uparrow, \psi_2^{\rm s}\downarrow\rangle$$

you should find $\psi_1^{\rm s} = \psi_{\rm L}$ and $\psi_2^{\rm s} = \psi_{\rm R}$ (or vice versa), which would produce a wave function

$$\frac{\psi_{\rm L}(\vec{r}_1)\psi_{\rm R}(\vec{r}_2)\uparrow(S_{z1})\downarrow(S_{z2}) - \psi_{\rm R}(\vec{r}_1)\psi_{\rm L}(\vec{r}_2)\downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}.$$

In both terms, if the first electron is around the one nucleus, the second electron is around the other. So this produces the correct energy, that of two neutral hydrogen atoms. But the spin is now all wrong. It is not a singlet state, but the combination of a singlet and a triplet state already written down earlier. Little in life is ideal, is it?

(Actually there is a dirty trick to fix this. Note that which of the two orbitals you give spin-up and which spin-down is physically immaterial. So there is a trivially different solution

$$|\det \psi_1^{\rm s}\downarrow, \psi_2^{\rm s}\uparrow\rangle$$

If you take a 50/50 combination of the original Slater determinant and minus the one above, you get the correct singlet spin state. And the spatial state will now be the correct average of $\psi_{\rm L}\psi_{\rm R}$ and $\psi_{\rm R}\psi_{\rm L}$ to boot. This spatial state is more accurate than just two neutral atoms if the distance between the nuclei decreases, chapter 5.2. All this for free! This sort of dirty trick in Hartree-Fock is called a "spin adapted configuration." It is usually used to deal with a few open shells in an otherwise closed-shell restricted Hartree-Fock configuration.)

All of the above may be much more than you ever wanted to hear about the wave function. The purpose was mainly to indicate that things are not as simple as you might initially suppose. As the examples showed, some understanding of the system that you are trying to model definitely helps. Or experiment with different approaches.

Let's go on to the next step: how to get the equations for the spatial orbitals $\psi_1^{\rm s}, \psi_2^{\rm s}, \ldots$ that give the most accurate approximation of a multi-electron problem. The expectation value of energy will be needed for that, and to get that, first the Hamiltonian is needed. That will be the subject of the next subsection.

9.3.2 The Hamiltonian

The nonrelativistic Hamiltonian of the system of I electrons consists of a number of contributions. First there is the kinetic energy of the electrons; the sum of the kinetic energy operators of the individual electrons:

$$\widehat{T}^{\rm E} = -\sum_{i=1}^{I} \frac{\hbar^2}{2m_{\rm e}} \nabla_i^2 = -\sum_{i=1}^{I} \frac{\hbar^2}{2m_{\rm e}} \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \right). \tag{9.17}$$

Next there is the potential energy due to the ambient electric field that the electrons move in. It will be assumed that this field is caused by J nuclei, numbered using an index j, and having charge $Z_j e$ (i.e. there are Z_j protons in nucleus number j). In that case, the total potential energy due to nucleus-electron attractions is, summing over all electrons and over all nuclei:

$$V^{\rm NE} = -\sum_{i=1}^{I} \left(\sum_{j=1}^{J} \frac{Z_j e^2}{4\pi\epsilon_0} \frac{1}{r^{\rm n}_{ij}} \right) \tag{9.18}$$

where $r^{\rm n}_{ij} \equiv |\vec{r}_i - \vec{r}^{\,\rm n}_j|$ is the distance between electron number i and nucleus number j, and $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is the permittivity of space.

And now for the black plague of quantum mechanics, the electron to electron repulsions. The potential energy for those repulsions is

$$V^{\rm EE} = \sum_{i=1}^{I} \sum_{\underline{i}>i}^{I} \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{i\underline{i}}} \tag{9.19}$$

where $r_{i\underline{i}} \equiv |\vec{r}_i - \vec{r}_{\underline{i}}|$ is the distance between electron number i and electron number $\underline{i}$. To avoid counting each repulsion energy twice, (the second time with reversed electron order), the second electron number is required to be larger than the first.

Without this repulsion between different electrons, you could solve for each electron separately, and all would be nice. But you do have it, and so you really need to solve for all electrons at once, usually an impossible task. You may recall that when chapter 5.9 examined the atoms heavier than hydrogen, those with more than one electron, the discussion cleverly threw out the electron to electron repulsion terms, by assuming that the effect of each neighboring electron is approximately like canceling out one proton in the nucleus. And you may also remember how this outrageous assumption led to all those wrong predictions that had to be corrected by various excuses. The Hartree-Fock approximation tries to do better than that.

It is helpful to split the Hamiltonian into the single electron terms and the troublesome interactions, as follows,

$$H = \sum_{i=1}^{I} h^{\rm e}_i + \sum_{i=1}^{I} \sum_{\underline{i}>i}^{I} v^{\rm ee}_{i\underline{i}} \tag{9.20}$$

where $h^{\rm e}_i$ is the single-electron Hamiltonian of electron i,

$$h^{\rm e}_i = -\frac{\hbar^2}{2m_{\rm e}} \nabla_i^2 + \sum_{j=1}^{J} \frac{Z_j e^2}{4\pi\epsilon_0} \frac{1}{r^{\rm n}_{ij}} \tag{9.21}$$

and $v^{\rm ee}_{i\underline{i}}$ is the electron i to electron $\underline{i}$ repulsion potential energy,

$$v^{\rm ee}_{i\underline{i}} = \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_{i\underline{i}}} \tag{9.22}$$

Note that $h^{\rm e}_1$, $h^{\rm e}_2$, ..., $h^{\rm e}_I$ all take the same general form; the difference is just in which electron you are talking about. That is not surprising because the electrons all have the same properties. Similarly, the difference between $v^{\rm ee}_{12}$, $v^{\rm ee}_{13}$, ..., $v^{\rm ee}_{(I-1)I}$ is just in which pair of electrons you talk about.

9.3.3 The expectation value of energy

As was discussed in more detail in section 9.1, to find the best possible Hartree-Fock approximation, the expectation value of energy will be needed. For example, the best approximation to the ground state is the one that has the smallest expectation value of energy.

The expectation value of energy $\langle E \rangle$ is defined as the inner product

$$\langle \Psi | H | \Psi \rangle$$

where H is the Hamiltonian as given in the previous subsection. There is a problem with using this expression mindlessly, though. Take once again the example of the arsenic atom. There are 33 electrons in this atom, so you could try to choose 33 promising single-electron wave functions to describe it. You could then try to multiply out the Slater determinant for Ψ, but that produces 33!, or about 4 10^{36}, Hartree products. If you put these 33! terms in both sides of the inner product, you get $(33!)^2$ or 7.5 10^{73} pairs of terms, each producing one inner product that must be integrated. Now since there are 3 coordinates for each of the positions of the 33 electrons, this means that each term requires integration over 99 scalar coordinates. Even using only 10 points in each direction, that would mean evaluating 10^{99} integration points for each of the 7.5 10^{73} pairs of terms. A computer that could do that is unimaginable. As of 2014, the fastest computer in the world can do no more than 10^{25} floating point computations if it stays at it for 10 years.

Fortunately, it turns out, {D.52}, that almost all of those integrations are trivial since the single-electron functions are orthonormal. If you sit down and identify what is really left, you find that only a few three-dimensional and six-dimensional inner products survive the weeding-out process.

In particular, the single-electron Hamiltonians from the previous subsection produce only single-electron energy expectation values of the general form

$$\boxed{E^{\rm e}_n \equiv \langle \psi^{\rm s}_n | h^{\rm e} | \psi^{\rm s}_n \rangle} \tag{9.23}$$

If you had only one single electron, and it was in the spatial single-particle state $\psi^{\rm s}_n(\vec{r})$, the above inner product would be its energy.

The combined single-electron energy for all I electrons is then

$$\sum_{n=1}^{I} E_n^{\rm e}$$

It is just as if you had electron 1 in state $\psi_1^{\rm s}$, electron 2 in state $\psi_2^{\rm s}$, etcetera. Of course, that is not really true. Antisymmetrization requires that all electrons are partly in all states. Indeed, if you look a bit closer at the math, you see that each of the I electrons contributes an equal fraction $1/I$ to each of the I terms above. But it does not make a real difference. Without electron-electron interactions, quantum mechanics would be so much easier!

But the repulsions are there. The Hamiltonians of the repulsions turn out to produce six-dimensional spatial inner products of two types. The inner products of the first type are called "Coulomb integrals:"

$$\boxed{J_{n\underline{n}} \equiv \left\langle \psi_n^{\rm s} \psi_{\underline{n}}^{\rm s} \middle| v^{\rm ee} \middle| \psi_n^{\rm s} \psi_{\underline{n}}^{\rm s} \right\rangle} \tag{9.24}$$

To understand the Coulomb integrals better, the inner product above can be written out explicitly as an integral, while also expanding $v^{\rm ee}$:

$$\int_{\text{all } \vec{r}} \int_{\text{all } \underline{\vec{r}}} |\psi_n^{\rm s}(\vec{r})|^2 |\psi_{\underline{n}}^{\rm s}(\underline{\vec{r}})|^2 \frac{e^2}{4\pi\epsilon_0} \frac{1}{|\vec{r} - \underline{\vec{r}}|} \, {\rm d}^3\vec{r} \, {\rm d}^3\underline{\vec{r}}$$

The integrand equals the probability of an electron in state $\psi_n^{\rm s}$ to be found near a position $\vec{r}$, times the probability of an electron in state $\psi_{\underline{n}}^{\rm s}$ to be found near a position $\underline{\vec{r}}$, times the Coulomb repulsion energy if the two electrons are at those positions. In short, $J_{n\underline{n}}$ is the expectation value of the Coulomb repulsion potential between an electron in state $\psi_n^{\rm s}$ and one in state $\psi_{\underline{n}}^{\rm s}$. Thinking again of electron 1 in state $\psi_1^{\rm s}$, electron 2 in state $\psi_2^{\rm s}$, etcetera, the total Coulomb repulsion energy would be

$$\sum_{n=1}^{I} \sum_{\underline{n}>n}^{I} J_{n\underline{n}}$$

which is indeed the correct combined sum of the Coulomb integrals.

Unfortunately, that is not the complete story for the repulsion energy. Recall that there are $I!$ different ways in which you can distribute the I electrons over the I single particle states. And the antisymmetrization requirement requires that the system wave function is an equal combination of all these $I!$ different possibilities. In terms of classical physics, it might still seem that this should make no difference: if any one of these $I!$ possibilities is true, then the others must be untrue. But quantum mechanics allows states in which the electrons are distributed in one way to interact with states in which they are distributed in another way. That produces the so-called "exchange integrals:"

$$\boxed{K_{n\underline{n}} \equiv \left\langle \psi_n^{\rm s} \psi_{\underline{n}}^{\rm s} \middle| v^{\rm ee} \middle| \psi_{\underline{n}}^{\rm s} \psi_n^{\rm s} \right\rangle} \tag{9.25}$$

Written out explicitly, that equals

$$\int_{\text{all } \vec{r}} \int_{\text{all } \underline{\vec{r}}} \psi_n^{\rm s}(\vec{r})^* \psi_{\underline{n}}^{\rm s}(\underline{\vec{r}})^* \frac{e^2}{4\pi\epsilon_0} \frac{1}{|\vec{r} - \underline{\vec{r}}|} \psi_{\underline{n}}^{\rm s}(\vec{r}) \psi_n^{\rm s}(\underline{\vec{r}}) \, {\rm d}^3\vec{r} \, {\rm d}^3\underline{\vec{r}}$$

It is an interaction of the possibility that the first electron is in state $\psi_n^{\rm s}$ and the second in state $\psi_{\underline{n}}^{\rm s}$ with the possibility that the second electron is in state $\psi_n^{\rm s}$ and the first in state $\psi_{\underline{n}}^{\rm s}$. This book likes to call terms like this "twilight terms," since in terms of classical physics they do not make sense.

It may be noted that a single Hartree product satisfying the Pauli exclusion principle would not produce exchange integrals; in such a wave function, there is no possibility for an electron to be in another state. But don't start thinking that the exchange integrals are there just because the wave function must be antisymmetric under electron exchange. They, and others, would show up in any reasonably general wave function. You can think of the exchange integrals instead as Coulomb integrals with the electrons in the right hand side of the inner product exchanged.

Adding it all up, the expectation energy of the complete system of I electrons can be written as

$$\boxed{\langle E \rangle = \sum_{n=1}^{I} E_n^{\rm e} + \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\underline{n}=1}^{I} J_{n\underline{n}} - \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\underline{n}=1}^{I} \left\langle \updownarrow_n \middle| \updownarrow_{\underline{n}} \right\rangle^2 K_{n\underline{n}}} \tag{9.26}$$

Note that the above expression sums over all values of $\underline{n}$, not just $\underline{n} > n$. That counts each pair of single-electron wave functions twice, so factors one-half have been added to compensate. It also adds terms in which $\underline{n} = n$, both electrons in the same state, which is not allowed by the Pauli principle. But since $J_{nn} = K_{nn}$, these additional terms cancel each other.

Note also the spin inner products multiplying the exchange terms. These are zero if the two states have opposite spin, so there are no exchange contributions between electrons in spin orbitals of opposite spins. And if the spin orbitals have the same spin, the spin inner product is 1, so the square is somewhat superfluous.

There are also some a priori things you can say about the Coulomb and exchange integrals, {D.53}; they are real, and additionally

$$J_{nn} = K_{nn} \qquad J_{n\underline{n}} = J_{\underline{n}n} \qquad K_{n\underline{n}} = K_{\underline{n}n} \qquad J_{n\underline{n}} \geqslant K_{n\underline{n}} \geqslant 0 \tag{9.27}$$

Note in particular that since the $K_{n\underline{n}}$ terms are positive, they lower the net expectation energy of the system. So a wave function consisting of a single Hartree product, which produces no exchange terms, cannot be the state of lowest energy. Even without the antisymmetrization requirement, you would need Hartree products with the electrons exchanged, simply to lower the energy.

It is actually somewhat tricky to prove that the K_{nn} terms are positive and so lower the energy, {D.53}. But there is a simple physical reason why you might *guess* that an antisymmetric wave function would lower the electron-electron repulsion energy compared to the individual Hartree products from which it is made up. In particular, the Coulomb repulsion between electrons becomes very large when they get close together. But for an anti-symmetric wave function, unlike for a single Hartree product, the relative probability of electrons of the same spin getting close together is vanishingly small. That prevents any strong Coulomb repulsion between electrons of the same spin.

(Recall that the relative probability for electrons to be at given positions and spins is given by the square magnitude of the wave function at those positions and spins. Now an antisymmetric wave function must be zero wherever any two electrons are at the same position with the same spin, making this impossible. After all, if you swap the two electrons, the antisymmetric wave function must change sign. But since neither electron changes position nor spin, the wave function cannot change either. Something can only change sign and stay the same if it is zero. See also {A.34}.)

The analysis given in this subsection can easily be extended to generalized orbitals that take the form

$$\psi^{\rm p}_n(\vec{r}, S_z) = \psi^{\rm s}_{n+}(\vec{r})\uparrow(S_z) + \psi^{\rm s}_{n-}(\vec{r})\downarrow(S_z).$$

However, the normal unrestricted spin-up or spin-down orbitals, in which either $\psi^{\rm s}_{n+}$ or $\psi^{\rm s}_{n-}$ is zero, already satisfy the variational requirement $\delta\langle E\rangle = 0$ even if generalized variations in the orbitals are allowed, {N.17}.

In any case, the expectation value of energy has been found.

9.3.4 The canonical Hartree-Fock equations

The previous subsection found the expectation value of energy for any electron wave function described by a single Slater determinant. The final step is to find the orbitals that produce the best approximation of the true wave function using such a single determinant. For the ground state, the best single determinant would be the one with the lowest expectation value of energy. But surely you would not want to guess spatial orbitals at random until you find some with really, really, low energy.

What you would like to have is specific equations for the best spatial orbitals that you can then solve in a methodical way. And you can have them using the methods of section 9.1, {D.54}. In unrestricted Hartree-Fock, for every spatial

orbital $\psi^{\rm s}_n(\vec{r})$ there is an equation of the form:

$$h^{\rm e}\psi^{\rm s}_n(\vec{r}) + \sum_{\underline{n}=1}^{I} \left\langle \psi^{\rm s}_{\underline{n}} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}} \right\rangle \psi^{\rm s}_n(\vec{r}) - \sum_{\underline{n}=1}^{I} \left\langle \updownarrow_{\underline{n}} \middle| \updownarrow_n \right\rangle^2 \left\langle \psi^{\rm s}_{\underline{n}} \middle| v^{\rm ee} \middle| \psi^{\rm s}_n \right\rangle \psi^{\rm s}_{\underline{n}}(\vec{r}) = \epsilon_n \psi^{\rm s}_n(\vec{r}) \tag{9.28}$$

These are called the "canonical Hartree-Fock equations." For equations valid for the restricted closed-shell and single-determinant open-shell approximations, see the derivation in {D.54}.

Recall that $h^{\rm e}$ is the single-electron Hamiltonian consisting of the electron's kinetic energy and potential energy due to nuclear attractions, and that $v^{\rm ee}$ is the potential energy of repulsion between the electron and another at a position $\underline{\vec{r}}$:

$$h^{\rm e} = -\frac{\hbar^2}{2m_{\rm e}}\nabla^2 - \sum_{j=1}^{J} \frac{Z_j e^2}{4\pi\epsilon_0} \frac{1}{r_j} \quad r_j \equiv |\vec{r} - \vec{r}^{\,\rm n}_j| \qquad v^{\rm ee} = \frac{e^2}{4\pi\epsilon_0} \frac{1}{\underline{r}} \quad \underline{r} \equiv |\vec{r} - \underline{\vec{r}}|$$

So, if there were no electron-electron repulsions, i.e. $v^{\rm ee} = 0$, the canonical equations above would be single-electron Hamiltonian eigenvalue problems of the form $h^{\rm e}\psi^{\rm s}_n = \epsilon_n \psi^{\rm s}_n$ where ϵ_n would be the energy of the single-electron orbital. This is really what happened in the approximate analysis of atoms in chapter 5.9: the electron to electron repulsions were ignored there in favor of nuclear strength reductions, and the result was single-electron hydrogen-atom orbitals.

In the presence of electron to electron repulsions, the equations for the orbitals can still *symbolically* be written as if they were single-electron eigenvalue problems,

$$\mathcal{F}\psi^{\rm s}_n(\vec{r})\updownarrow_n(S_z) = \epsilon_n \psi^{\rm s}_n(\vec{r})\updownarrow_n(S_z)$$

where $\mathcal{F}$ is called the "Fock operator," and is written out further as:

$$\mathcal{F} = h^{\rm e} + v^{\rm HF}.$$

The first term in the Fock operator is the single-electron Hamiltonian. The mischief is in the innocuous-looking second term $v^{\rm HF}$. Supposedly, this is the potential energy related to the repulsion by the other electrons. What is it? Well, it will have to be the terms in the canonical equations (9.28) not described by the single-electron Hamiltonian $h^{\rm e}$:

$$v^{\rm HF}\psi^{\rm s}(\vec{r})\updownarrow(S_z) = \sum_{\underline{n}=1}^{I} \left\langle \psi^{\rm s}_{\underline{n}} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}} \right\rangle \psi^{\rm s}(\vec{r})\updownarrow(S_z) - \sum_{\underline{n}=1}^{I} \left\langle \updownarrow_{\underline{n}} \middle| \updownarrow \right\rangle \left\langle \psi^{\rm s}_{\underline{n}} \middle| v^{\rm ee} \middle| \psi^{\rm s}(\underline{\vec{r}}) \right\rangle \psi^{\rm s}_{\underline{n}}(\vec{r})\updownarrow_{\underline{n}}(S_z)$$

The definition of the Fock operator is unavoidably in terms of spin rather than just spatial orbitals: the spin of the state on which it operates must be known to evaluate the final term.

Note that the above expression did not give an expression for $v^{\rm HF}$ by itself, but only for $v^{\rm HF}$ applied to an arbitrary single-electron function $\psi^{\rm s}\updownarrow$. The reason is that $v^{\rm HF}$ is not a normal potential at all: the second term, the one due to the exchange integrals, does not multiply $\psi^{\rm s}\updownarrow$ by a potential function, it shoves it into an inner product! The Hartree-Fock "potential" $v^{\rm HF}$ is an *operator*, not a normal potential energy. Given a single-electron function including spin, it produces another single-electron function including spin.

Actually, even that is not quite true. The Hartree-Fock "potential" is only an operator *after* you have found the orbitals $\psi^{\rm s}_1\updownarrow_1$, $\psi^{\rm s}_2\updownarrow_2$, ..., $\psi^{\rm s}_{\underline{n}}\updownarrow_{\underline{n}}$, ..., $\psi^{\rm s}_I\updownarrow_I$ appearing in it. While you are still trying to find them, the Fock "operator" is not even an operator, it is just a "thing." However, *given* the orbitals, at least the Fock operator is a Hermitian one, one that can be taken to the other side if it appears in an inner product, and that has real eigenvalues and a complete set of eigenfunctions, {D.55}.

So how do you solve the canonical Hartree-Fock equations for the orbitals $\psi^{\rm s}_n$? If the Hartree-Fock potential $v^{\rm HF}$ was a known operator, you would have only linear, single-electron eigenvalue problems to solve. That would be relatively easy, as far as those things come. But since the operator $v^{\rm HF}$ contains the unknown orbitals, you do not have a linear problem at all; it is a system of coupled cubic equations in infinitely many unknowns. The usual way to solve it is iteratively: you guess an approximate form of the orbitals and plug it into the Hartree-Fock potential. With this guessed potential, the orbitals may then be found from solving linear eigenvalue problems. If all goes well, the obtained orbitals, though not perfect, will at least be better than the ones that you guessed at random. So plug those improved orbitals into the Hartree-Fock potential and solve the eigenvalue problems again. Still better orbitals should result. Keep going until you get the correct solution to within acceptable accuracy.

You will know when you have got the correct solution since the Hartree-Fock potential will no longer change; the potential that you used to compute the final set of orbitals is really the potential that those final orbitals produce. In other words, the final Hartree-Fock potential that you compute is consistent with the final orbitals. Since the potential would be a field if it was not an operator, that explains why such an iterative method to compute the Hartree-Fock solution is called a "self-consistent field method." It is like calling an iterative scheme for the Laplace equation on a mesh a "self-consistent neighbors method," instead of "point relaxation." Surely the equivalent for Hartree-Fock, like "iterated potential" or "potential relaxation" would have been much clearer to a general audience?

9.3.5 Additional points

This brief section was not by any means a tutorial of the Hartree-Fock method. The purpose was only to explain the basic ideas in terms of the notations and coverage of this book. If you actually want to apply the method, you will need to take up a book written by experts who know what they are talking about. The book by Szabo and Ostlund [46] was the main reference for this section, and is recommended as a well written introduction. Below are some additional concepts you may want to be aware of.

9.3.5.1 Meaning of the orbital energies

In the single electron case, the "orbital energy" ϵ_n in the canonical Hartree-Fock equation

$$h^{\rm e}\psi^{\rm s}_n(\vec{r}) + \sum_{\underline{n}=1}^{I} \left\langle \psi^{\rm s}_{\underline{n}} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}} \right\rangle \psi^{\rm s}_n(\vec{r}) - \sum_{\underline{n}=1}^{I} \left\langle \updownarrow_{\underline{n}} \middle| \updownarrow_n \right\rangle^2 \left\langle \psi^{\rm s}_{\underline{n}} \middle| v^{\rm ee} \middle| \psi^{\rm s}_n \right\rangle \psi^{\rm s}_{\underline{n}}(\vec{r}) = \epsilon_n \psi^{\rm s}_n(\vec{r})$$

represents the actual energy of the electron. It also represents the ionization energy, the energy required to take the electron away from the nuclei and leave it far away at rest. This subsubsection will show that in the multiple electron case, the "orbital energies" ϵ_n are not orbital energies in the sense of giving the contributions of the orbitals to the total expectation energy. However, they can still be taken to be approximate ionization energies. This result is known as "Koopman's theorem."

To verify the theorem, a suitable equation for ϵ_n is needed. It can be found by taking an inner product of the canonical equation above with $\psi^{\rm s}_n(\vec{r})$, i.e. by putting $\psi^{\rm s}_n(\vec{r})^*$ to the left of both sides and integrating over $\vec{r}$. That produces

$$\epsilon_n = E^{\rm e}_n + \sum_{\underline{n}=1}^{I} J_{n\underline{n}} - \sum_{\underline{n}=1}^{I} \left\langle \updownarrow_n \middle| \updownarrow_{\underline{n}} \right\rangle^2 K_{n\underline{n}} \tag{9.29}$$

which consists of the single-electron energy $E^{\rm e}_n$, Coulomb integrals $J_{n\underline{n}}$ and exchange integrals $K_{n\underline{n}}$ as defined in subsection 9.3.3. It can already be seen that if all the ϵ_n are summed together, it does not produce the total expectation energy (9.26), because that one includes a factor $\frac{1}{2}$ in front of the Coulomb and exchange integrals. So, ϵ_n cannot be seen as the part of the system energy associated with orbital $\psi^{\rm s}_n\updownarrow_n$ in any meaningful sense.

However, ϵ_n can still be viewed as an approximate ionization energy. Assume that the electron is removed from orbital $\psi^{\rm s}_n\updownarrow_n$, leaving the electron at infinite distance at rest. No, scratch that; all electrons share orbital $\psi^{\rm s}_n\updownarrow_n$, not just one.

Assume that one electron is removed from the system and that the remaining $I-1$ electrons stay out of the orbital $\psi^{\mathrm{s}}_n{\uparrow_n}$. Then, *if it is assumed that the other orbitals do not change,* the new system's Slater determinant is the same as the original system's, except that column n and a row have been removed. The expectation energy of the new state then equals the original expectation energy, except that E^{e}_n and the n-th column plus the n-th row of the Coulomb and exchange integral matrices have been removed. The energy removed is then exactly ϵ_n above. (While ϵ_n only involves the n-th row of the matrices, not the n-th column, it does not have the factor $\frac{1}{2}$ in front of them like the expectation energy does. And rows equal columns in the matrices, so half the row in ϵ_n counts as the half column in the expectation energy and the other half as the half row. This counts the element $\underline{n} = n$ twice, but that is zero anyway since $J_{nn} = K_{nn}$.)

So by the removal of the electron "from" (read: and) orbital $\psi^{\mathrm{s}}_n{\uparrow_n}$, an amount of energy ϵ_n has been removed from the expectation energy. Better put, a positive amount of energy $-\epsilon_n$ has been added to the expectation energy. So the ionization energy is $-\epsilon_n$ if the electron is removed from orbital $\psi^{\mathrm{s}}_n{\uparrow_n}$ according to this story.

Of course, the assumption that the other orbitals do not change after the removal of one electron and orbital is dubious. If you were a lithium electron in the expansive 2s state, and someone removed one of the two inner 1s electrons, would you not want to snuggle up a lot more closely to the now much less shielded three-proton nucleus? On the other hand, in the more likely case that someone removed the 2s electron, it would probably not seem like that much of an event to the remaining two 1s electrons near the nucleus, and the assumption that the orbitals do not change would appear more reasonable. And normally, when you say ionization energy, you are talking about removing the electron from the highest energy state.

But still, you should really recompute the remaining two orbitals from the canonical Hartree-Fock equations for a two-electron system to get the best, lowest, energy for the new $I-1$ electron ground state. The energy you get by not doing so and just sticking with the original orbitals will be too high. Which means that all else being the same, the ionization energy will be too high too.

However, there is another error of importance here, the error in the Hartree-Fock approximation itself. If the original and final system would have the same Hartree-Fock error, then it would not make a difference and ϵ_n would overestimate the ionization energy as described above. But Szabo and Ostlund [46, p. 128] note that Hartree-Fock tends to overestimate the energy for the original larger system more than for the final smaller one. The difference in Hartree-Fock error tends to compensate for the error you make by not recomputing the final orbitals, and in general the orbital energies provide reasonable first approximations to the experimental ionization energies.

The opposite of ionization energy is "electron affinity," the energy with which

the atom or molecule will bind an additional free electron [in its valence shell], {N.19}. It is not to be confused with electronegativity, which has to do with willingness to take on electrons in chemical bonds, rather than free electrons.

To compute the electron affinity of an atom or molecule with I electrons using the Hartree-Fock method, you can either recompute the $I+1$ orbitals with the additional electron from scratch, or much easier, just use the Fock operator of the I electrons to compute one more orbital $\psi^{\rm s}_{I+1}\updownarrow_{I+1}$. In the later case however, the energy of the final system will again be higher than Hartree-Fock, and it being the larger system, the Hartree-Fock energy will be too high compared to the I-electron system already. So now the errors add up, instead of subtract as in the ionization case. If the final energy is too high, then the computed binding energy will be too low, so you would expect ϵ_{I+1} to underestimate the electron affinity relatively badly. That is especially so since affinities tend to be relatively small compared to ionization energies. Indeed Szabo and Ostlund [46, p. 128] note that while many neutral molecules will take up and bind a free electron, producing a stable negative ion, the orbital energies almost always predict negative binding energy, hence no stable ion.

9.3.5.2 Asymptotic behavior

The exchange terms in the Hartree-Fock potential are not really a potential, but an operator. It turns out that this makes a major difference in how the probability of finding an electron decays with distance from the system.

Consider again the Fock eigenvalue problem, but with the single-electron Hamiltonian identified in terms of kinetic energy and nuclear attraction,

$$-\frac{\hbar^2}{2m_{\rm e}}\nabla^2\psi^{\rm s}_n(\vec{r}) + v^{\rm Ne}\psi^{\rm s}_n(\vec{r}) + \sum_{\underline{n}=1}^{I}\left\langle\psi^{\rm s}_{\underline{n}}\middle|v^{\rm ee}\middle|\psi^{\rm s}_{\underline{n}}\right\rangle\psi^{\rm s}_n(\vec{r}) - \sum_{\underline{n}=1}^{I}\left\langle\updownarrow_{\underline{n}}\middle|\updownarrow_n\right\rangle^2\left\langle\psi^{\rm s}_{\underline{n}}\middle|v^{\rm ee}\middle|\psi^{\rm s}_n\right\rangle\psi^{\rm s}_{\underline{n}}(\vec{r}) = \epsilon_n\psi^{\rm s}_n(\vec{r})$$

Now consider the question which of these terms dominate at large distance from the system and therefore determine the large-distance behavior of the solution.

The first term that can be thrown out is $v^{\rm Ne}$, the Coulomb potential due to the nuclei; this potential decays to zero approximately inversely proportional to the distance from the system. (At large distance from the system, the distances between the nuclei can be ignored, and the potential is then approximately the one of a single point charge with the combined nuclear strengths.) Since ϵ_n in the right hand side does not decay to zero, the nuclear term cannot survive compared to it.

Similarly the third term, the Coulomb part of the Hartree-Fock potential, cannot survive since it too is a Coulomb potential, just with a charge distribution given by the orbitals in the inner product.

However, the final term in the left hand side, the exchange part of the Hartree-Fock potential, is more tricky, because the various parts of this sum have other orbitals outside of the inner product. This term can still be ignored for the slowest-decaying spin-up and spin-down states, because for them none of the other orbitals is any larger, and the multiplying inner product still decays like a Coulomb potential (faster, actually). Under these conditions the kinetic energy will have to match the right hand side, implying

$$\text{slowest decaying orbitals: } \psi_n^{\mathrm{s}}(\vec{r}) \sim \exp(-\sqrt{-2m_{\mathrm{e}}\epsilon_n}r/\hbar + \ldots)$$

From this expression, it can also be seen that the ϵ_n values must be negative, or else the slowest decaying orbitals would not have the exponential decay with distance of a bound state.

The other orbitals, however, cannot be less than the slowest decaying one of the same spin by more than algebraic factors: the slowest decaying orbital with the same spin appears in the exchange term sum and will have to be matched. So, with the exchange terms included, all orbitals normally decay slowly, raising the chances of finding electrons at significant distances. The decay can be written as

$$\psi_n^{\mathrm{s}}(\vec{r}) \sim \exp(-\sqrt{2m_{\mathrm{e}}|\epsilon_m|_{\text{min, same spin, no ss}}}r/\hbar + \ldots) \tag{9.30}$$

where ϵ_m is the ϵ value of smallest magnitude (absolute value) among all the orbitals with the same spin.

However, in the case that $\psi_n^{\mathrm{s}}(\vec{r})$ is spherically symmetric, (i.e. an s state), exclude other s-states as possibilities for ϵ_m. The reason is a peculiarity of the Coulomb potential that makes the inner product appearing in the exchange term exponentially small at large distance for two orthogonal, spherically symmetric states. (For the incurably curious, it is a result of Maxwell's first equation applied to a spherically symmetric configuration like figure 13.1, but with multiple spherically distributed charges rather than one, and the net charge being zero.)

9.3.5.3 Hartree-Fock limit

The Hartree-Fock approximation greatly simplifies finding a many-dimensional wave function. But really, solving the "eigenvalue problems" (9.28) for the orbitals iteratively is not that easy either. Typically, what one does is to write the orbitals ψ_n^{s} as sums of *chosen* single-electron functions $f_1, f_2, \ldots$. You can then precompute various integrals in terms of those functions. Of course, the number of chosen single-electron functions will have to be a lot more than the number of orbitals I; if you are only using I chosen functions, it really means that you are choosing the orbitals ψ_n^{s} rather than computing them.

But you do not want to choose too many functions either, because the required numerical effort will go up. So there will be an error involved; you will

not get as close to the true best orbitals as you can. One thing this means is that the actual error in the ground state energy will be even larger than true Hartree-Fock would give. For that reason, the Hartree-Fock value of the ground state energy is called the "Hartree-Fock limit:" it is how close you could come to the correct energy if you were able to solve the Hartree-Fock equations exactly.

In short, to compute the Hartree-Fock solution accurately, you want to select a large number of single-electron functions to represent the orbitals. But don't start using zillions of them. The problem is that even the exact Hartree-Fock solution still has a finite error; a wave function cannot in general be described accurately using only a single Slater determinant. So what would the point in computing the very inaccurate numbers to ten digits accuracy?

9.3.5.4 Correlation energy

As the previous subsubsection noted, the Hartree-Fock solution, even if computed exactly, will still have a finite error. You might think that this error would be called something like "Hartree-Fock error." Or maybe "representation error" or "single-determinant error," since it is due to an incomplete representation of the true wave function using a single Slater determinant.

However, the Hartree-Fock error in energy is called "correlation energy." The reason is because there is a energizing correlation between the more impenetrable and poorly defined your jargon, and the more respect you will get for doing all that incomprehensible stuff.

And of course the word "error" should never be used in the first place, God forbid. Or those hated non-experts might figure out that Hartree-Fock has an error in energy so big that it makes the base approximation pretty much useless for chemistry.

To understand what physicists are referring to with "correlation," reconsider the form of the Hartree-Fock wave function, as described in subsection 9.3.1. It consisted of a single Slater determinant. However, that Slater determinant in turn consisted of a lot of Hartree products, the first of which was

$$\psi_1^{\rm s}(\vec{r}_1)\updownarrow_1(S_{z1})\psi_2^{\rm s}(\vec{r}_2)\updownarrow_2(S_{z2})\psi_3^{\rm s}(\vec{r}_3)\updownarrow_3(S_{z3})\ldots\psi_I^{\rm s}(\vec{r}_I)\updownarrow_I(S_{zI})$$

The other Hartree products were different only in the order in which the electrons appear in the product. And since the electrons are all the same, the order does not make a difference: each of these Hartree products has the same expectation energy. Each also satisfies the Pauli exclusion principle but, by itself, *not* the antisymmetrization requirement.

Now, consider what the Born statistical interpretation says about the single Hartree product above. It says that the probability of electron 1 to be within a vicinity of volume $\mathrm{d}^3\vec{r}_1$ around a given position $\vec{r}_1$ with given spin S_{z1}, *and* electron 2 to be within a vicinity of volume $\mathrm{d}^3\vec{r}_2$ around a given position $\vec{r}_2$ with

given spin S_{z2}, etcetera, is given by

$$\begin{aligned} &|\psi_1^{\rm s}(\vec{r}_1)\updownarrow_1(S_{z1})|^2\,{\rm d}^3\vec{r}_1 \\ &\quad\times |\psi_2^{\rm s}(\vec{r}_2)\updownarrow_2(S_{z2})|^2\,{\rm d}^3\vec{r}_2 \\ &\quad\quad\times |\psi_3^{\rm s}(\vec{r}_3)\updownarrow_3(S_{z3})|^2\,{\rm d}^3\vec{r}_3 \\ &\quad\quad\quad\ldots \end{aligned}$$

This takes the form of a probability for electron 1 to be in the given state that is *independent* of where the other electrons are, times a probability for electron 2 to be in the given state that is *independent* of where the other electrons are, etcetera. In short, in a single Hartree product the electrons do not care where the other electrons are. Their positions are "uncorrelated."

Uncorrelated positions would be OK if the electrons did not repel each other. In that case, each electron would indeed not care where the other electrons are. Then all Hartree products would have the same energy, which would also be the energy of the complete Slater determinant.

But electrons do repel each other. So, if electron 1 is at a given position $\vec{r}_1$, electron 2 can reduce its potential energy by preferring positions farther away from that position. It cannot overdo it, as that will increase its kinetic energy too much, but there is some room for improvement. So the exact wave function will have correlations between the positions of different electrons. Based on arguments like that, physicists then come up with the term correlation energy.

Not so fast, physicists! For one, a Slater determinant is not a single Hartree product but already includes some electron correlations. Also, "correlation energy" is not the same as "error in energy caused by incorrect correlations." And "error in energy caused by incorrect correlations" is not the same as "error in energy for an incorrect solution, including incorrect correlations." And the last is what the Hartree-Fock error really is. Note that while there is some rough qualitative relation between potential energy and electron position correlations, you cannot find the potential energy by pontificating about electrons trying to stay away from each other. And the correct energy state is found by delicately balancing subtle reductions in potential energy against subtle increases in kinetic energy. The kinetic energy does not even care about electron correlations. However, the kinetic energy is wrong too when applied on a single Slater determinant.

See note {N.18} for more.

9.3.5.5 Configuration interaction

Since the base Hartree-Fock approximation has an error that is far too big for typical chemistry applications, the next question is what can be done about it. The basic answer is simple: use more that I orbitals, i.e. single-particle wave

functions. As already noted in section 5.7, if you include enough orthonormal basis functions, using all their possible Slater determinants, you can approximate any function to arbitrary accuracy.

After the I, (or $I/2$ in the restricted closed-shell case,) spatial orbitals have been found, the Hartree-Fock operator becomes just a Hermitian operator, and can be used to compute further orthonormal orbitals $\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}, \psi^{\rm s}_{I+2}{\uparrow\downarrow}_{I+2}, \ldots$. You can add these to the mix, say to get a better approximation to the true ground state wave function of the system.

You might want to try to start small. If you include just one more orbital $\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}$, you can already form I more Slater determinants: you can replace any of the I orbitals in the original determinant by the new function $\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}$. So you can now approximate the true wave function by the more general expression

$$\begin{aligned}\Psi \;=\; a_0\Big(&\big|\det \psi^{\rm s}_1{\uparrow\downarrow}_1, \psi^{\rm s}_2{\uparrow\downarrow}_2, \psi^{\rm s}_3{\uparrow\downarrow}_3, \ldots, \psi^{\rm s}_I{\uparrow\downarrow}_I\big\rangle \\ &+ a_1\big|\det \psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}, \psi^{\rm s}_2{\uparrow\downarrow}_2, \psi^{\rm s}_3{\uparrow\downarrow}_3, \ldots, \psi^{\rm s}_I{\uparrow\downarrow}_I\big\rangle \\ &+ a_2\big|\det \psi^{\rm s}_1{\uparrow\downarrow}_1, \psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}, \psi^{\rm s}_3{\uparrow\downarrow}_3, \ldots, \psi^{\rm s}_I{\uparrow\downarrow}_I\big\rangle \\ &+ \ldots \\ &+ a_I\big|\det \psi^{\rm s}_1{\uparrow\downarrow}_1, \psi^{\rm s}_2{\uparrow\downarrow}_2, \psi^{\rm s}_3{\uparrow\downarrow}_3, \ldots, \psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}\big\rangle\Big)\end{aligned}$$

where the coefficients $a_1, a_2, \ldots$ are to be chosen to approximate the ground state energy more closely and a_0 is a normalization constant.

The additional I Slater determinants are called "excited determinants". For example, the first excited state

$$\big|\det \psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}, \psi^{\rm s}_2{\uparrow\downarrow}_2, \psi^{\rm s}_3{\uparrow\downarrow}_3, \ldots, \psi^{\rm s}_I{\uparrow\downarrow}_I\big\rangle$$

is like a state where you excited an electron out of the lowest state $\psi^{\rm s}_1{\uparrow\downarrow}_1$ into an elevated energy state $\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}$.

(However, note that if you really wanted to satisfy the variational requirement $\delta\langle E\rangle = 0$ for such a state, you would have to recompute the orbitals from scratch, using $\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}$ in the Fock operator instead of $\psi^{\rm s}_1{\uparrow\downarrow}_1$. That is not what you want to do here; you do not want to create totally new orbitals, just more of them.)

It may seem that this must be a winner: as much as I more determinants to further minimize the energy. Unfortunately, now you pay the price for doing such a great job with the single determinant. Since, hopefully, the Slater determinant is the best single determinant that can be formed, any changes that are equivalent to simply changing the determinant's orbitals will do no good. And it turns out that the $I+1$-determinant wave function above is equivalent to the single-determinant wave function

$$\Psi = a_0\big|\det \psi^{\rm s}_1{\uparrow\downarrow}_1 + a_1\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}, \psi^{\rm s}_2{\uparrow\downarrow}_2 + a_2\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}, \ldots, \psi^{\rm s}_I{\uparrow\downarrow}_I + a_I\psi^{\rm s}_{I+1}{\uparrow\downarrow}_{I+1}\big\rangle$$

as you can check with some knowledge of the properties of determinants. Since you already have the best single determinant, all your efforts are going to be wasted if you try this.

You might try forming another set of I excited determinants by replacing one of the orbitals in the original Hartree-Fock determinant by $\psi^{\rm s}_{I+2}\updownarrow_{I+2}$ instead of $\psi^{\rm s}_{I+1}\updownarrow_{I+1}$, but the fact is that the infinitesimal variational condition $\delta\langle E\rangle = 0$ is still going to be satisfied when the wave function is the original Hartree-Fock one. For small changes in wave function, the additional determinants can still be pushed inside the Hartree-Fock one. To ensure a decrease in energy, you want to include determinants that allow a nonzero decrease in energy even for small changes from the original determinant, and that requires "doubly" excited determinants, in which two different original states are replaced by excited ones like $\psi^{\rm s}_{I+1}\updownarrow_{I+1}$ and $\psi^{\rm s}_{I+2}\updownarrow_{I+2}$.

Note that you can form $I(I-1)$ such determinants; the number of determinants rapidly explodes when you include more and more orbitals. And a mathematically convergent process would require an asymptotically large set of orbitals, compare chapter 5.7. How big is your computer?

Most people would probably call improving the wave function representation using multiple Slater determinants something like "multiple-determinant representation," or "excited-determinant correction.". However, it is called "configuration interaction." The reason is that every hated non-expert will wonder whether the physicist is talking about the configuration of the nuclei or the electrons, and what it is interacting with.

(Actually, "configuration" refers to the *practitioner* "configuring" all those determinants, no kidding. The interaction is with the computer used to do so. Suppose you were creating the numerical mesh for some finite difference or finite element computation. If you called that "configuration interaction" instead of "mesh generation," because it required you to "configure" all those mesh points through interacting with your computer, some people might doubt your sanity. But in physics, the standards are not so high.)

Chapter 10

Solids

Quantum mechanics is essential to make sense out of the properties of solids. Some of the most important properties of solids were already discussed in chapter 6. It is a good idea to review these sections before reading this chapter.

The discussion will remain restricted to solids that have a "crystal structure." In a crystal the atoms are packed together in a regular manner. Some important materials, like glass and plastic, are amorphous, they do not have such a regular crystal structure, and neither do liquids, so not all the ideas will apply to them.

10.1 Molecular Solids

The hydrogen molecule is the most basic example in quantum mechanics of how atoms can combine into molecules in order to share electrons. So, the question suggests itself whether, if hydrogen molecules are brought close together in a solid, will the atoms start sharing their electrons not just with one other atom, but with all surrounding atoms? The answer under normal conditions is no. Metals do that, but hydrogen under normal conditions does not. Hydrogen atoms are very happy when combined in pairs, and have no desire to reach out to further atoms and weaken the strong bond they have already created. Normally hydrogen is a gas, not a metal.

However, if you cool hydrogen way down to 20 K, it will eventually condense into a liquid, and if you cool it down even further to 14 K, it will then freeze into a solid. That solid still consists of hydrogen molecules, so it is called a molecular solid. (Note that solidified noble gases, say frozen neon, are called molecular solids too, even though they are made up of atoms rather than molecules.)

The forces that glue the hydrogen molecules together in the liquid and solid phases are called Van der Waals forces, and more specifically, they are called London forces. (Van der Waals forces are often understood to be all intermolecular forces, not just London forces.) London forces are also the only forces that can glue noble gas atoms together. These forces are weak.

It is exactly because these forces are so weak that hydrogen must be cooled down so much to condense it into liquid and finally freeze it. At the time of this writing, that is a significant issue in the "hydrogen economy." Unless you go to very unusual temperatures and pressures, hydrogen is a very thin gas, hence extremely bulky.

Helium is even worse; it must be cooled down to 4 K to condense it into a liquid, and under normal pressure it will not freeze into a solid at all. These two, helium and hydrogen are the worst elements of them all, and the reason is that their atoms are so small. Van der Waals forces increase with size.

To explain why the London forces occur is easy; there are in fact two explanations that can be given. There is a simple, logical, and convincing explanation that can easily be found on the web, and that is also completely wrong. And there is a weird quantum explanation that is also correct, {A.33}.

If you are the audience that this book is primarily intended for, you may already know the London forces under the guise of the Lennard-Jones potential. London forces produce an attractive potential between atoms that is proportional to $1/d^6$ where d is a scaled distance between the molecules. So the Lennard-Jones potential is taken to be

$$V_{\rm LJ} = C\left(d^{-12} - d^{-6}\right) \tag{10.1}$$

where C is a constant. The second term represents the London forces.

The first term in the Lennard-Jones potential is there to model the fact that when the atoms get close enough, they rapidly start repelling instead of attracting each other. (See section 5.10 for more details.) The power 12 is computationally convenient, since it makes the first term just the square of the second one. However, theoretically it is not very justifiable. A theoretically more reasonable repulsion would be one of the form $\bar{C}e^{-d/c}/d^n$, with $\bar{C}$, c, and n suitable constants, since that reflects the fact that the strength of the electron wave functions ramps up exponentially when you get closer to an atom. But practically, the Lennard-Jones potential works very well; the details of the first term make no big difference as long as the potential ramps up quickly.

It may be noted that at very large distances, the London force takes the Casimir-Polder form $1/d^7$ rather than $1/d^6$. Charged particles do not really interact directly as a Coulomb potential assumes, but through photons that move at the speed of light. At large separations, the time lag makes a difference, [26]. The separation at which this happens can be ballparked through dimensional arguments. The frequency of a typical photon corresponding to transitions between energy states is given by $\hbar\omega = E$ with E the energy difference between the states. The frequency for light to bounce back and forwards between the molecules is given by c/d, with c the speed of light. It follows that the frequency for light to bounce back and forward is no longer large compared to ω when $Ed/\hbar c$ becomes order one. For hydrogen, E is about 10 eV and $\hbar c$ is about 200

eV nm. That makes the typical separation at which the $1/d^6$ relation breaks down about 20 nm, or 200 Å.

Molecular solids may be held together by other Van der Waals forces besides London forces. Many molecules have an charge distribution that is inherently asymmetrical. If one side is more negative and the other more positive, the molecule is said to have a "dipole strength." The molecules can arrange themselves so that the negative sides of the molecules are close to the positive sides of neighboring molecules and vice versa, producing attraction. (Even if there is no net dipole strength, there will be some electrostatic interaction if the molecules are very close and are not spherically symmetric like noble gas atoms are.)

Chemguide [[1]] notes: "Surprisingly dipole-dipole attractions are fairly minor compared to dispersion [London] forces, and their effect can only really be seen if you compare two molecules with the same number of electrons and the same size." One reason is that thermal motion tends to kill off the dipole attractions by messing up the alignment between molecules. But note that the dipole forces act on top of the London ones, so everything else being the same, the molecules with a dipole strength will be bound together more strongly.

When more than one molecular species is around, species with inherent dipoles can induce dipoles in other molecules that normally do not have them.

Another way molecules can be kept together in a solid is by what are called "hydrogen bonds." In a sense, they too are dipole-dipole forces. In this case, the molecular dipole is created when the electrons are pulled away from hydrogen atoms. This leaves a partially uncovered nucleus, since an hydrogen atom does not have any other electrons to shield it. Since it allows neighboring molecules to get very close to a nucleus, hydrogen bonds can be strong. They remain a lot weaker than a typical chemical bond, though.

Key Points

- Even neutral molecules that do not want to create other bonds can be glued together by various "Van der Waals forces."
- These forces are weak, though hydrogen bonds are much less so.
- The London type Van Der Waals forces affects all molecules, even noble gas atoms.
- London forces can be modeled using the Lennard-Jones potential.
- London forces are one of these weird quantum effects. Molecules with inherent dipole strength feature a more classically understandable version of such forces.

10.2 Ionic Solids

A typical example of a ionic solid is ordinary salt, NaCl. There is little quantitative quantum mechanics required to describe either the salt molecule or solid salt. Still, there are some important qualitative points, so it seems useful to include a discussion in this book. Both molecule and solid will be described in this subsection, since the ideas are very similar.

To form a NaCl salt molecule, a clorine atom takes the loosely bound lone 3s electron away from a natrium (sodium) atom and puts it in its single still vacant 3p position. That leaves a negative chlorine ion with filled K, L, and M shells and a positive natrium ion with just filled K and L shells. Since the combined electron distribution of filled shells is spherically symmetric, you can reasonably think of the two ions as somewhat soft billiard balls. Since they have opposite charge, they stick together into a salt molecule as sketched in figure 10.1. The natrium ion is a bit less than two Å in diameter, the clorine one a bit less than four.

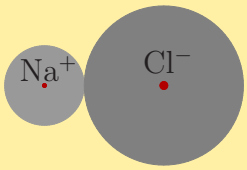


Figure 10.1: Billiard-ball model of the salt molecule.

The energetics of this process is rather interesting. Assume that you start out with a neutral natrium atom and a neutral chlorine atom that are far apart. To take the lone 2s electron out of the natrium atom, and leave it at rest at a position far from either the natrium or the chlorine atom, takes an amount of energy called the "ionization energy" of natrium. Its value is 5.14 eV (electron volts).

To take that free electron at rest and put it into the vacant 3p position of the chlorine ion gives back an amount of energy called the "electron affinity" of clorine. Its value is 3.62 eV.

(Electron affinity, the willingness to take on free electrons, is not to be confused with "electronegativity," the willingness to take on electrons in chemical bonds. Unlike electronegativity, electron affinity varies wildly from element to element in the periodic table. There is some system in it, still, especially within single columns. It may also be noted that there seems to be some disagreement about the definition of electronegativity, in particular for atoms or molecules that cannot stably bind a free electron, {N.19}.)

Anyway, since it takes 5.14 eV to take the electron out of natrium, and you get only 3.62 eV back by putting it into clorine, you may wonder how a salt molecule could ever be stable. But the described picture is very misleading.

It does not really take 5.14 eV to take the electron *out of* natrium; most of that energy is used to pull the liberated electron and positive ion far apart. In the NaCl molecule, they are not pulled far apart; the positive natrium ion and negative chlorine ion stick together as in figure 10.1.

In other words, to create the widely separated positive natrium ion and negative chlorine ion took $5.14 - 3.62$ eV, but watch the energy that is recovered when the two ions are brought together to their correct 2.36 Å separation distance in the molecule. It is approximately given by the Coulomb expression

$$\frac{e^2}{4\pi\epsilon_0}\frac{1}{d}$$

where $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is the permittivity of space and d is the 2.36 Å distance between the nuclei. Putting in the numbers, dropping an e to get the result in eV, this energy is 6.1 eV. That gives the total binding energy as $-5.14 + 3.62 + 6.1$, or 4.58 eV. That is not quite right, but it is close; the true value is 4.26 eV.

There are a few reasons why it is slightly off, but one is that the Coulomb expression above is only correct if the ions were billiard balls that would move unimpeded towards each other until they hit. Actually, the atoms are somewhat softer than billiard balls; their mutual repulsion force ramps up quickly, but not instantaneously. That means that the repulsion force will do a small amount of negative work during the final part of the approach of the ions. Also, the uncertainty principle does not allow the localized ions to have exactly zero kinetic energy. But as you see, these are small effects. It may also be noted that the repulsion between the ions is mostly Pauli repulsion, as described in section 5.10.

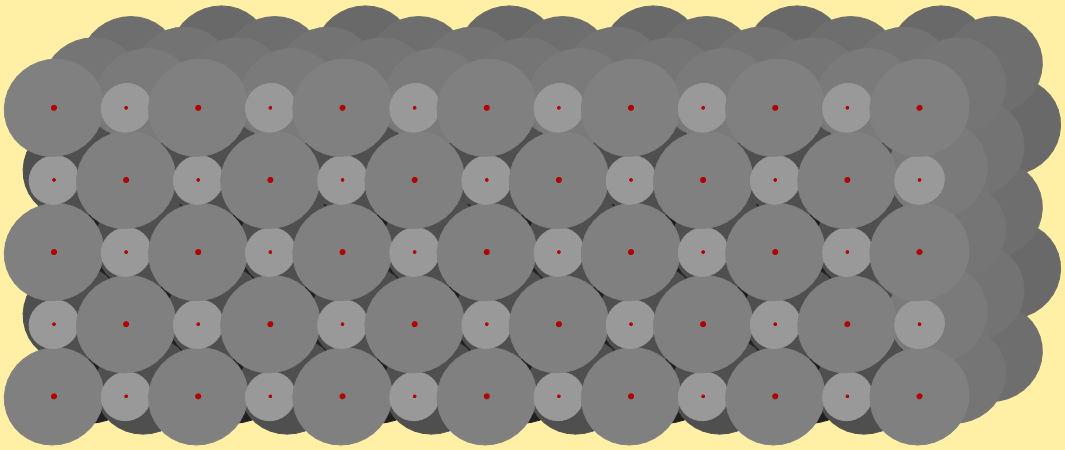


Figure 10.2: Billiard-ball model of a salt crystal.

Now the electrostatic force that keeps the two ions together in the molecule is omni-directional. That means that if you bring a lot of salt molecules together,

the clorine ions will also attract the natrium ions of other molecules and vice versa. As a result, under normal conditions, salt molecules pack together into solid salt crystals, as shown in figure 10.2. The ions arrange themselves very neatly into a pattern that allows each ion to be surrounded by as many attracting ions of the opposite kind as possible. In fact, as figure 10.2 indicates, each ion is surrounded by six ions of the opposite kind: four in the same vertical plane, a fifth behind it, and a sixth in front of it. A more detailed description of the crystal structure will be given next, but first consider what it means for the energy.

Since when the molecules pack into a solid, each ion gets next to six ions of the opposite type, the simplest guess would be that the 6.1 eV Coulomb attraction of the ions in the molecule would increase by a factor 6 in the solid. But that is a bad approximation: in the solid, each ion is not just surrounded by six attracting ions of the opposite kind, but also by twelve repelling ions of the same kind that are only slightly further away, then again eight attracting ions still a bit further away, etcetera. The net effect is that the Coulomb attraction is only 1.75 times higher in the solid than the lone molecules would have. The factor 1.75 is called the "Madelung constant. So, all else being the same, by forming a salt crystal the salt molecules would raise their Coulomb attraction to 1.75×6.1 or 10.7 eV.

That is still not quite right, because in the solid, the ions are farther apart than in the molecule. Recall that in the solid, each attracting ion is surrounded by repelling ions of the opposite kind, reducing the attraction between pairs. In the solid, opposite ions are 2.82 Å apart instead of 2.36, so the Coulomb energy reduces to $10.7 \times 2.36/2.82$ or 8.93 eV. Still, the bottom line is that the molecules pick up about 2.8 eV more Coulomb energy by packing together into salt crystals, and that is quite a bit of energy. So it should not come as a surprise that salt must be heated as high as 801 °C to melt it, and as high as 1 465 °C to boil it.

Finally, consider the crystal structure that the molecules combine into. One way of thinking of it is as a three-dimensional chess board structure. In figure 10.2, think of the frontal plane as a chess board of black and white cubes, with a natrium nucleus in the center of each white cube and a clorine nucleus in the center of each black one. The next plane of atoms can similarly be considered to consists of black and white cubes, where the back cubes are behind the white cubes of the frontal plane and vice-versa. And the same way for further planes.

However, this is not how a material scientist would think about the structure. A material scientist likes to describe a crystal in terms copies of a simple unit, called the "basis," that are stacked together in a regular manner. One possible choice for the basis in salt is a single natrium ion plus a single clorine ion to the right of it, like the molecule of figure 10.1. In figure 10.3 the ions of the salt crystal have been moved far apart to make the actual structure visible, and the two atoms of the basis units have been joined by a blue line. Note that the

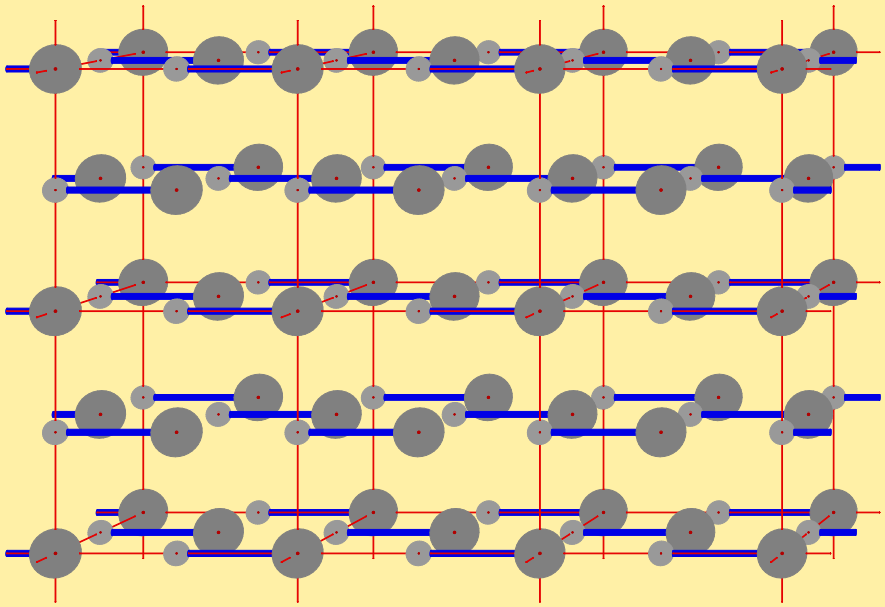


Figure 10.3: The salt crystal disassembled to show its structure.

entire structure consists of these basis units.

But also note that the molecules lose their identity in a ionic solid. You could just as well build up the crystal from vertical "molecules," say, instead of horizontal ones. In fact, there are six reasonable choices of basis, depending on which of its six surrounding chlorine ions you want to associate each natrium ion with. There are of course always countless unreasonable ones...

The regular way in which the bases are stacked together to form the complete crystal structure is called the "lattice." You can think of the volume of the salt crystal as consisting of little cubes called "unit cells" indicated by the red frames in figure 10.3. There are clorine atoms at the corners of the cubes as well as at the center points of the faces of the cubes. That is the reason the salt lattice is called the "face centered cubic" (FCC) lattice. Also note that if you shift the unit cells half a cell to the left, it will be the *natrium* ions that are at the corners and face centers of the cubes. In general, every point of a basis is arranged in the crystal according to the same lattice.

You will agree that it sounds much more professional to say that you have studied the face-centered cubic arrangement of the basis in a NaCl crystal than to say that you have studied the three-dimensional chess board structure of salt.

Key Points

- In a fully ionic bond like NaCl, one atom takes an electron away from another.
- The positive and negative ions stick together by electrostatic force, creating a molecule.

- Because of the same electrostatic force, molecules clump together into strong ionic solids.
- The crystal structure of NaCl consists of copies of a two-atom NaCL basis arranged in a face-centered cubic lattice.

10.3 Metals

Metals are unique in the sense that there is no true molecular equivalent to the way the atoms are bound together in metals. In a metal, the valence electrons are shared on crystal scales, rather than between pairs of atoms. This and subsequent sections will discuss what this really means in terms of quantum mechanics.

10.3.1 Lithium

The simplest metal is lithium. Before examining solid lithium, first consider once more the free lithium atom. Figure 10.4 gives a more realistic picture of the atom than the simplistic analysis of chapter 5.9 did. The atom is really made up of two tightly bound electrons in "$|1s\rangle$" states very close to the nucleus, plus a loosely bound third "valence" electron in an expansive "$|2s\rangle$" state. The core, consisting of the nucleus and the two closely bound 1s electrons, resembles an helium atom that has picked up an additional proton in its nucleus. It will be referred to as the "atom core." As far as the 2s electron is concerned, this entire atom core is not that much different from an hydrogen nucleus: it is compact and has a net charge equivalent to one proton.

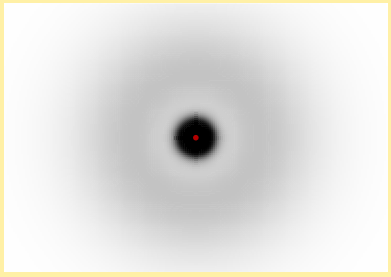


Figure 10.4: The lithium atom, scaled more correctly than before.

One obvious question is then why under normal circumstances lithium is a solid metal and hydrogen is a thin gas. The quantitative difference is that a single-charge core has a favorite distance at which it would like to hold its electron, the Bohr radius. In the hydrogen atom, the electron is about at the Bohr radius, and hydrogen holds onto it tightly. It is willing to share electrons

with one other hydrogen atom, but after that, it is satisfied. It is not looking for any other hydrogen molecules to share electrons with; that would weaken the bond it already has. On the other hand, the 2s electron in the lithium atom is only loosely attached and readily given up or shared among multiple atoms.

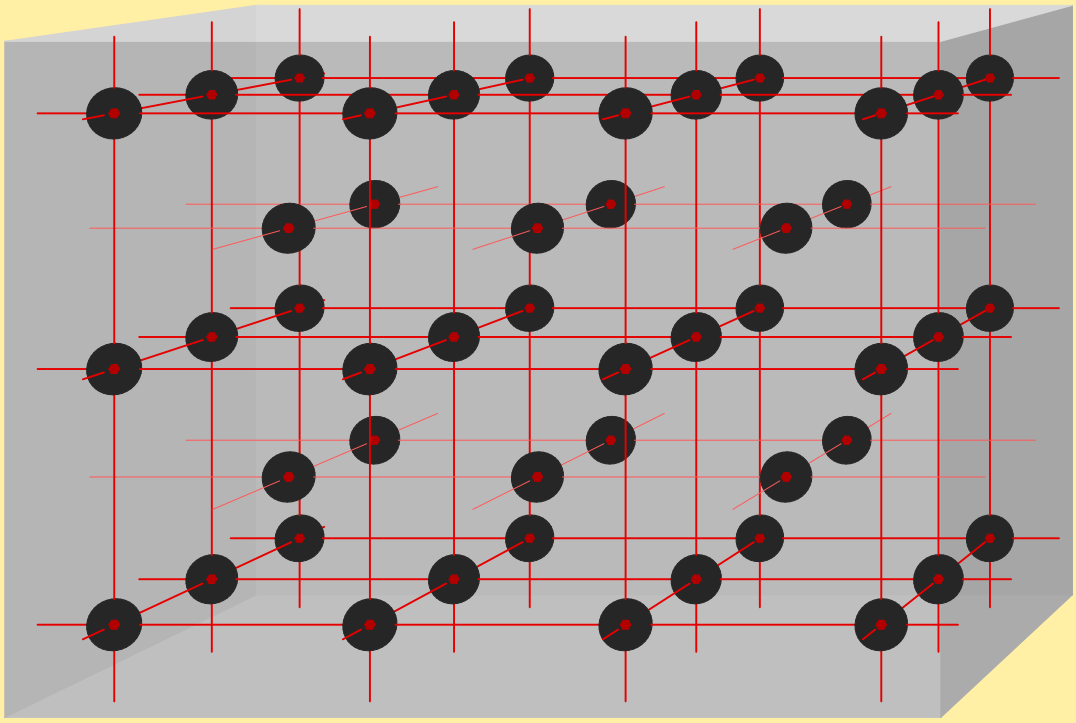


Figure 10.5: Body-centered-cubic (BCC) structure of lithium.

Now consider solid lithium. A perfect lithium crystal would look as sketched in figure 10.5. The atom cores arrange themselves in a regular, repeating, pattern called the " crystal structure." As indicated in the figure by the thick red lines, you can think of the total crystal volume as consisting of many identical little cubes called "(unit) cells.". There are atom cores at all eight corners of these cubes and there is an additional core in the center of the cubic cell. In solid mechanics, this arrangement of positions is referred to as the "body-centered cubic" (BCC) lattice. The crystal "basis" for lithium is a single lithium atom, (or atom core, really); if you put a single lithium atom at every point of the BCC lattice, you get the complete lithium crystal.

Around the atom cores, the 2s electrons form a fairly homogeneous electron density distribution. In fact, the atom cores get close enough together that a typical 2s electron is no closer to the atom core to which it supposedly "belongs" than to the surrounding atom cores. Under such conditions, the model of the 2s electrons being associated with any particular atom core is no longer really meaningful. It is better to think of them as belonging to the solid as a whole, moving freely through it like an electron "gas."

Under normal conditions, bulk lithium is "poly-crystalline," meaning that it consists of many microscopically small crystals, or "grains," each with the above BCC structure. The "grain boundaries" where different crystals meet are crucial to understand the mechanical properties of the material, but not so much to understand its electrical or heat properties, and their effects will be ignored. Only perfect crystals will be discussed.

Key Points

- Lithium can meaningfully be thought of as an atom core, with a net charge of one proton, and a 2s valence electron around it.
- In the solid, the cores arrange themselves into a "body-centered cubic" (BCC) lattice.
- The 2s electrons form an "electron gas" around the cores.
- Normally the solid, like other solids, does not have the same crystal lattice throughout, but consists of microscopic grains, each crystalline, (i.e. with its lattice oriented its own way).
- The grain structure is critical for mechanical properties like strength and plasticity. But that is another book.

10.3.2 One-dimensional crystals

Even the quantum mechanics of a perfect crystal like the lithium one described above is not very simple. So it is a good idea to start with an even simpler crystal. The easiest example would be a "crystal" consisting of only two atoms, but two lithium atoms do not make a lithium crystal, they make a lithium molecule.

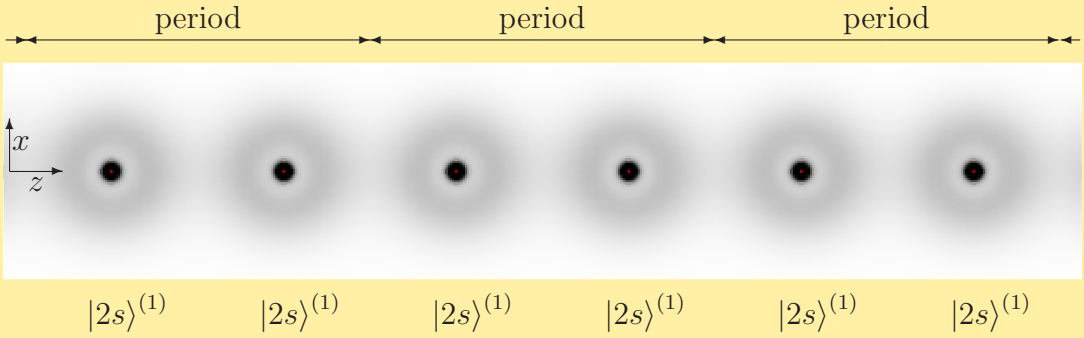


Figure 10.6: Fully periodic wave function of a two-atom lithium "crystal."

Fortunately, there is a dirty trick to get a "crystal" with only two atoms: assume that nature keeps repeating itself as indicated in figure 10.6. Mathematically, this is called "using periodic boundary conditions." It assumes that

after moving towards the left over a distance called the period, you are back at the same point as you started, as if you are walking around in a circle and the period is the circumference.

Of course, this is an outrageous assumption. If nature repeats itself at all, and that is doubtful at the time of this writing, it would be on a cosmological scale, not on the scale of two atoms. But the fact remains that if you make the assumption that nature repeats, the two-atom model gives a much better description of the mathematics of a true crystal than a two-atom molecule would. And if you add more and more atoms, the point where nature repeats itself moves further and further away from the typical atom, making it less and less of an issue for the local quantum mechanics.

Key Points

- Periodic boundary conditions are very artificial.
- Still, for crystal lattices, periodic boundary conditions often work very well.
- And nobody is going to put any *real* grain boundaries into any basic model of solids anyway.

10.3.3 Wave functions of one-dimensional crystals

To describe the energy eigenstates of the electrons in one-dimensional crystals in simple terms, a further assumption must be made: that the detailed interactions between the electrons can be ignored, except for the exclusion principle. Trying to correctly describe the complex interactions between the large numbers of electrons found in a macroscopic solid is simply impossible. And it is not really such a bad assumption as it may appear. In a metal, electron wave functions overlap greatly, and when they do, electrons see other electrons in all directions, and effects tend to cancel out. The equivalent in classical gravity is where you go down far below the surface of the earth. You would expect that gravity would become much more important now that you are surrounded by big amounts of mass at all sides. But they tend to cancel each other out, and gravity is actually reduced. Little gravity is left at the center of the earth. It is not recommended as a vacation spot anyway due to excessive pressure and temperature.

In any case, it will be assumed that for any single electron, the net effect of the atom cores and smeared-out surrounding 2s electrons produces a periodic potential that near every core resembles that of an isolated core. In particular, if the atoms are spaced far apart, the potential near each core is exactly the one of a free lithium atom core. For an electron in this two atom "crystal," the intuitive eigenfunctions would then be where it is around either the first or the second core in the 2s state, (or rather, taking the periodicity into account, around every

first or every second core in each period.) Alternatively, since these two states are equivalent, quantum mechanics allows the electron to hedge its bets and to be about each of the two cores at the same time with some probability.

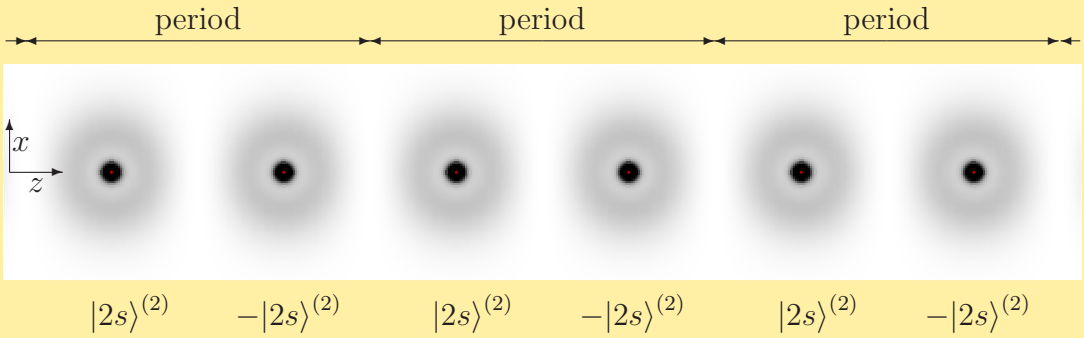


Figure 10.7: Flip-flop wave function of a two-atom lithium "crystal."

But as soon as the atoms are close enough to start noticeably affecting each other, only two true energy eigenfunctions remain, and they are ones in which the electron is around both cores with equal probability. There is one eigenfunction that is exactly the same around both of the atom cores. This eigenfunction is sketched in figure 10.6; it is periodic from core to core, rather than merely from pair of cores to pair of cores. The second eigenfunction is the same from core to core except for a change of sign, call it a flip-flop eigenfunction. It is shown in figure 10.7. Since the grey-scale electron probability distribution only shows the magnitude of the wave function, it looks periodic from atom to atom, but the actual wave function is only the same after moving along two atoms.

To avoid the grey fading away, the shown wave functions have not been normalized; the darkness level is as if the 2s electrons of both the atoms are in that state.

As long as the atoms are far apart, the wave functions around each atom closely resemble the isolated-atom $|2s\rangle$ state. But when the atoms get closer together, differences start to show up. Note for example that the flip-flop wave function is exactly zero half way in between two cores, while the fully periodic one is not. To indicate the deviations from the true free-atom $|2s\rangle$ wave function, parenthetical superscripts will be used.

A one-dimensional crystal made up from four atoms is shown in figure 10.8. Now there are four energy eigenstates. The energy eigenstate that is the same from atom to atom is still there, as is the flip-flop one. But there is now also an energy eigenstate that changes by a factor i from atom to atom, and one that changes by a factor $-$i. They change more slowly from atom to atom than the flip-flop one: it takes two atom distances for them to change sign. Therefore it takes a distance of four atoms, rather than two, for them to return to the same values.

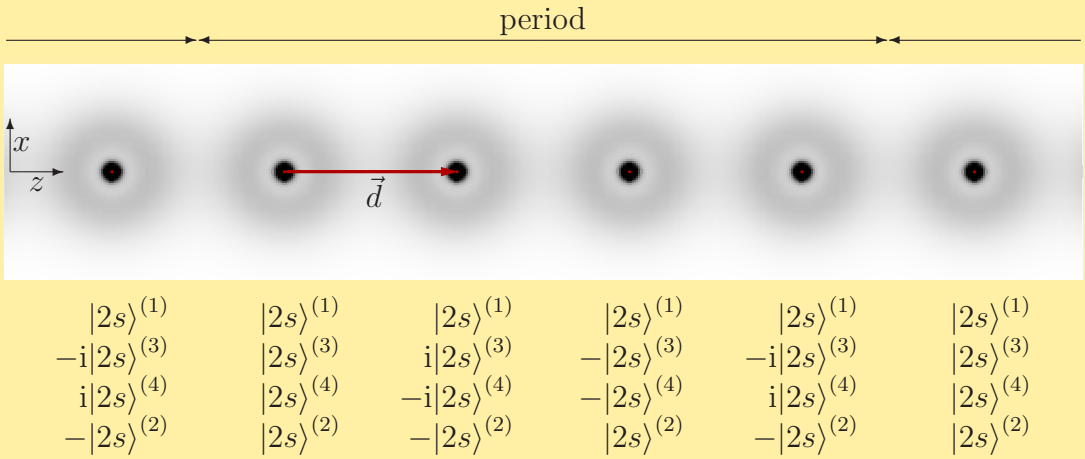


Figure 10.8: Wave functions of a four-atom lithium "crystal." The actual picture is that of the fully periodic mode.

Key Points

- The electron energy eigenfunctions in a metal like lithium extend over the entire crystal.
- If the cores are relatively far apart, near each core the energy eigenfunction of an electron still resembles the 2s state of the free lithium atom.
- However, the magnitude near each core is of course much less, since the electron is spread out over the entire crystal.
- Also, from core to core, the wave function changes by a factor of magnitude one.
- The extreme cases are the fully periodic wave function that changes by a factor one (stays the same) from core to core, versus the flip-flop mode that changes sign completely from one core to the next.
- The other eigenfunctions change by an amount in between these two extremes from core to core.

10.3.4 Analysis of the wave functions

There is a pattern to the wave functions of one-dimensional crystals as discussed in the previous subsection. First of all, while the spatial energy eigenfunctions of the crystal are different from those of the individual atoms, their number is the same. Four free lithium atoms would each have one $|2s\rangle$ spatial state to put their one 2s electron in. Put them in a crystal, and there are still four spatial states to put the four 2s electrons in. But the four spatial states in the crystal

are no longer single atom states; each now extends over the entire crystal. The atoms share all the electrons. If there were eight atoms, the eight atoms would share the eight 2s electrons in eight possible crystal-wide states. And so on.

To be very precise, a similar thing is true of the inner 1s electrons. But since the $|1s\rangle$ states remain well apart, the effects of sharing the electrons are trivial, and describing the 1s electrons as belonging pair-wise to a single lithium nucleus is fine. In fact, you may recall that the antisymmetrization requirement of electrons requires every electron in the universe to be slightly present in every occupied state around every atom. Obviously, you would not want to consider that in the absence of a nontrivial need.

The reason that the energy eigenfunctions take the form shown in figure 10.8 is relatively simple. It follows from the fact that the Hamiltonian commutes with the "translation operator" that shifts the entire wave function over one atom spacing $\vec{d}$. After all, because the potential energy is exactly the same after such a translation, it does not make a difference whether you evaluate the energy before or after you shift the wave function over.

Now commuting operators have a common set of eigenfunctions, so the energy eigenfunctions can be taken to be also eigenfunctions of the translation operator. The eigenvalue must have magnitude one, since periodic wave functions cannot change in overall magnitude when translated. So the eigenvalue describing the effect of an atom-spacing translation on an energy eigenfunction can be written as $e^{\mathrm{i}2\pi\nu}$ with ν a real number. (The factor 2π does nothing except rescale the value of ν. Apparently, crystallographers do not even put it in. This book does so that you do not feel short-changed because other books have factors 2π and yours does not.)

This can be verified for the example energy eigenfunctions shown in figure 10.8. For the fully periodic eigenfunction $\nu = 0$, making the translation eigenvalue $e^{\mathrm{i}2\pi\nu}$ equal to one. So this eigenfunction is multiplied by one under a translation by one atom spacing d: it is the same after such a translation. For the flip-flop mode, $\nu = \frac{1}{2}$; this mode changes by $e^{\mathrm{i}\pi} = -1$ under a translation over an atom spacing d. That means that it changes sign when translated over an atom spacing d. For the two intermediate eigenfunctions $\nu = \pm\frac{1}{4}$, so, using the Euler formula (2.5), they change by factors $e^{\pm\mathrm{i}\pi/2} = \pm\mathrm{i}$ for each translation over a distance d.

In general, for an J-atom periodic crystal, there will be J values of ν in the range $-\frac{1}{2} < \nu \leqslant \frac{1}{2}$. In particular for an even number of atoms J:

$$\nu = \frac{j}{J} \quad \text{for} \quad j = -\frac{J}{2}+1,\ -\frac{J}{2}+2,\ -\frac{J}{2}+3,\ \ldots,\ \frac{J}{2}-1,\ \frac{J}{2}$$

Note that for these values of ν, if you move over J atom spacings, $e^{\mathrm{i}2\pi\nu J} = 1$ as it should; according to the imposed periodic boundary conditions, the wave functions must be the same after J atoms. Also note that it suffices for j to be restricted to the range $-J/2 < j \leqslant J/2$, hence $-\frac{1}{2} < \nu \leqslant \frac{1}{2}$: if j is outside that

range, you can always add or subtract a whole multiple of J to bring it back in that range. And changing j by a whole multiple of J does absolutely nothing to the eigenvalue $e^{\mathrm{i}2\pi\nu}$ since $e^{\mathrm{i}2\pi J/J} = e^{\mathrm{i}2\pi} = 1$.

10.3.5 Floquet (Bloch) theory

Mathematically it is awkward to describe the energy eigenfunctions piecewise, as figure 10.8 does. To arrive at a better way, it is helpful first to replace the axial Cartesian coordinate z by a new "crystal coordinate" u defined by

$$\boxed{z\hat{k} = u\vec{d}} \tag{10.2}$$

where $\vec{d}$ is the vector shown in figure 10.8 that has the length of one atom spacing d. Material scientists call this vector the "primitive translation vector" of the crystal lattice. Primitive vector for short.

The advantage of the crystal coordinate u is that if it changes by one unit, it changes the z-position by exactly one atom spacing. As noted in the previous subsection, such a translation should multiply an energy eigenfunction by a factor $e^{\mathrm{i}2\pi\nu}$. A *continuous* function that does that is the exponential $e^{\mathrm{i}2\pi\nu u}$. And that means that if you factor out that exponential from the energy eigenfunction, what is left does not change under the translation; it will be periodic on atom scale. In other words, the energy eigenfunctions can be written in the form

$$\psi^{\mathrm{p}} = e^{\mathrm{i}2\pi\nu u}\psi^{\mathrm{p}}_{\mathrm{p}}$$

where $\psi^{\mathrm{p}}_{\mathrm{p}}$ is a function that is periodic on the atom scale d; it is the same in each successive interval d.

This result is part of what is called "Floquet theory:"

> *If the Hamiltonian is periodic of period d, the energy eigenfunctions are not in general periodic of period d, but they do take the form of exponentials times functions that are periodic of period d.*

In physics, this result is known as "Bloch's theorem," and the Floquet-type wave function solutions are called "Bloch functions" or "Bloch waves," because Floquet was just a mathematician, and the physicists' hero is Bloch, the physicist who succeeded in doing it too, half a century later. {N.20}.

The periodic part $\psi^{\mathrm{p}}_{\mathrm{p}}$ of the energy eigenfunctions is *not* the same as the $|2\mathrm{s}\rangle^{(.)}$ states of figure 10.8, because $e^{\mathrm{i}2\pi\nu u}$ varies continuously with the crystal position $z = ud$, unlike the factors shown in figure 10.8. However, since the magnitude of $e^{\mathrm{i}2\pi\nu u}$ is one, the magnitudes of $\psi^{\mathrm{p}}_{\mathrm{p}}$ and the $|2\mathrm{s}\rangle^{(.)}$ states are the same, and therefore, so are their grey scale electron probability pictures.

It is often more convenient to have the energy eigenfunctions in terms of the Cartesian coordinate z instead of the crystal coordinate u, writing them in the form

$$\boxed{\psi^{\mathrm{p}}_k = e^{\mathrm{i}kz}\psi^{\mathrm{p}}_{\mathrm{p},k} \text{ with } \psi^{\mathrm{p}}_{\mathrm{p},k} \text{ periodic on the atom scale } d} \tag{10.3}$$

The constant k in the exponential is called the wave number, and subscripts k have been added to $\psi^{\rm p}$ and $\psi^{\rm p}_{\rm p}$ just to indicate that they will be different for different values of this wave number. Since the exponential must still equal $e^{{\rm i}2\pi\nu u}$, clearly the wave number k is proportional to ν. Indeed, substituting $z = ud$ into $e^{{\rm i}kz}$, k can be traced back to be

$$\boxed{k = \nu D \qquad D = \frac{2\pi}{d} \qquad -\tfrac{1}{2} < \nu \leqslant \tfrac{1}{2}} \tag{10.4}$$

10.3.6 Fourier analysis

As the previous subsection explained, the energy eigenfunctions in a crystal take the form of a Floquet exponential times a periodic function $\psi^{\rm p}_{{\rm p},k}$. This periodic part is not normally an exponential. However, it is generally possible to write it as an infinite *sum* of exponentials:

$$\boxed{\psi^{\rm p}_{{\rm p},k} = \sum_{m=-\infty}^{\infty} c_{km} e^{{\rm i}k_m z} \qquad k_m = mD \text{ for } m \text{ an integer}} \tag{10.5}$$

where the c_{km} are constants whose values will depend on x and y, as well as on k and the integer m.

Writing the periodic function $\psi^{\rm p}_{{\rm p},k}$ as such a sum of exponentials is called "Fourier analysis," after another French mathematician. That it is *possible* follows from the fact that these exponentials are the atom-scale-periodic eigenfunctions of the z-momentum operator $p_z = \hbar\partial/{\rm i}\partial z$, as is easily verified by straight substitution. Since the eigenfunctions of an Hermitian operator like p_z are complete, *any* atom-scale-periodic function, including $\psi^{\rm p}_{{\rm p},k}$, can be written as a sum of them. See also {D.8}.

10.3.7 The reciprocal lattice

As the previous two subsections discussed, the energy eigenfunctions in a one-dimensional crystal take the form of a Floquet exponential $e^{{\rm i}kz}$ times a periodic function $\psi^{\rm p}_{{\rm p},k}$. That periodic function can be written as a sum of Fourier exponentials $e^{{\rm i}k_m z}$. It is a good idea to depict all those k-values graphically, to keep them apart. That is done in figure 10.9.

The Fourier k values, $k_m = mD$ with m an integer, form a lattice of points spaced a distance D apart. This lattice is called the "reciprocal lattice." The spacing of the reciprocal lattice, $D = 2\pi/d$, is proportional to the reciprocal of the atom spacing d in the physical lattice. Since on a macroscopic scale the atom spacing d is very small, the spacing of the reciprocal lattice is very large.

The Floquet k value, $k = \nu D$ with $-\frac{1}{2} < \nu \leqslant \frac{1}{2}$, is somewhere in the grey range in figure 10.9. This range is called the first "Brillouin zone.' It is an

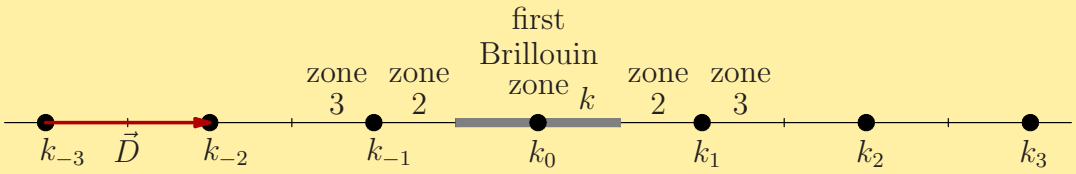


Figure 10.9: Reciprocal lattice of a one-dimensional crystal.

interval, a unit cell if you want, of length D around the origin. The first Brillouin zone is particularly important in the theory of solids. The fact that the Floquet k value may be assumed to be in it is but one reason.

To be precise, the Floquet k value could in principle be in an interval of length D around any wave number k_m, not just the origin, but if it is, you can shift it to the first Brillouin zone by splitting off a factor $e^{\mathrm{i}k_m z}$ from the Floquet exponential $e^{\mathrm{i}kz}$. The $e^{\mathrm{i}k_m z}$ can be absorbed in a redefinition of the Fourier series for the periodic part $\psi^{\mathrm{p}}_{\mathrm{p},k}$ of the wave function, and what is left of the Floquet k value is in the first zone. Often it is good to do so, but not always. For example, in the analysis of the free-electron gas done later, it is critical *not* to shift the k value to the first zone because you want to keep the (there trivial) Fourier series intact.

The first Brillouin zone are the points that are closest to the origin on the k-axis, and similarly the second zone are the points that are second closest to the origin. The points in the interval of length $D/2$ in between k_{-1} and the first Brillouin zone make up half of the second Brillouin zone: they are closest to k_{-1}, but second closest to the origin. Similarly, the other half of the second Brillouin zone is given by the points in between k_1 and the first Brillouin zone. In one dimension, the boundaries of the Brillouin zone fragments are called the "Bragg points." They are either reciprocal lattice points or points half way in between those.

10.3.8 The energy levels

Valence band. Conduction band. Band gap. Crystal. Lattice. Basis. Unit cell. Primitive vector. Bloch wave. Fourier analysis. Reciprocal lattice. Brillouin zones. These are the jargon of solid mechanics; now they have all been defined. (Though certainly not fully discussed.) But jargon is not physics. The physically interesting question is what are the energy levels of the energy eigenfunctions.

For the two-atom crystal of figures 10.6 and 10.7, the answer is much like that for the hydrogen molecular ion of chapter 4.6 and hydrogen molecule of chapter 5.2. In particular, when the atom cores are far apart, the $|2\mathrm{s}\rangle^{(\cdot)}$ states are the same as the free lithium atom wave function $|2\mathrm{s}\rangle$. In either the fully periodic or the flip-flop mode, the electron is with 50% probability in that state around each of the two cores. That means that at large spacing d between

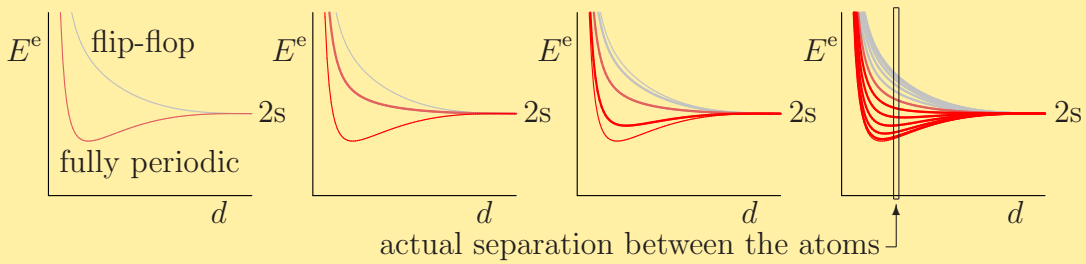


Figure 10.10: Schematic of energy bands.

the cores, the energy is the 2s free lithium atom energy, whether it is the fully periodic or flip-flop mode. That is shown in the left graph of figure 10.10.

When the distance d between the atoms decreases so that the 2s wave functions start to noticeably overlap, things change. As the same left graph in figure 10.10 shows, the energy of the flip-flop state increases, but that of the fully periodic state initially decreases. The reasons for the latter are similar to those that gave the symmetric hydrogen molecular ion and hydrogen molecule states lower energy. In particular, the electrons pick up more effective space to move in, decreasing their uncertainty-principle demanded kinetic energy. Also, when the electron clouds start to merge, the repulsion between electrons is reduced, allowing the electrons to lose potential energy by getting closer to the nuclei of the neighboring atoms. (Note however that the simple model used here would not faithfully reproduce that since the repulsion between the electrons is not correctly modeled.)

Next consider the case of a four-atom crystal, as shown in the second graph of figure 10.10. The fully periodic and flip flop states are unchanged, and so are their energies. But there are now two additional states. Unlike the fully periodic state, these new states vary from atom, but less rapidly than the flip flop mode. As you would then guess, their energy is somewhere in between that of the fully periodic and flip-flop states. Since the two new states have equal energy, it is shown as a double line in 10.10. The third graph in that figure shows the energy levels of an 8 atom crystal, and the final graph that of a 24 atom crystal. When the number of atoms increases, the energy levels become denser and denser. By the time you reach a one hundredth of an inch, one-million atom one-dimensional crystal, you can safely assume that the energy levels within the band have a continuous, rather than discrete distribution.

Now recall that the Pauli exclusion principle allows up to two electrons in a single spatial energy state. Since there are an equal number of spatial states and electrons, that means that the electrons can pair up in the lowest half of the states. The upper states will then be unoccupied. Further, the actual separation distance between the atoms will be the one for which the total energy of the crystal is smallest. The energy spectrum at this actual separation distance is found inside the vanishingly narrow vertical frame in the rightmost graph of

figure 10.10. It shows that lithium forms a metal with a partially-filled band.

The partially filled band means that lithium conducts electricity well. As was already discussed earlier in chapter 6.20, an applied voltage does not affect the band structure at a given location. For an applied voltage to do that, it would have to drop an amount comparable to volts *per atom*. The current that would flow in a metal under such a voltage would vaporize the metal instantly. Current occurs because electrons get excited to states of slightly higher energy that produce motion in a preferential direction.

10.3.9 Merging and splitting bands

The explanation of electrical conduction in metals given in the previous subsection is incomplete. It incorrectly seems to show that beryllium, (and similarly other metals of valence two,) is an insulator. Two valence electrons per atom will completely fill up all 2s states. With all states filled, there would be no possibility to excite electrons to states of slightly higher energy with a preferential direction of motion. There would be no such states. All states would be red in figure 10.10, so nothing could change.

What is missing is consideration of the 2p atom states. When the atoms are far enough apart not to affect each other, the 2p energy levels are a bit higher than the 2s ones and not involved. However, as figure 10.11 shows, when the atom spacing decreases to the actual one in a crystal, the widening bands merge together. With this influx of 300% more states, valence-two metals have plenty of free states to excite electrons to. Beryllium is actually a better conductor than lithium.

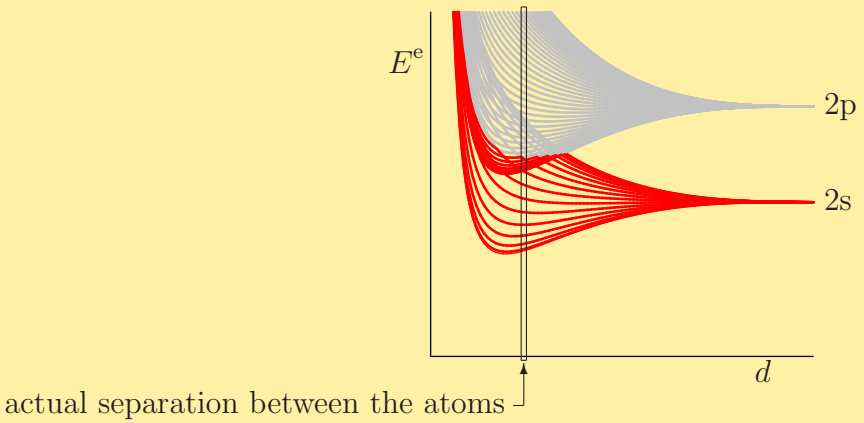


Figure 10.11: Schematic of merging bands.

Hydrogen is a more complicated story. Solid hydrogen consists of molecules and the attractions between different molecules are weak. The proper model of hydrogen is not a series of equally spaced atoms, but a series of pairs of atoms joined into molecules, and with wide gaps between the molecules. When the two

atoms in a single molecule are brought together, the energy varies with distance between the atoms much like the left graph in figure 10.10. The wave function that is the same for the two atoms in the current simple model corresponds to the normal covalent bond in which the electrons are symmetrically shared; the flip-flop function that changes sign describes the "anti-bonding" state in which the two electrons are antisymmetrically shared. In the ground state, both electrons go into the state corresponding to the covalent bond, and the anti-bonding state stays empty. For multiple molecules, each of the two states turns into a band, but since the interactions between the molecules are weak, these two bands do not fan out much. So the energy spectrum of solid hydrogen remains much like the left graph in figure 10.10, with the bottom curve becoming a filled band and the top curve an empty one. An equivalent way to think of this is that the 1s energy level of hydrogen does not fan out into a single band like the 2s level of lithium, but into two half bands, since there are two spacings involved; the spacing between the atoms in a molecule and the spacing between molecules. In any case, because of the band gap energy required to reach the empty upper half 1s band, hydrogen is an insulator.

10.3.10 Three-dimensional metals

The ideas of the previous subsections generalize towards three-dimensional crystals in a relatively straightforward way.

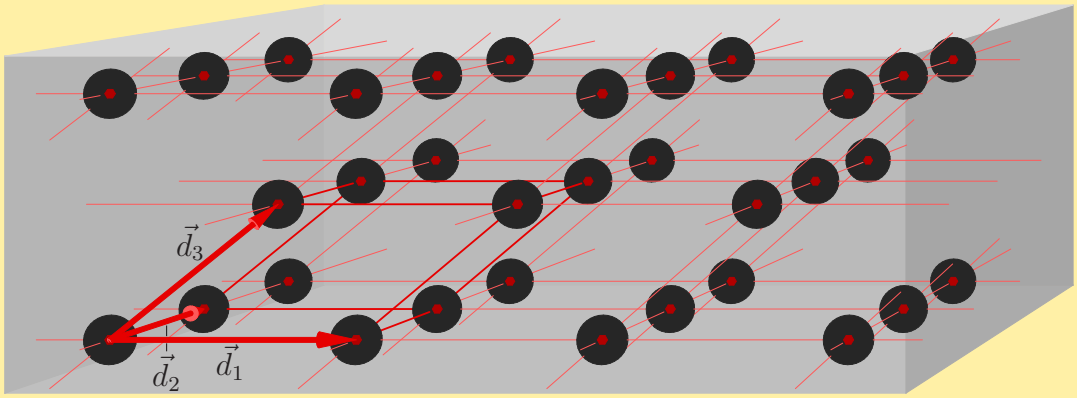


Figure 10.12: A primitive cell and primitive translation vectors of lithium.

As the lithium crystal of figure 10.12 illustrates, in a three-dimensional crystal there are three "primitive translation vectors." The three-dimensional Cartesian position $\vec{r}$ can be written as

$$\boxed{\vec{r} = u_1\vec{d}_1 + u_2\vec{d}_2 + u_3\vec{d}_3} \tag{10.6}$$

where if any of the "crystal coordinates" u_1, u_2, or u_3 changes by exactly one unit, it produces a physically completely equivalent position.

Note that the vectors $\vec{d}_1$ and $\vec{d}_2$ are two bottom sides of the "cubic unit cell" defined earlier in figure 10.5. However, $\vec{d}_3$ is *not* the vertical side of the cube. The reason is that primitive translation vectors must be chosen to allow you to reach *any* point of the crystal from any equivalent point in whole steps. Now $\vec{d}_1$ and $\vec{d}_2$ allow you to step from any point in a horizontal plane to any equivalent point in the same plane. But if $\vec{d}_3$ was vertically upwards like the side of the cubic unit cell, stepping with $\vec{d}_3$ would miss every second horizontal plane. With $\vec{d}_1$ and $\vec{d}_2$ defined as in figure 10.12, $\vec{d}_3$ *must* point to an equivalent point in an immediately adjacent horizontal plane, not a horizontal plane farther away.

Despite this requirement, there are still many ways of choosing the primitive translation vectors other than the one shown in figure 10.12. The usual way is to choose all three to extend towards adjacent cube centers. However, then it gets more difficult to see that no lattice point is missed when stepping around with them.

The parallelepiped shown in figure 10.12, with sides given by the primitive translation vectors, is called the "primitive cell." It is the smallest building block that can be stacked together to form the total crystal. The cubic unit cell from figure 10.5 is not a primitive cell since it has twice the volume. The cubic unit cell is instead called the "conventional cell."

Since the primitive vectors are not unique, the primitive cell they define is not either. These primitive cells are purely mathematical quantities; an arbitrary choice for the smallest single volume element from which the total crystal volume can be build up. The question suggests itself whether it would not be possible to define a primitive cell that has some physical meaning; whose definition is unique, rather than arbitrary. The answer is yes, and the unambiguously defined primitive cell is called the "Wigner-Seitz cell." The Wigner-Seitz cell around a lattice point is the vicinity of locations that are closer to that lattice point than to any other lattice point.

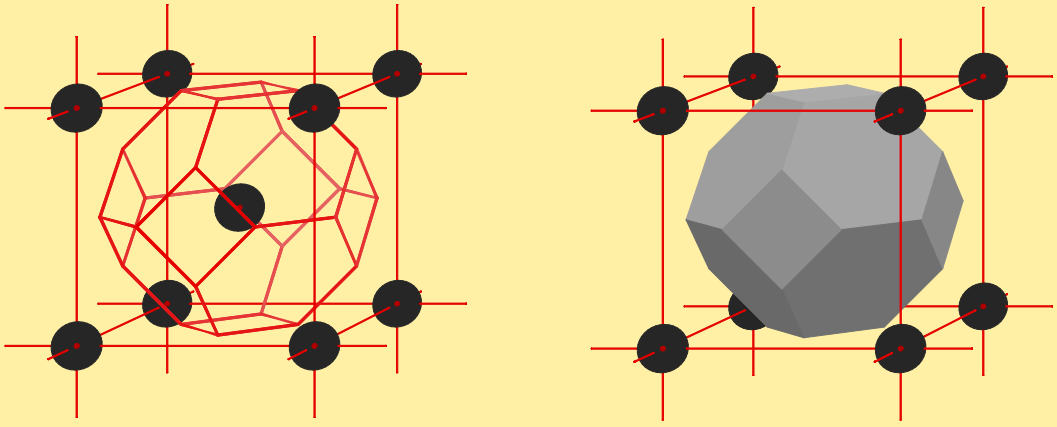


Figure 10.13: Wigner-Seitz cell of the BCC lattice.

Figure 10.13 shows the Wigner-Seitz cell of the BCC lattice. To the left, it is

shown as a wire frame, and to the right as an opaque volume element. To put it within context, the atom around which this Wigner-Seitz cell is centered was also put in the center of a conventional cubic unit cell. Note how the Wigner-Seitz primitive cell is much more spherical than the parallelepiped-shaped primitive cell shown in figure 10.12. The outside surface of the Wigner-Seitz cell consists of hexagonal planes on which the points are just on the verge of getting closer to a corner atom of the conventional unit cell than to the center atom, and of squares on which the points are just on the verge of getting closer to the center atom of an adjacent conventional unit cell. The squares are located within the faces of the conventional unit cell.

The reason that the entire crystal volume can be build up from Wigner-Seitz cells is simple: every point must be closest to some lattice point, so it must be in some Wigner-Seitz cell. When a point is equally close to two nearest lattice points, it is on the boundary where adjacent Wigner-Seitz cells meet.

Turning to the energy eigenfunctions, they can now be taken to be eigenfunctions of three translation operators; they will change by some factor $e^{\mathrm{i}2\pi\nu_1}$ when translated over $\vec{d}_1$, by $e^{\mathrm{i}2\pi\nu_2}$ when translated over $\vec{d}_2$, and by $e^{\mathrm{i}2\pi\nu_3}$ when translated over $\vec{d}_3$. All that just means that they must take the Floquet (Bloch) function form

$$\psi^{\mathrm{p}} = e^{\mathrm{i}2\pi(\nu_1 u_1 + \nu_2 u_2 + \nu_3 u_3)} \psi^{\mathrm{p}}_{\mathrm{p}},$$

where $\psi^{\mathrm{p}}_{\mathrm{p}}$ is periodic on atom scales, exactly the same after one unit change in any of the crystal coordinates u_1, u_2 or u_3.

It is again often convenient to write the Floquet exponential in terms of normal Cartesian coordinates. To do so, note that the relation giving the physical position $\vec{r}$ in terms of the crystal coordinates u_1, u_2, and u_3,

$$\vec{r} = u_1 \vec{d}_1 + u_2 \vec{d}_2 + u_3 \vec{d}_3$$

can be inverted to give the crystal coordinates in terms of the physical position, as follows:

$$\boxed{u_1 = \frac{1}{2\pi} \vec{D}_1 \cdot \vec{r} \quad u_2 = \frac{1}{2\pi} \vec{D}_2 \cdot \vec{r} \quad u_3 = \frac{1}{2\pi} \vec{D}_3 \cdot \vec{r}} \tag{10.7}$$

(Again, factors 2π have been thrown in merely to fully satisfy even the most demanding quantum mechanics reader.) To find the vectors $\vec{D}_1$, $\vec{D}_2$, and $\vec{D}_3$, simply solve the expression for $\vec{r}$ in terms of u_1, u_2, and u_3 using linear algebra procedures. In particular, they turn out to be the rows of the inverse of matrix $(\vec{d}_1, \vec{d}_2, \vec{d}_3)$.

If you do not know linear algebra, it can be done geometrically: if you dot the expression for $\vec{r}$ above with $\vec{D}_1/2\pi$, you must get u_1; for that to be true, the first three conditions below are required:

$$\boxed{\begin{array}{lll} \vec{d}_1 \cdot \vec{D}_1 = 2\pi, & \vec{d}_2 \cdot \vec{D}_1 = 0, & \vec{d}_3 \cdot \vec{D}_1 = 0, \\ \vec{d}_1 \cdot \vec{D}_2 = 0, & \vec{d}_2 \cdot \vec{D}_2 = 2\pi, & \vec{d}_3 \cdot \vec{D}_2 = 0, \\ \vec{d}_1 \cdot \vec{D}_3 = 0, & \vec{d}_2 \cdot \vec{D}_3 = 0, & \vec{d}_3 \cdot \vec{D}_3 = 2\pi. \end{array}} \tag{10.8}$$

The second set of three equations is obtained by dotting with $\vec{D}_2/2\pi$ to get u_2 and the third by dotting with $\vec{D}_3/2\pi$ to get u_3. From the last two equations in the first row, it follows that vector $\vec{D}_1$ must be orthogonal to both $\vec{d}_2$ and $\vec{d}_3$. That means that you can get $\vec{D}_1$ by first finding the vectorial cross product of vectors $\vec{d}_2$ and $\vec{d}_3$ and then adjusting the length so that $\vec{d}_1 \cdot \vec{D}_1 = 2\pi$. In similar ways, $\vec{D}_2$ and $\vec{D}_3$ may be found.

If the expressions for the crystal coordinates are substituted into the exponential part of the Bloch functions, the result is

$$\boxed{\psi^{\rm p}_{\vec{k}} = e^{{\rm i}\vec{k}\cdot\vec{r}}\psi^{\rm p}_{{\rm p},\vec{k}} \quad \vec{k} = \nu_1\vec{D}_1 + \nu_2\vec{D}_2 + \nu_3\vec{D}_3} \tag{10.9}$$

So, in three dimensions, a wave number k becomes a "wave number vector" $\vec{k}$.

Just like for the one-dimensional case, the periodic function $\psi^{\rm p}_{{\rm p},\vec{k}}$ too can be written in terms of exponentials. Converted from crystal to physical coordinates, it gives:

$$\boxed{\begin{array}{c} \psi^{\rm p}_{{\rm p},\vec{k}} = \sum_{m_1}\sum_{m_2}\sum_{m_3} c_{p,\vec{k}\vec{m}} e^{{\rm i}\vec{k}_{\vec{m}}\cdot\vec{r}} \\ \vec{k}_{\vec{m}} = m_1\vec{D}_1 + m_2\vec{D}_2 + m_3\vec{D}_3 \text{ for } m_1,\ m_2, \text{ and } m_3 \text{ integers} \end{array}} \tag{10.10}$$

If these wave number vectors $\vec{k}_{\vec{m}}$ are plotted three-dimensionally, it again forms a lattice called the "reciprocal lattice," and its primitive vectors are $\vec{D}_1$, $\vec{D}_2$, and $\vec{D}_3$. Remarkably, the *reciprocal* lattice to lithium's BCC physical lattice turns out to be the FCC lattice of NaCl fame!

And now note the beautiful symmetry in the relations (10.8) between the primitive vectors $\vec{D}_1$, $\vec{D}_2$, and $\vec{D}_3$ of the reciprocal lattice and the primitive vectors $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$ of the physical lattice. Because these relations involve both sets of primitive vectors in exactly the same way, if a physical lattice with primitive vectors $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$ has a reciprocal lattice with primitive vectors $\vec{D}_1$, $\vec{D}_2$, and $\vec{D}_3$, then a physical lattice with primitive vectors $\vec{D}_1$, $\vec{D}_2$, and $\vec{D}_3$ has a reciprocal lattice with primitive vectors $\vec{d}_1$, $\vec{d}_2$, and $\vec{d}_3$. Which means that since NaCl's FCC lattice is the reciprocal to lithium's BCC lattice, lithium's BCC lattice is the reciprocal to NaCl's FCC lattice. You now see where the word "reciprocal" in reciprocal lattice comes from. Lithium and NaCl borrow each other's lattice to serve as their lattice of wave number vectors.

Finally, how about the definition of the "Brillouin zones" in three dimensions? In particular, how about the first Brillouin zone to which you often prefer to move the Floquet wave number vector $\vec{k}$? Well, it is the magnitude of the wave number vector that is important, so the first Brillouin zone is defined to be the Wigner-Seitz cell around the origin in the reciprocal lattice. Note that this means that in the first Brillouin zone, ν_1, ν_2, and ν_3 are not simply

numbers in the range from $-\frac{1}{2}$ to $\frac{1}{2}$ as in one dimension; that would give a parallelepiped-shaped primitive cell instead.

Solid state physicists may tell you that the other Brillouin zones are also reciprocal lattice Wigner-Seitz cells, [29, p. 38], but if you look closer at what they are actually doing, the higher zones consist of *fragments* of reciprocal lattice Wigner-Seitz cells that can be assembled together to produce a Wigner-Seitz cell shape. Like for the one-dimensional crystal, the second zone are again the points that are second closest to the origin, etcetera.

The boundaries of the Brillouin zone fragments are now planes called "Bragg planes." Each is a perpendicular bisector of a lattice point and the origin. That is so because the locations where points stop being first/, second/, third/, ... closest to the origin and become first/, second/, third/, ... closest to some other reciprocal lattice point must be on the bisector between that lattice point and the origin. Sections 10.5.1 and 10.6 will give Bragg planes and Brillouin zones for a simple cubic lattice.

The qualitative story for the valence electron energy levels is the same in three dimensions as in one. Sections 10.5 and 10.6 will look a bit closer at them quantitatively.

10.4 Covalent Materials

In covalent materials, the atoms are held together by covalent chemical bonds. Such bonds are strong. Note that the classification is somewhat vague; many crystals, like quartz (silicon dioxide), have partly ionic, partly covalent binding. Another ambiguity occurs for graphite, the stable form of carbon under normal condition. Graphite consists of layers of carbon atoms arranged in a hexagonal pattern. There are four covalent bonds binding each carbon to three neighboring atoms in the layer: three sp^2 hybrid bonds in the plane and a fourth π-bond normal it. The π-electrons are delocalized and will conduct electricity. (When rolled into carbon nanotubes, this becomes a bit more complicated.) As far as the binding of the solid is concerned, however, the point is that different layers of graphite are only held together with weak Van der Waals forces, rather than covalent bonds. This makes graphite one of the softest solids known.

Under pressure, carbon atoms can form diamond rather than graphite, and diamond is one of the hardest substances known. The diamond structure is a very clean example of purely covalent bonding, and this section will have a look at its nature. Other group IV elements in the periodic table, in particular silicon, germanium, and grey tin also have the diamond structure. All these, of course, are very important for engineering applications.

One question that suggests itself in view of the earlier discussion of metals is why these materials are not metals. Consider carbon for example. Compared to beryllium, it has four rather than two electrons in the second, L, shell. But

the merged 2s and 2p bands can hold eight electrons, so that cannot be the explanation. In fact, tin comes in two forms under normal conditions: covalent grey tin is stable below 13 °C; while above that temperature, metallic white tin is the stable form. It is often difficult to guess whether a particular element will form a metallic or covalent substance near the middle of the periodic table.

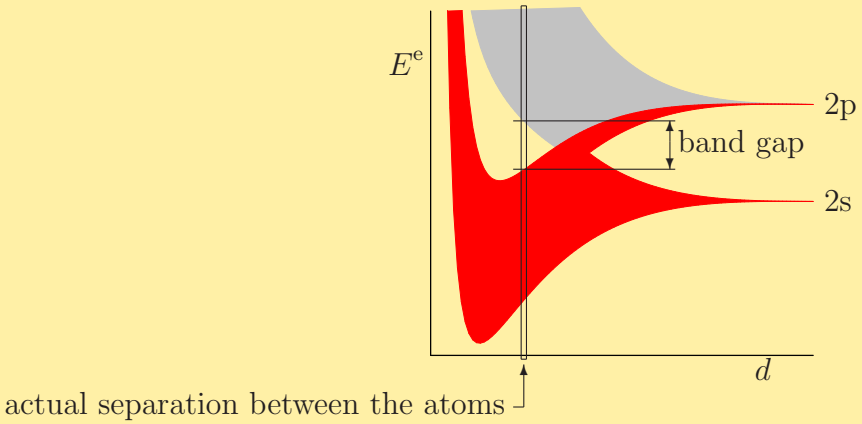


Figure 10.14: Schematic of crossing bands.

Figure 10.14 gives a schematic of the energy band structure for a diamond-type crystal when the spacing between the atoms is artificially changed. When the atoms are far apart, i.e. d is large, the difference from beryllium is only that carbon has two electrons in 2p states versus beryllium none. But when the carbon atoms start coming closer, they have a group meeting and hit upon the bright idea to reduce their energy even more by converting their one 2s and three 2p spatial states into four hybrid sp^3 states. This allows them to share pairs of electrons symmetrically in as much as four strong covalent bonds. And it does indeed work very well for lowering the energy of these states, filled to the gills with electrons. But it does not work well at all for the "anti-bonding" states that share the electrons antisymmetrically, (as discussed for the hydrogen molecule in chapter 5.2.4), and who do not have a single electron to support their case at the meeting. So a new energy gap now opens up.

At the actual atom spacing of diamond, this band gap has become as big as 5.5 eV, making it an electric insulator (unlike graphite, which is a semi-metal). For silicon however, the gap is a much smaller 1.1 eV, similar to the one for germanium of 0.7 eV; grey tin is considerably smaller still; recent authoritative sources list it as zero. These smaller band gaps allow noticeable numbers of electrons to get into the empty conduction band by thermal excitation, so these materials are semiconductors at room temperature.

The crystal structure of these materials is rather interesting. It must allow each atom core to connect to 4 others to form the hybrid covalent bonds. That requires the rather spacious structure sketched in figure 10.15. For simplicity

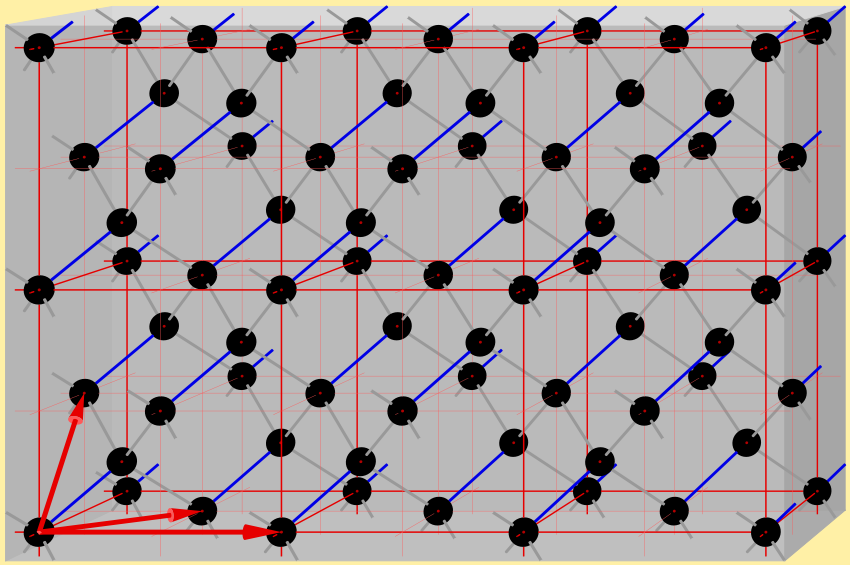


Figure 10.15: Ball and stick schematic of the diamond crystal.

and clarity, the four hybrid bonds that attach each atom core to its four neighbors are shown as blue or dark grey sticks rather than as a distribution of grey tones.

Like for lithium, you can think of the spheres as representing the inner electrons. The grey gas represents the outer electrons, four per atom.

To understand the figure beyond that, first note that it turns out to be impossible to create the diamond crystal structure from a basis of a single atom. It is simply not possible to distribute clones of a single carbon atom around using a *single* set of three primitive vectors, and produce all the atoms in the diamond crystal. A basis of a pair of atoms is needed. The choice of which pair is quite arbitrary, but in figure 10.15 the clones of the chosen pair are linked by blue lines. Notice how the entire crystal is build up from such clones. (Physically, the choice of basis is artificial, and the blue sticks indicate hybrid bonds just like the grey ones.) One possible choice for a set of three primitive translation vectors is shown in the figure. The more usual choice is to take the one in the front plane to the atom located at 45 degrees instead.

Now notice that the lower members of these pairs are located at the corners and face centers of the cubic volume elements indicated by the fat red lines. Yes, diamond is another example of a face-centered cubic lattice. What is different from the NaCl case is the basis; two carbon atoms at some weird angle, instead of a natrium and a chlorine ion sensibly next to each other. Actually, if you look a bit closer, you will notice that in terms of the *half-size* cubes indicated by thin red frames, the structure is not that illogical. It is again that of a three-dimensional chess board, where the centers of the black cubes contain the upper carbon of a basis clone, while the centers of the white cubes are empty.

But of course, you would not want to tell people that. They might think you spend your time playing games, and terminate your support.

If you look at the massively cross-linked diamond structure, it may not come as that much of a surprise that diamond is the hardest substance to occur naturally. Under normal conditions, diamond will supposedly degenerate extremely slowly into graphite, but without doubt, diamonds are forever.

10.5 Free-Electron Gas

Chapter 6 discussed the model of noninteracting electrons in a periodic box. This simple model, due to Sommerfeld, is a first starting point for much analysis of solids. It was used to provide explanations of such effects as the incompressibility of solids and liquids, and of electrical conduction. This section will use the model to explain some of the analytical methods that are used to analyze electrons in crystals. A free-electron gas is a model for electrons in a crystal when the physical effect of the crystal structure on the electrons is ignored. The assumption is that the crystal structure is still there, but that it does not actually do anything to the electrons.

The single-particle energy eigenfunctions of a periodic box are given by

$$\psi^{\rm p}_{\vec{k}}(\vec{r}) = \frac{1}{\sqrt{\mathcal{V}}} e^{{\rm i}\vec{k}\cdot\vec{r}} = \frac{1}{\sqrt{\mathcal{V}}} e^{{\rm i}(k_x x + k_y y + k_z z)} \tag{10.11}$$

Here the wave numbers are related to the box dimensions as

$$k_x = n_x \frac{2\pi}{\ell_x} \qquad k_y = n_y \frac{2\pi}{\ell_y} \qquad k_z = n_z \frac{2\pi}{\ell_z} \tag{10.12}$$

where the quantum numbers n_x, n_y, and n_z are integers. This section will use the wave number vector, rather than the quantum numbers, to indicate the individual eigenfunctions.

Note that each of these eigenfunctions can be regarded as a Bloch wave: the exponentials are the Floquet ones, and the periodic parts are trivial constants. The latter reflects the fact the periodic potential itself is trivially constant (zero) for a free-electron gas.

Of course, there is a spin-up version $\psi^{\rm p}_{\vec{k}}\downarrow$ and a spin-down version $\psi^{\rm p}_{\vec{k}}\uparrow$ of each eigenfunction above. However, spin will not be much of an issue in the analysis here.

The Floquet exponentials have not been shifted to any first Brillouin zone. In fact, since the electrons experience no forces, as far as they are concerned, there is no crystal structure, hence no Brillouin zones.

10.5.1 Lattice for the free electrons

As far as the mathematics of free electrons is concerned, the box in which they are confined may as well be empty. However,it is useful to put the results in

context of a surrounding crystal lattice anyway. That will allow some of the basic concepts of the solid mechanics of crystals to be defined within a simple setting.

It will therefore be assumed that there is a crystal lattice, but that its potential is zero. So the lattice does not affect the motion of the electrons. An appropriate choice for this lattice must now be made. The plan is to keep the same Floquet wave number vectors as for the free electrons in a rectangular periodic box. Those wave numbers form a rectangular grid in wave number space as shown in figure 6.17 of chapter 6.18. To preserve these wave numbers, it is best to figure out a suitable reciprocal lattice first.

To do so, compare the general expression for the Fourier $\vec{k}_{\vec{m}}$ values that make up the reciprocal lattice:

$$\vec{k}_{\vec{m}} = m_1\vec{D}_1 + m_2\vec{D}_2 + m_3\vec{D}_3$$

in which m_1, m_2, and m_3 are integers, with the Floquet $\vec{k}$ values,

$$\vec{k} = \nu_1\vec{D}_1 + \nu_2\vec{D}_2 + \nu_3\vec{D}_3$$

(compare section 10.3.10.) Now ν_1 is of the form $\nu_1 = j_1/J_1$ where j_1 is an integer just like m_1 is an integer, and J_1 is the number of lattice cells in the direction of the first primitive vector. For a macroscopic crystal, J_1 will be a very large number, so the conclusion must be that the Floquet wave numbers are spaced much more closely together than the Fourier ones. And so they are in the other two directions.

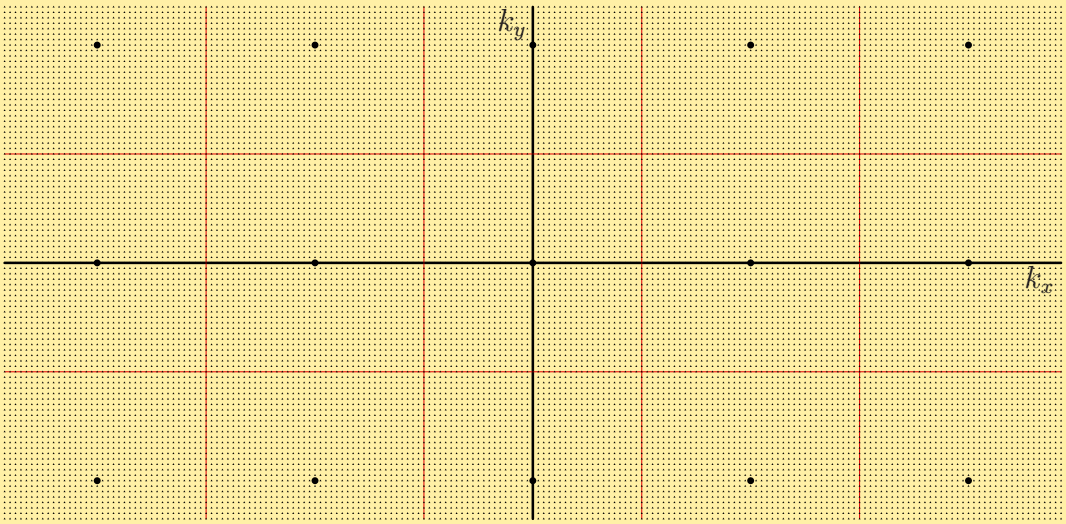


Figure 10.16: Assumed simple cubic reciprocal lattice, shown as black dots, in cross-section. The boundaries of the surrounding primitive cells are shown as thin red lines.

In particular, if it is assumed that there are an equal number of cells in each primitive direction, $J_1 = J_2 = J_3 = J$, then the Fourier wave numbers are spaced farther apart than the Floquet ones by a factor J in each direction. Such a reciprocal lattice is shown as fat black dots in figure 10.16.

Note that in this section, the wave number space will be shown only in the $k_z = 0$ cross-section. A full three-dimensional space, like the one of figure 6.17, would get very messy when crystal structure effects are added.

A lattice like the one shown in figure 10.16 is called a "simple cubic lattice," and it is the easiest lattice that you can define. The primitive vectors are orthonormal, just a multiple of the Cartesian unit vectors $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$. Each lattice point can be taken to be the center of a primitive cell that is a cube, and this cubic primitive cell just happens to be the Wigner-Seitz cell too.

It is of course not that strange that the simple cubic lattice would work here, because the assumed wave number vectors were derived for electrons in a rectangular periodic box.

How about the physical lattice? That is easy too. The simple cubic lattice is its own reciprocal. So the physical crystal too consists of cubic cells stacked together. (Atomic scale ones, of course, for a physical lattice.) In particular, the wave numbers as shown in figure 10.16 correspond to a crystal that is macroscopically a cube with equal sides 2ℓ, and that on atomic scale consists of $J \times J \times J$ identical cubic cells of size $d = 2\ell/J$. Here J, the number of atom-scale cells in each direction, will be a very large number, so d will be very small.

In $\vec{k}$-space, J is the number of Floquet points in each direction within a unit cell. Figure 10.16 would correspond to a physical crystal that has only 40 atoms in each direction. A real crystal would have many thousands, and the Floquet points would be much more densely spaced than could be shown in a figure like figure 10.16.

It should be pointed out that the simple cubic lattice, while definitely simple, is not that important physically unless you happen to be particularly interested in polonium or compounds like cesium chloride or beta brass. But the mathematics is really no different for other crystal structures, just messier, so the simple cubic lattice makes a good example. Furthermore, many other lattices feature cubic unit cells, even if these cells are a bit larger than the primitive cell. That means that the assumption of a potential that has cubic periodicity on an atomic scale is quite widely applicable.

10.5.2 Occupied states and Brillouin zones

The previous subsection chose the reciprocal lattice in wave number space to be the simple cubic one. The next question is how the occupied states show up in it. As usual, it will be assumed that the crystal is in the ground state, corresponding to zero absolute temperature.

As shown in figure 6.17, in the ground state the energy levels occupied by electrons form a sphere in wave number space. The surface of the sphere is the Fermi surface. The corresponding single-electron energy is the Fermi energy.

Figure 10.17 shows the occupied states in $k_z = 0$ cross section if there are one, two, and three valence electrons per physical lattice cell. (In other words, if there are J^3, $2J^3$, and $3J^3$ valence electrons.) For one valence electron per lattice cell, the spherical region of occupied states stays within the first Brillouin zone, i.e. the Wigner-Seitz cell around the origin, though just barely. There are J^3 spatial states in a Wigner-Seitz cell, the same number as the number of physical lattice cells, and each can hold two electrons, (one spin up and one spin down,) so half the states in the first Brillouin zone are filled. For two electrons per lattice cell, there are just as many occupied spatial states as there are states within the first Brillouin zone. But since in the ground state, the occupied free electron states form a spherical region, rather than a cubic one, the occupied states spill over into immediately adjacent Wigner-Seitz cells. For three valence electrons per lattice cell, the occupied states spill over into still more neighboring Wigner-Seitz cells. (It is hard to see, but the diameter of the occupied sphere is slightly larger than the diagonal of the Wigner-Seitz cell cross-section.)

However, these results may show up presented in a different way in literature. The reason is that a Bloch-wave representation is not unique. In terms of Bloch waves, the free-electron exponential solutions as used here can be represented in the form

$$\psi^{\mathrm{p}}_{\vec{k}} = e^{\mathrm{i}\vec{k}\cdot\vec{r}}\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$$

where the atom-scale periodic part $\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$ of the solution is a trivial constant. In addition, the Floquet wave number $\vec{k}$ can be in any Wigner-Seitz cell, however far away from the origin. Such a description is called an "extended zone scheme".

This free-electron way of thinking about the solutions is often not the best way to understand the physics. Seen within a single physical lattice cell, a solution with a Floquet wave number in a Wigner-Seitz cell far from the origin looks like an extremely rapidly varying exponential. However, all of that *atom-scale* physics is in the *crystal-scale* Floquet exponential; the lattice-cell scale part $\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$ is a trivial constant. It may be better to shift the Floquet wave number to the Wigner-Seitz cell around the origin, the first Brillouin zone. That will turn the crystal-scale Floquet exponential into one that varies relatively slowly over the physical lattice cell; the rapid variation will now be absorbed into the lattice-cell part $\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$. This idea is called the "reduced zone scheme." As long as the Floquet wave number vector is shifted to the first Brillouin zone by whole amounts of the primitive vectors of the reciprocal lattice, $\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$ will remain an atom-scale-periodic function; it will just become nontrivial. This shifting of the Floquet wave numbers to the first Brillouin zone is illustrated in figures 10.18*a*

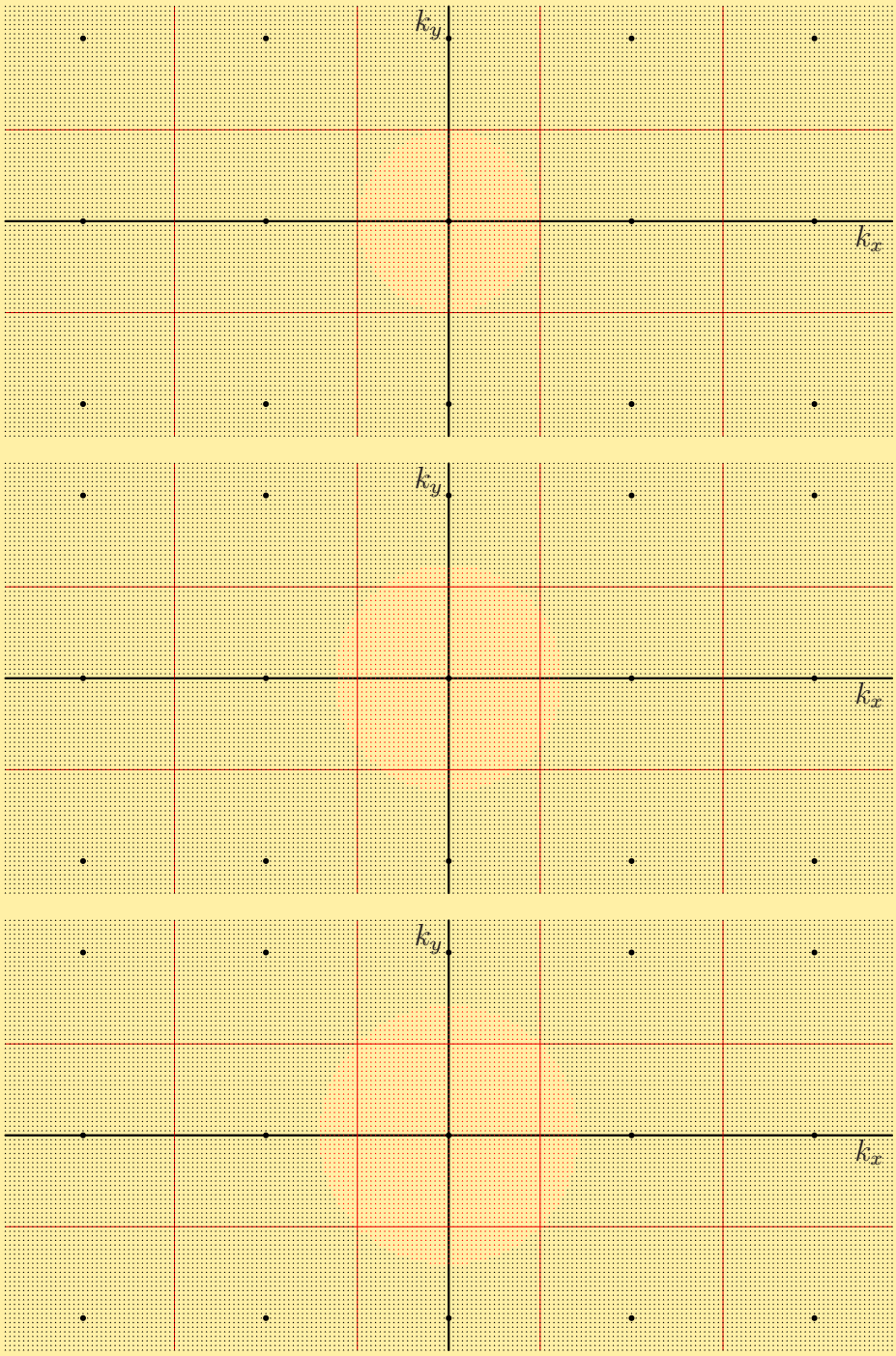


Figure 10.17: Occupied states for one, two, and three free electrons per physical lattice cell.

and 10.18*b*. The figures are for the case of three valence electrons per lattice cell, but with a slightly increased radius of the sphere to avoid visual ambiguity.

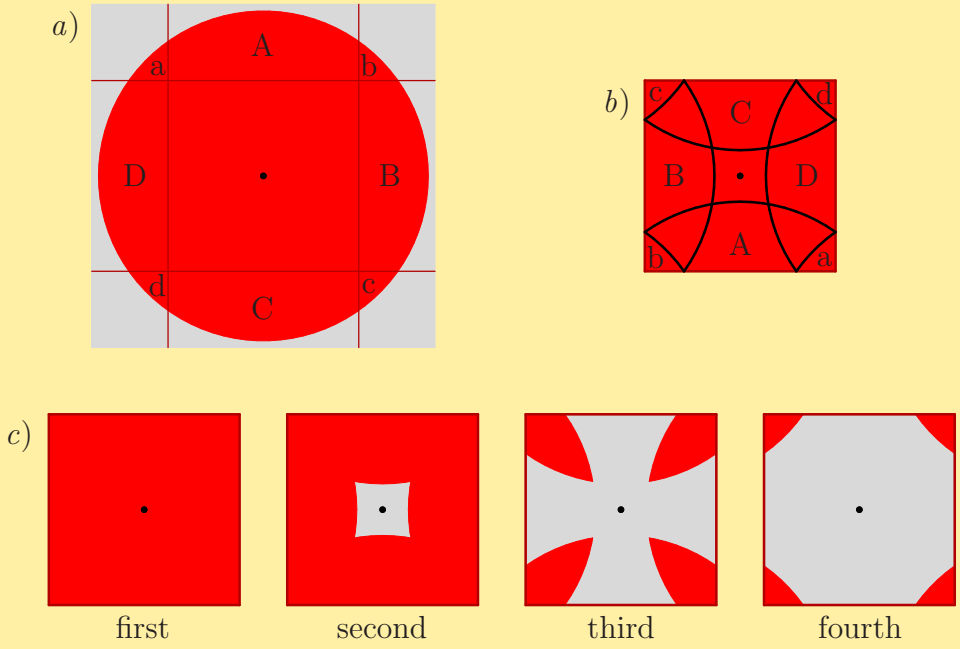


Figure 10.18: Redefinition of the occupied wave number vectors into Brillouin zones.

Now each Floquet wave number vector in the first Brillouin zone does no longer correspond to just one spatial energy eigenfunction like in the extended zone scheme. There will now be multiple spatial eigenfunctions, distinguished by different lattice-scale variations $\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$. Compare that with the earlier approximation of one-dimensional crystals as widely separated atoms. That was in terms of different atomic wave functions like the 2s and 2p ones, not a single one, that were modulated by Floquet exponentials that varied relatively slowly over an atomic cell. In other words, the reduced zone scheme is the natural one for widely spaced atoms: the lattice scale parts $\psi^{\mathrm{p}}_{\mathrm{p},\vec{k}}$ correspond to the different atomic energy eigenfunctions. And since they take care of the nontrivial variations within each lattice cell, the Floquet exponentials become slowly varying ones.

But you might rightly feel that the critical Fermi surface is messed up pretty badly in the reduced zone scheme figure 10.18*b*. That does not seem to be such a hot idea, since the electrons near the Fermi surface are critical for the properties of metals. However, the picture can now be taken apart again to produce separate Brillouin zones. There is a construction credited to Harrison that is illustrated in figure 10.18*c*. For points that are covered by at least one fragment of the original sphere, (which means all points, here,) the first covering is moved into the first Brillouin zone. For points that are covered by at least two

fragments of the original sphere, the second covering is moved into the second Brillouin zone. And so on.

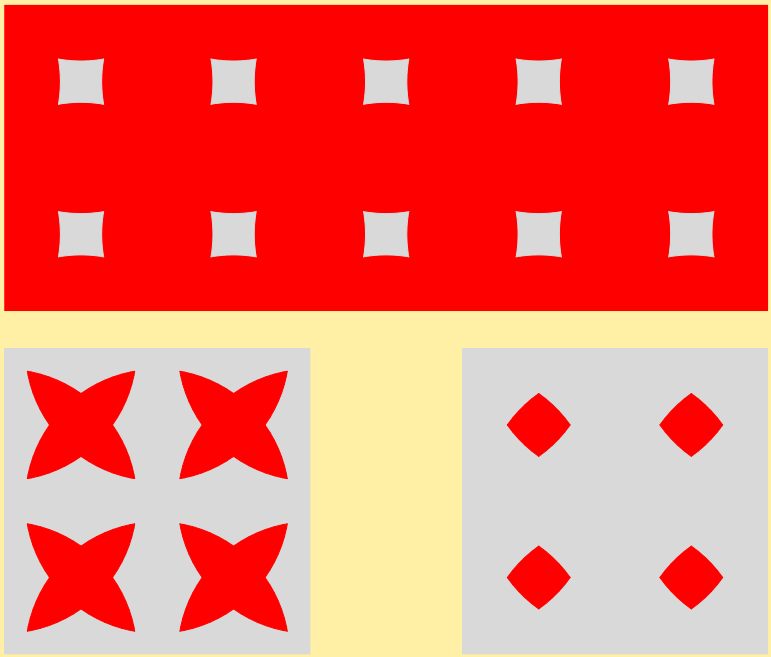


Figure 10.19: Second, third, and fourth Brillouin zones seen in the periodic zone scheme.

Remember that in say electrical conduction, the electrons change occupied states near the Fermi surfaces. To simplify talking about that, physicist like to extend the pictures of the Brillouin zones periodically, as illustrated in figure 10.19. This is called the "periodic zone scheme." In this scheme, the boundaries of the Wigner-Seitz cells, which are normally not Fermi surfaces, are no longer a distracting factor. It may be noted that a bit of a lattice potential will round off the sharp corners in figure 10.19, increasing the esthetics.

10.6 Nearly-Free Electrons

The free-electron energy spectrum does not have bands. Bands only form when some of the forces that the ambient solid exerts on the electrons are included. In this section, some of the mechanics of that process will be explored. The only force considered will be one given by a periodic lattice potential. The discussion will still ignore true electron-electron interactions, time variations of the lattice potential, lattice defects, etcetera.

In addition, to simplify the mathematics it will be assumed that the lattice potential is weak. That makes the approach here diametrically opposite to the one followed in the discussion of the one-dimensional crystals. There the

starting point was electrons tightly bound to widely spaced atoms; the atom energy levels then corresponded to infinitely concentrated bands that fanned out when the distance between the atoms was reduced. Here the starting idea is free electrons in closely packed crystals for which the bands are completely fanned out so that there are no band gaps left. But it will be seen that when a bit of nontrivial lattice potential is added, energy gaps will appear.

The analysis will again be based on the Floquet energy eigenfunctions for the electrons. As noted in the previous section, they correspond to periodic boundary conditions for periods $2\ell_x$, $2\ell_y$, and $2\ell_z$. In case that the energy eigenfunctions for confined electrons are desired, they can be obtained from the Bloch solutions to be derived in this section in the following way: Take a Bloch solution and flip it over around the $x = 0$ plane, i.e. replace x by $-x$. Subtract that from the original solution, and you have a solution that is zero at $x = 0$. And because of periodicity and odd symmetry, it will also be zero at $x = \ell_x$. Repeat these steps in the y and z directions. It will produce energy eigenfunctions for electrons confined to a box $0 < x < \ell_x$, $0 < y < \ell_y$, $0 < z < \ell_z$. This method works as long as the lattice potential has enough symmetry that it does not change during the flip operations.

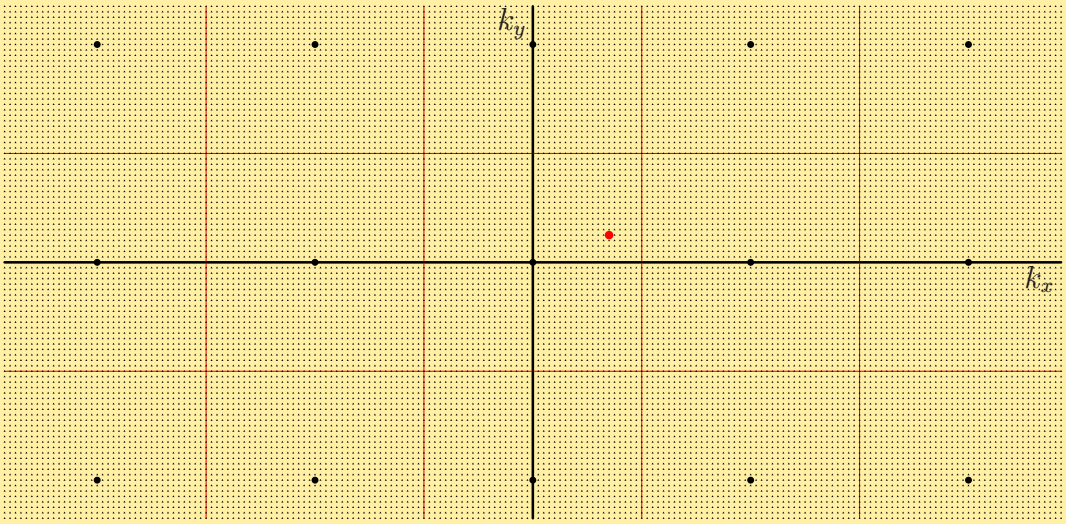


Figure 10.20: The red dot shows the wavenumber vector of a sample free electron wave function. It is to be corrected for the lattice potential.

The approach will be to start with the solutions for force-free electrons and see how they change if a small, but nonzero lattice potential is added to the motion. It will be a "*nearly*-free electron model." Consider a sample Floquet wave number as shown by the red dot in the wave number space figure 10.20. If there is no lattice potential, the corresponding energy eigenfunction is the

free-electron one,

$$\psi^{\rm p}_{\vec{k},0} = \frac{1}{\sqrt{8\ell_x\ell_y\ell_z}} e^{{\rm i}(k_x x + k_y y + k_z z)}$$

where the subscript zero merely indicates that the lattice potential is zero. (This section will use the extended zone scheme because it is mathematically easiest.) If there is a lattice potential, the eigenfunction will change into a Bloch one of the form

$$\psi^{\rm p}_{\vec{k}} = \psi^{\rm p}_{{\rm p},\vec{k}} e^{{\rm i}(k_x x + k_y y + k_z z)}$$

where $\psi^{\rm p}_{{\rm p},\vec{k}}$ is periodic on an atomic scale. If the lattice potential is weak, as assumed here,

$$\psi^{\rm p}_{{\rm p},\vec{k}} \approx \frac{1}{\sqrt{8\ell_x\ell_y\ell_z}}$$

Also, the energy will be almost the free-electron one:

$$E^{\rm e}_{\vec{k}} \approx E^{\rm e}_{\vec{k},0} = \frac{\hbar^2}{2m_{\rm e}} k^2$$

However, that is not good enough. The interest here is in the *changes* in the energy due to the lattice potential, even if they are weak. So the first thing will be to figure out these energy changes.

10.6.1 Energy changes due to a weak lattice potential

Finding the energy changes due to a small change in a Hamiltonian can be done by a mathematical technique called "perturbation theory." A full description and derivation are in {A.38} and {D.79}. This subsection will simply state the needed results.

The effects of a small change in a Hamiltonian, here being the weak lattice potential, are given in terms of the so-called "Hamiltonian perturbation coefficients" defined as

$$H_{\vec{k}\underline{\vec{k}}} \equiv \langle \psi^{\rm p}_{\vec{k},0} | V \psi^{\rm p}_{\underline{\vec{k}},0} \rangle \tag{10.13}$$

where V is the lattice potential, and the $\psi^{\rm p}_{\vec{k},0}$ are the free-electron energy eigenfunctions.

In those terms, the energy of the eigenfunction $\psi_{\vec{k}}$ with Floquet wave number $\vec{k}$ is

$$E^{\rm e}_{\vec{k}} \approx E^{\rm e}_{\vec{k},0} + H_{\vec{k}\vec{k}} - \sum_{\underline{\vec{k}} \neq \vec{k}} \frac{|H_{\vec{k}\underline{\vec{k}}}|^2}{E^{\rm e}_{\underline{\vec{k}},0} - E^{\rm e}_{\vec{k},0}} + \ldots \tag{10.14}$$

Here $E^{\rm e}_{\vec{k},0}$ is the free-electron energy. The dots stand for contributions that can be ignored for sufficiently weak potentials.

The first correction to the free-electron energy is the Hamiltonian perturbation coefficient $H_{\vec{k}\vec{k}}$. However, by writing out the inner product, it is seen

that this perturbation coefficient is just the average lattice potential. Such a constant energy change is of no particular physical interest; it can be eliminated by redefining the zero level of the potential energy.

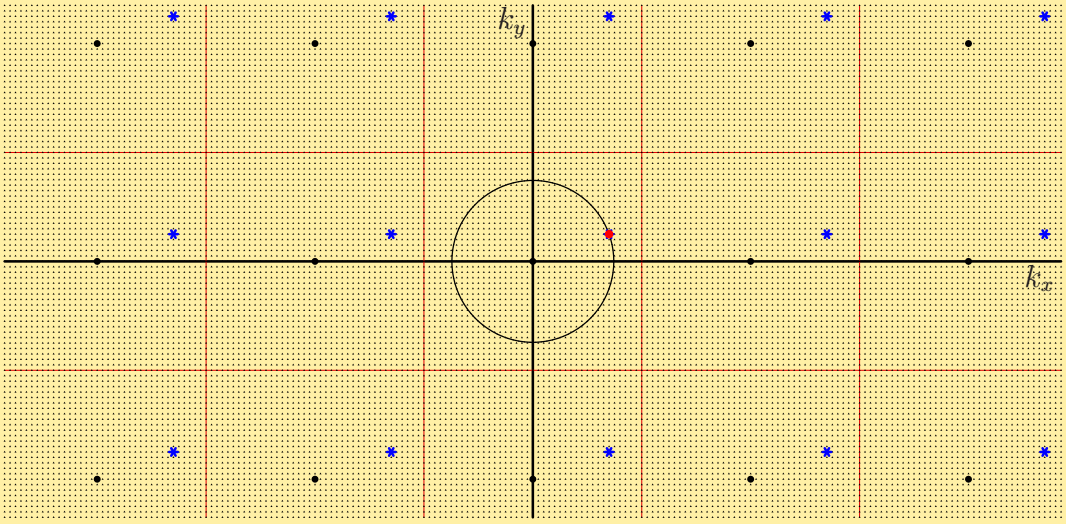


Figure 10.21: The grid of nonzero Hamiltonian perturbation coefficients and the problem sphere in wave number space.

That makes the sum in (10.14) the physically interesting change in energy. Now, unlike it seems from the given expression, it is not really necessary to sum over *all* free-electron energy eigenfunctions $\psi_{\vec{\underline{k}},0}$. The only Hamiltonian perturbation coefficients that are nonzero occur for the $\vec{\underline{k}}$ values shown in figure 10.21 as blue stars. They are spaced apart by amounts J in each direction, where J is the large number of physical lattice cells in that direction. These claims can be verified by writing the lattice potential as a Fourier series and then integrating the inner product. More elegantly, you can use the observation from addendum {A.38.3} that the only eigenfunctions that need to be considered are those with the same eigenvalues under displacement over the primitive vectors of the lattice. (Since the periodic lattice potential is the same after such displacements, these displacement operators commute with the Hamiltonian.)

The correct expression for the energy change has therefore now been identified. There is one caveat in the whole story, though. The above analysis is not justified if there are eigenfunctions $\psi^{\rm p}_{\vec{\underline{k}},0}$ on the grid of blue stars that have the same free-electron energy $E^{\rm e}_{\vec{k},0}$ as the eigenfunction $\psi^{\rm p}_{\vec{k},0}$. You can infer the problem from (10.14); you would be dividing by zero if that happened. You would have to fix the problem by using so-called "singular perturbation theory," which is much more elaborate.

Fortunately, since the grid is so widely spaced, the problem occurs only for relatively few energy eigenfunctions $\psi^{\rm p}_{\vec{k}}$. In particular, since the free-electron

energy $E^{\rm e}_{\vec{k},0}$ equals $\hbar^2k^2/2m_{\rm e}$, the square magnitude of $\underline{\vec{k}}$ would have to be the same as that of $\vec{k}$. In other words, $\underline{\vec{k}}$ would have to be on the same spherical surface around the origin as point $\vec{k}$. So, as long as the grid has no points other than $\vec{k}$ on the spherical surface, all is OK.

10.6.2 Discussion of the energy changes

The previous subsection determined how the energy changes from the free-electron gas values due to a small lattice potential. It was found that an energy level $E^{\rm e}_{\vec{k},0}$ without lattice potential changes due to the lattice potential by an amount:

$$\Delta E^{\rm e}_{\vec{k}} = -\sum_{\underline{\vec{k}}\neq\vec{k}} \frac{|H_{\vec{k}\underline{\vec{k}}}|^2}{E^{\rm e}_{\underline{\vec{k}},0} - E^{\rm e}_{\vec{k},0}} \tag{10.15}$$

where the $H_{\vec{k}\underline{\vec{k}}}$ were coefficients that depend on the details of the lattice potential; $\vec{k}$ was the wave number vector of the considered free-electron gas solution, shown as a red dot in the wavenumber space figure 10.21, $\underline{\vec{k}}$ was an summation index over the blue grid points of that figure, and $E^{\rm e}_{\underline{\vec{k}},0}$ and $E^{\rm e}_{\vec{k},0}$ were proportional to the square distances from the origin to points $\underline{\vec{k}}$, respectively $\vec{k}$. $E^{\rm e}_{\vec{k},0}$ is also the energy level of the eigenfunction without lattice potential.

The expression above for the energy change is not valid when $E^{\rm e}_{\underline{\vec{k}},0} = E^{\rm e}_{\vec{k},0}$, in which case it would incorrectly give infinite change in energy. However, it is does apply when $E^{\rm e}_{\underline{\vec{k}},0} \approx E^{\rm e}_{\vec{k},0}$, in which case it predicts unusually large changes in energy. The condition $E^{\rm e}_{\underline{\vec{k}},0} \approx E^{\rm e}_{\vec{k},0}$ means that a blue star $\underline{\vec{k}}$ on the grid in figure 10.21 is almost the same distance from the origin as the red point $\vec{k}$ itself.

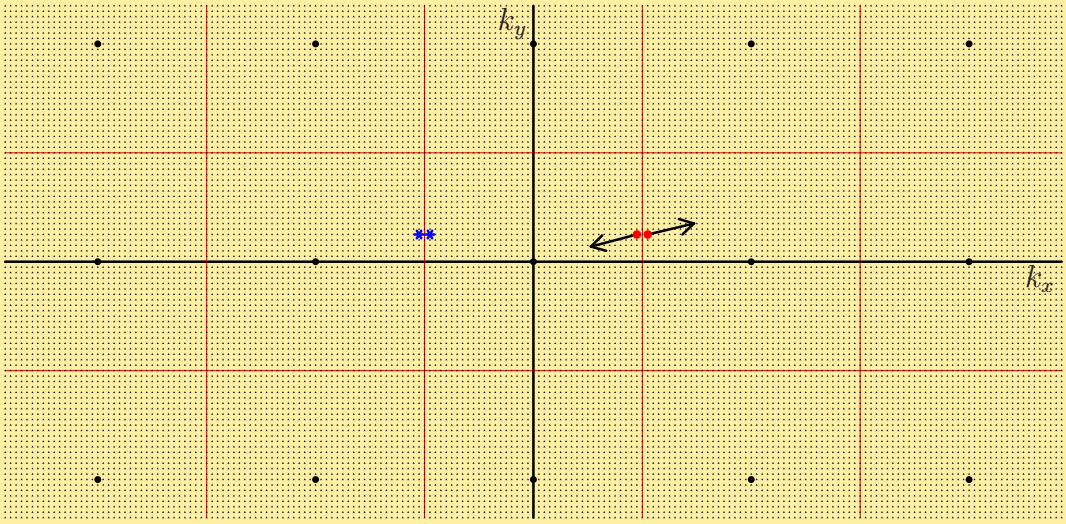


Figure 10.22: Tearing apart of the wave number space energies.

One case for which this happens is when the wave number vector $\vec{k}$ is right next to one of the boundaries of the Wigner-Seitz cell around the origin. Whenever a $\vec{k}$ is on the verge of leaving this cell, one of its lattice points is on the verge of getting in. As an example, figure 10.22 shows two neighboring states $\vec{k}$ straddling the right-hand vertical plane of the cell, as well as their lattice $\underline{\vec{k}}$ values that cause the unusually large energy changes.

For the left of the two states, $E^{\rm e}_{\underline{\vec{k}},0}$ is just a bit larger than $E^{\rm e}_{\vec{k}.0}$, so the energy change (10.15) due to the lattice potential is large and negative. All energy decreases will be represented graphically by moving the points towards the origin, in order that the distance from the origin continues to indicate the energy of the state. That means that the left state will move strongly towards the origin. Consider now the other state just to the right; $E^{\rm e}_{\underline{\vec{k}},0}$ for that state is just a bit less than $E^{\rm e}_{\vec{k},0}$, so the energy change of this state will be large and positive; graphically, this point will move strongly away from the origin. The result is that the energy levels are torn apart along the surface of the Wigner-Seitz cell.

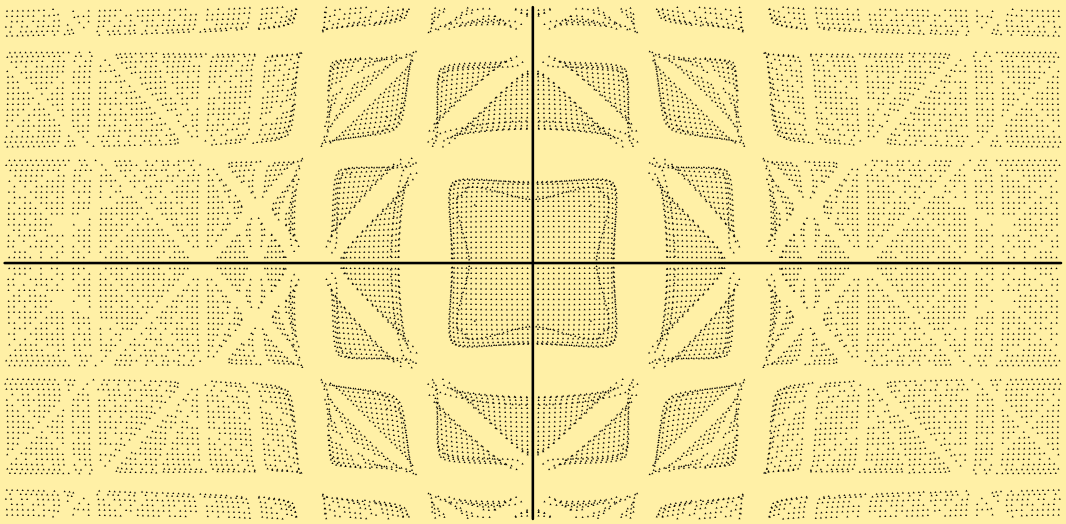


Figure 10.23: Effect of a lattice potential on the energy. The energy is represented by the square distance from the origin, and is relative to the energy at the origin.

That is illustrated for an arbitrarily chosen example lattice potential in figure 10.23. It is another reason why the Wigner-Seitz cell around the origin, i.e. the first Brillouin zone, is particularly important. For different lattices than the simple cubic one considered here, it is still the distance from the origin that is the deciding factor, so in general, it is the Wigner-Seitz cell, rather than some parallelepiped-shaped primitive cell along whose surfaces the energies get torn apart.

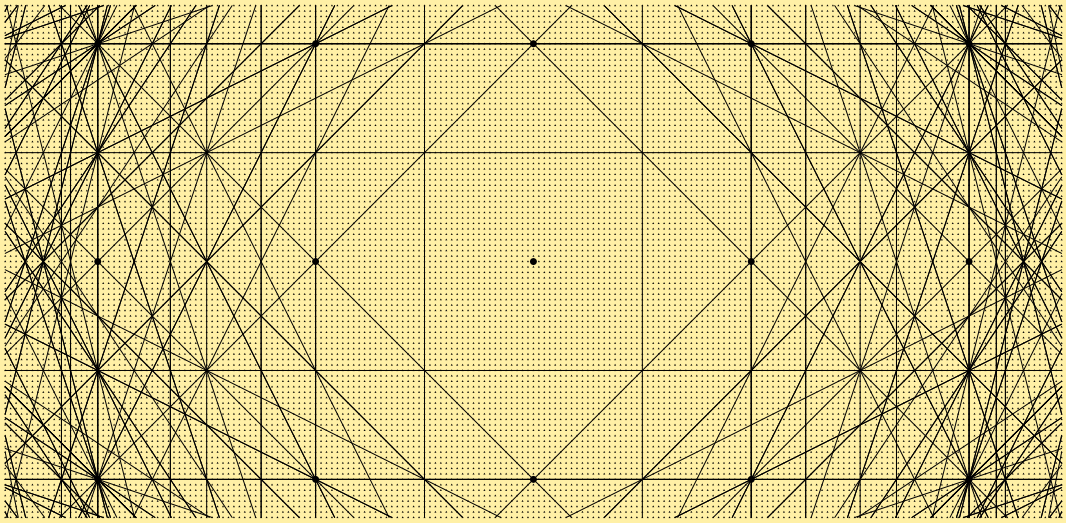


Figure 10.24: Bragg planes seen in wave number space cross section.

But notice in figure 10.23 that the energy levels get torn apart along many more surfaces than just the surface of the first Brillouin zone. In general, it can be seen that tears occur in wave number space along all the perpendicular bisector planes, or Bragg planes, between the points of the reciprocal lattice and the origin. Figure 10.24 shows their intersections with the cross section $k_z = 0$ as thin black lines. The k_x and k_y axes were left away to clarify that they do not hide any lines.

Recall that the Bragg planes are also the boundaries of the fragments that make up the various Brillouin zones. In fact the first Brillouin zone is the cube or Wigner-Seitz cell around the origin; (the square around the origin in the cross section figure 10.24). The second zone consists of six pyramid-shaped regions whose bases are the faces of the cube; (the four triangles sharing a side with the square in the cross section figure 10.24). They can be pushed into the first Brillouin zone using the fundamental translation vectors to combine into a Wigner-Seitz cell shape.

For a sufficiently strong lattice potential like the one in figure 10.23, the energy levels in the first Brillouin zone, the center patch, are everywhere lower than in the remaining areas. Electrons will then occupy these states first, and since there are $J \times J \times J$ spatial states in the zone, two valence electrons per physical lattice cell will just fill it, figure 10.25. That produces an insulator whose electrons are stuck in a filled valence band. The electrons must jump an finite energy gap to reach the outlying regions if they want to do anything nontrivial. Since no particular requirements were put onto the lattice potential, the forming of bands is self-evidently a very general process.

The wave number space in the right half of figure 10.25 also illustrates that a lattice potential can change the Floquet wave number vectors that get occupied.

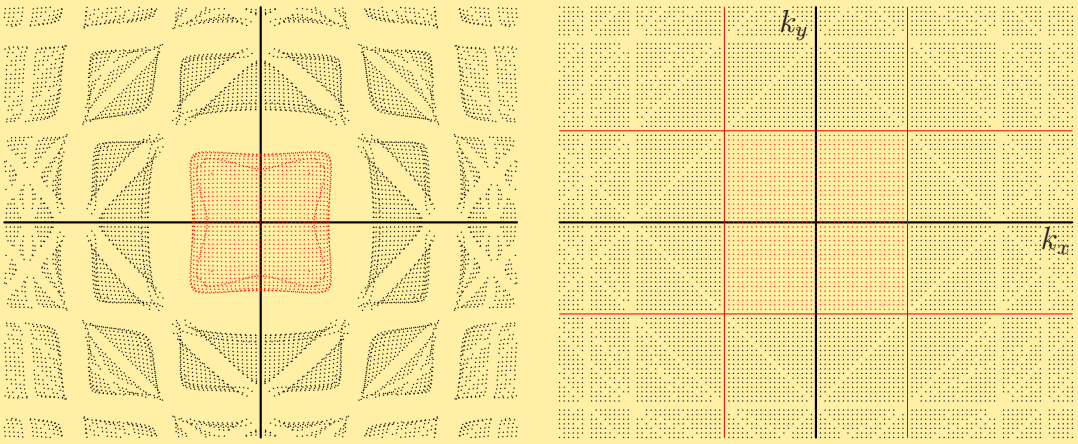


Figure 10.25: Occupied states for the energies of figure 10.23 if there are two valence electrons per lattice cell. Left: energy. Right: wave numbers.

For the free-electron gas, the occupied states formed a spherical region in terms of the wave number vectors, as shown in the middle of figure 10.17, but here the occupied states have become a cube, the Wigner-Seitz cell around the origin. The Fermi surface seen in the extended zone scheme is now no longer a spherical surface, but consists of the six faces of this cell.

But do not take this example too literally: the small-perturbation analysis is invalid for the strong potential required for an insulator, and the real picture would look quite different. In particular, the "roll-over" of the states at the edge of the first Brillouin zone in the energy plot is a clear indication that the accuracy is poor. The error in the perturbation analysis is the largest for states immediately next to the Bragg planes. The example is given just to illustrate that the nearly-free electron model can indeed describe band gaps if taken far enough.

The nearly-free electron model is more reasonable for the smaller lattice forces experienced by valence electrons in metals. For example, at reduced strength, the same potential as before produces figure 10.26. Now the electrons have no trouble finding states of slightly higher energy, as it should be for a metal. Note, incidentally, that the Fermi surfaces in the right-hand graphs seem to meet the Bragg planes much more normally than the spherical free-electron surface. That leads to smoothing out of the corners of the surface seen in the periodic zone scheme. For example, imagine the center zone of the one valence electron wave number space periodically continued.

10.7 Additional Points

This section mentions a couple of additional very basic issues in the quantum mechanics of solids.

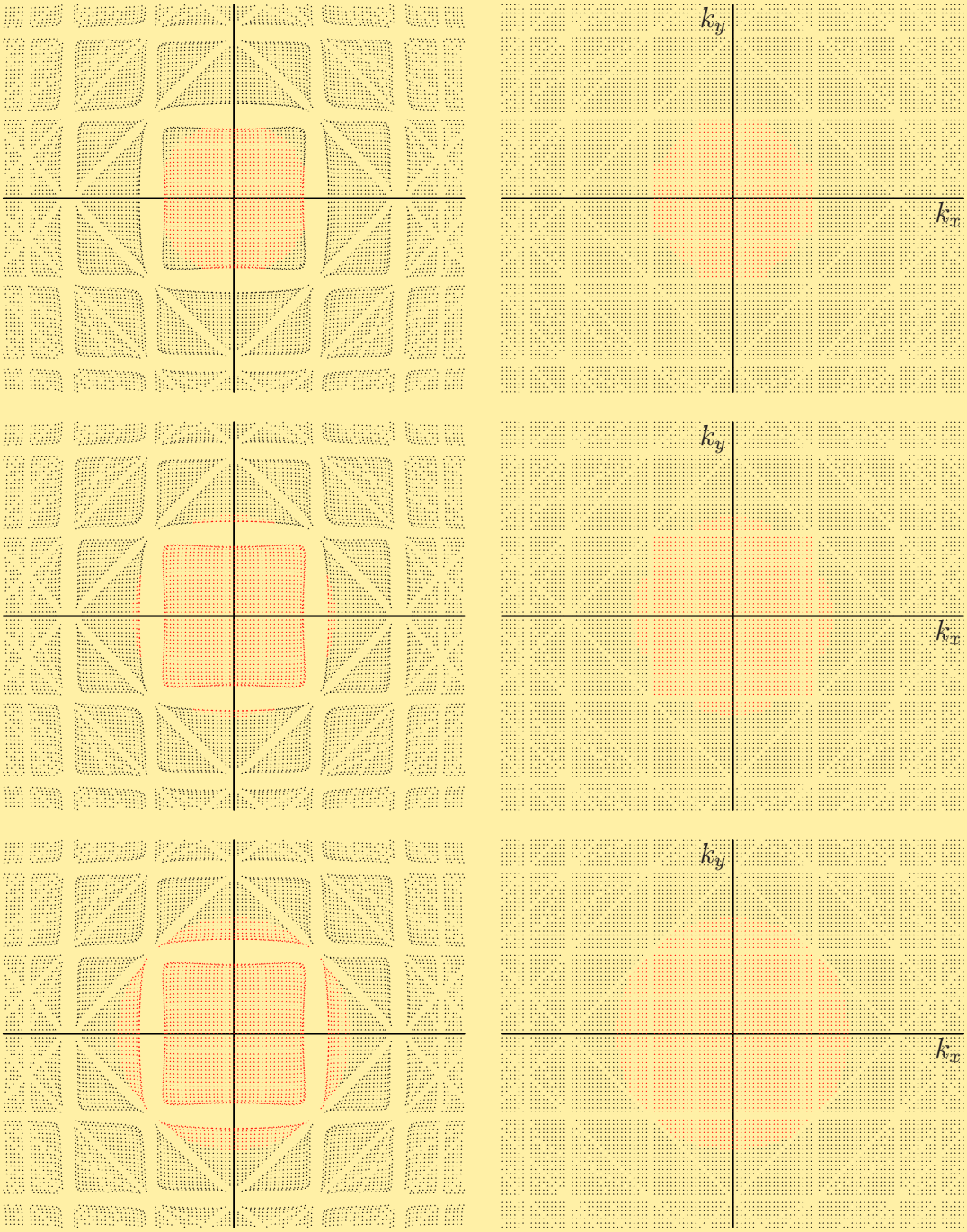


Figure 10.26: Smaller lattice potential. From top to bottom shows one, two and three valence electrons per lattice cell. Left: energy. Right: wave numbers.

10.7.1 About ferromagnetism

Magnetism in all its myriad forms and complexity is far beyond the scope of this book. But there is one very important fundamental quantum mechanics issue associated with ferromagnetism that has not yet been introduced.

Ferromagnetism is the plain variety of magnetism, like in refrigerator magnets. Ferromagnetic solids like iron are of great engineering interest. They can significantly increase a magnetic field and can stay permanently magnetized even in the absence of a field. The fundamental quantum mechanics issue has to do with why they produce magnetic fields in the first place.

The source of the ferromagnetic field is the electrons. Electrons have spin, and just like a classical charged particle that is spinning around in a circle produces a magnetic field, so do electrons act as little magnets. A free iron atom has 26 electrons, each with spin $\frac{1}{2}$. But two of these electrons are in the 1s states, the K shell, where they combine into a singlet state with zero net spin which produces no magnetic field. Nor do the two 2s electrons and the six 2p electrons in the L shell, and the two 3s electrons and six 3p electrons in the M shell and the two 4s electrons in the N shell produce net spin. All of that lack of net spin is a result of the Pauli exclusion principle, which says that if electrons want to go two at a time into the lowest available energy states, they must do it as singlet spin states. And these filled subshells produce no net orbital angular momentum either, having just as many positive as negative orbital momentum states filled in whatever way you look at it.

However, iron has a final six electrons in 3d states, and the 3d states can accommodate ten electrons, five for each spin direction. So only two out of the six electrons need to enter the same spatial state as a zero spin singlet. The other four electrons can each go into their private spatial state. And the electrons do want to do so, since by going into different spatial states, they can stay farther away from each other, minimizing their mutual Coulomb repulsion energy.

According to the simplistic model of noninteracting electrons that was used to describe atoms in chapter 5.9, these last four electrons can then have equal or opposite spin, whatever they like. But that is wrong. The four electrons interact through their Coulomb repulsion, and it turns out that they achieve the smallest energy when their spatial wave function is antisymmetric under particle exchange.

(This is just the opposite of the conclusion for the hydrogen molecule, where the symmetric spatial wave function had the lowest energy. The difference is that for the hydrogen molecule, the dominant effect is the reduction of the kinetic energy that the symmetric state achieves, while for the single-atom states, the dominant effect is the reduction in electron to electron Coulomb repulsion that the antisymmetric wave function achieves. In the antisymmetric spatial wave function, the electrons stay further apart on average.)

If the spatial wave function of the four electrons takes care of the antisymmetrization requirement, then their spin state cannot change under particle exchange; they all must have the same spin. This is known as "Hund's first rule:" electron interaction makes the net spin as big as the exclusion principle allows. The four unpaired 3d electrons in iron minimize their Coulomb energy at the price of having to align all four of their spins. Which means their spin magnetic moments add up rather than cancel each other. {A.34}.

Hund's second rule says that the electrons will next maximize their orbital angular momentum as much as is still possible. And according to Hund's third rule, this orbital angular momentum will add to the spin angular momentum since the ten 3d states are more than half full. It turns out that iron's 3d electrons have the same amount of orbital angular momentum as spin, however, orbital angular momentum is only about half as effective at creating a magnetic dipole.

In addition, the magnetic properties of orbital angular momentum are readily messed up when atoms are brought together in a solid, and more so for transition metals like iron than for the lanthanoid series, whose unfilled 4f states are buried much deeper inside the atoms. In most of the common ferromagnets, the orbital contribution is negligible small, though in some rare earths there is an appreciable orbital contribution.

Guessing just the right amounts of net spin angular momentum, net orbital angular momentum, and net combined angular momentum for an atom can be tricky. So, in an effort make quantum mechanics as readily accessible as possible, physicists provide the data in an intuitive hieroglyph. For example

$$^5D_4$$

gives the angular momentum of the iron atom. The 5 indicates that the spin angular momentum is 2. To arrive at 5, the physicists multiply by 2, since spin can be half integer and it is believed that many people doing quantum mechanics have difficulty with fractions. Next 1 is added to keep people from cheating and mentally dividing by 2 – you must subtract 1 first. (Another quick way of getting the actual spin: write down all possible values for the spin in increasing order, and then count until the fifth value. Start counting from 1, of course, because counting from 0 is so computer science.) The D intimates that the orbital angular momentum is 2. To arrive at D, physicists write down the intuitive sequence of letters $S, P, D, F, G, H, I, K, \ldots$ and then count, starting from zero, to the orbital angular momentum. Unlike for spin, here it is not the count, but the object being counted that is listed in the hieroglyph; unfortunately the object being counted is letters, not angular momentum. Physicists assume that after having practiced counting spin states and letters, your memory is refreshed about fractions, and the combined angular momentum is simply listed by value, 4 for iron. Listing spin and combined angular momentum in two different formats achieves that the class won't notice the error if the physics

professor misstates the spin or combined angular momentum for an atom with zero orbital momentum.

On to the solid. The atoms act as little magnets because of their four aligned electron spins and net orbital angular momentum, but why would different atoms want to align their magnetic poles in the same direction in a solid? If they don't, there is not going to be any macroscopically significant magnetic field. The logical reason for the electron spins of different atoms to align would seem to be that it minimizes the magnetic energy. However, if the numbers are examined, any such aligning force is far too small to survive random heat motion at normal temperatures.

The primary reason is without doubt again the same weird quantum mechanics as for the single atom. Nature does not care about magnetic alignment or not; it is squirming to minimize its *Coulomb* energy under the massive constraints of the antisymmetrization requirement. By aligning electron spins globally, it achieves that electrons can stay farther apart spatially. {N.22}.

It is a fairly small effect; among the pure elements, it really only works under normal operating temperatures for cobalt and its immediate neighbors in the periodic table, iron and nickel. And alignment is normally not achieved throughout a bulk solid, but only in microscopic zones, with different zones having different alignment. But any electrical engineer will tell you it is a very important effect anyway. For one since the zones can be manipulated with a magnetic field.

And it clarifies that nature does not necessarily select singlet states of opposite spin to minimize the energy, despite what the hydrogen molecule and helium atom might suggest. Much of the time, aligned spins are preferred.

10.7.2 X-ray diffraction

You may wonder how so much is known about the crystal structure of solids in view of the fact that the atoms are much too small to be seen with visible light. In addition, because of the fact that the energy levels get smeared out into bands, like in figure 10.11, solids do not have those tell-tale line spectra that are so useful for analyzing atoms and molecules.

To be precise, while the energy levels of the outer electrons of the atoms get smeared out, those of the inner electrons do not do so significantly, and these do produce line spectra. But since the energy levels of the inner electrons are very high, transitions involving inner electrons do not produce visible light, but X-rays.

There is a very powerful other technique for studying the crystal structure of atoms, however, and it also involves X-rays. In this technique, called X-ray diffraction, an X-ray is trained on a crystal from various angles, and the way the crystal scatters the X-ray is determined.

There is no quantum mechanics needed to describe how this works, but a brief description may be of value anyway. If you want to work in nanotechnology, you will inevitably run up against experimental work, and X-ray diffraction is a key technique. Having some idea of how it works and what it can do can be useful.

First a very basic understanding is needed of what is an X-ray. An X-ray is a propagating wave of electromagnetic radiation just like a beam of visible light. The only difference between them is that an X-ray is much more energetic. Whether it is light or an X-ray, an electromagnetic wave is physically a combination of electric and magnetic fields that propagate in a given direction with the speed of light.

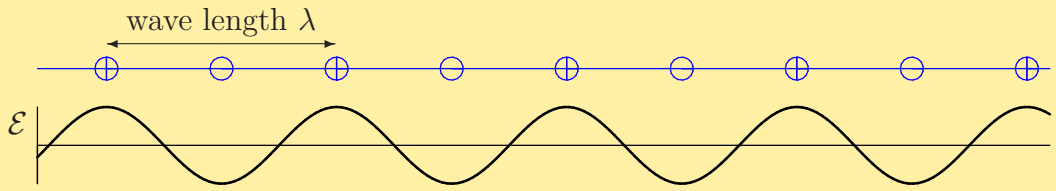


Figure 10.27: Depiction of an electromagnetic ray.

Figure 10.27 gives a sketch of how the strength of the electric field varies along the propagation direction of a simple monochromatic wave; the magnetic field is similar, but 90 degrees out of phase. Above that, a sketch is given how such rays will be visualized in this subsection: the positive maxima will be indicated by encircled plus signs, and the negative minima by encircled minus signs. Both these maxima and minima propagate along the line with the speed of light; the picture is just a snapshot at an arbitrary time.

The distance between two successive maxima is called the wave length λ. If the wave length is in the narrow range from about 4 000 to 7 000 Å, it is visible light. But such a wave length is much too large to distinguish atoms, since atom sizes are in the order of a few Å. Electromagnetic waves with the required wave lengths of a few Å fall in what is called the X-ray range.

The wave number κ is the reciprocal of the wave length within a normalization factor 2π: $\kappa = 2\pi/\lambda$. The wave number vector $\vec{\kappa}$ has the magnitude of the wave number κ and points in the direction of propagation of the wave.

Next consider a plane of atoms in a crystal, and imagine that it forms a perfectly flat mirror, as in figure 10.28. No, there are no physical examples of flat atoms known to science. But just imagine there would be, OK? Now shine an X-ray from the left onto this crystal layer and examine the diffracted wave that comes back from it. Assume Huygens' principle that the scattered rays come off in all directions, and that the scattering is elastic, meaning that the energy, hence wave length, stays the same.

Under those conditions, a detector A, placed at a position to catch the rays scattered to the same angle as the angle θ of the incident beam, will observe a

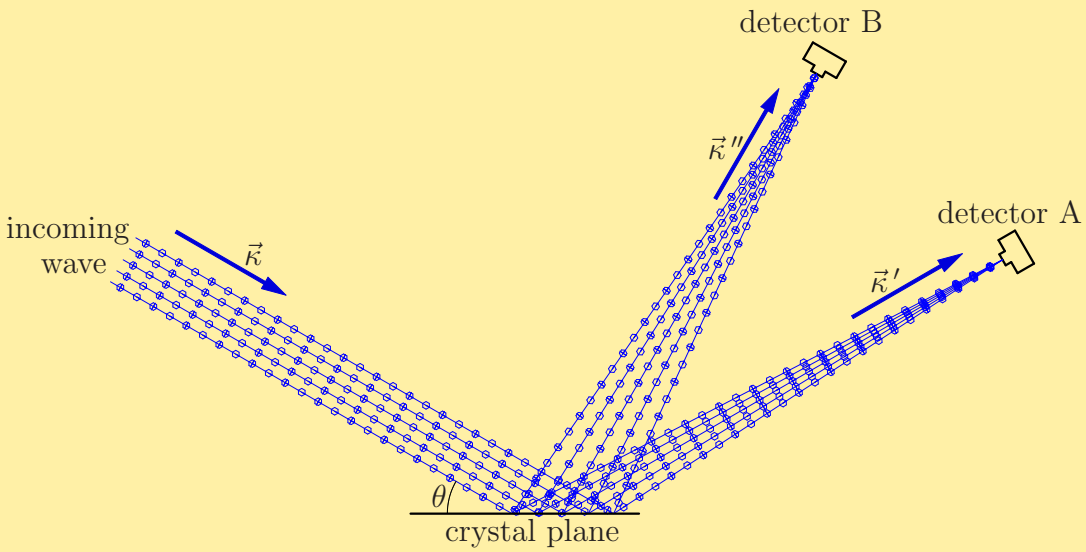


Figure 10.28: Law of reflection in elastic scattering from a plane.

strong signal. All the maxima in the electric field of the rays arrive at detector A at the same time, reinforcing each other. They march in lock-step. So a strong positive signal will exist at detector A at their arrival. Similarly, the minima march in lock-step, arriving at A at the same time and producing a strong signal, now negative. Detector A will record a strong, fluctuating, electric field.

Detector B, at a position where the angle of reflection is unequal to the angle of incidence, receives similar rays, but both positive and negative values of the electric field arrive at B at the same time, killing each other off. So detector B will not see an observable signal. That is the law of reflection: there is only a detectable diffracted wave at a position where the angle of reflection equals the angle of incidence. (Those angles are usually measured from the normal to the surface instead of from the surface itself, but not in Bragg diffraction.)

For visible light, this is actually a quite reasonable analysis of a mirror, since an atom-size surface roughness is negligible compared to the wave length of visible light. For X-rays, it is not so hot, partly because a layer of atoms is not flat on the scale of the wave length of the X-ray. But worse, a single layer of atoms does not reflect an X-ray by any appreciable amount. That is the entire point of medical X-rays; they can penetrate millions of layers of atoms to show what is below. A single layer is nothing to them.

For X-rays to be diffracted in an appreciable amount, it must be done by many parallel layers of atoms, not just one, as in figure 10.29. The layers must furthermore have a very specific spacing d for the maxima and minima from different layers to arrive at the detector at the same time. Note that the angular position of the detector is already determined by the law of reflection, in order to get whatever little there can be gotten from each plane separately.

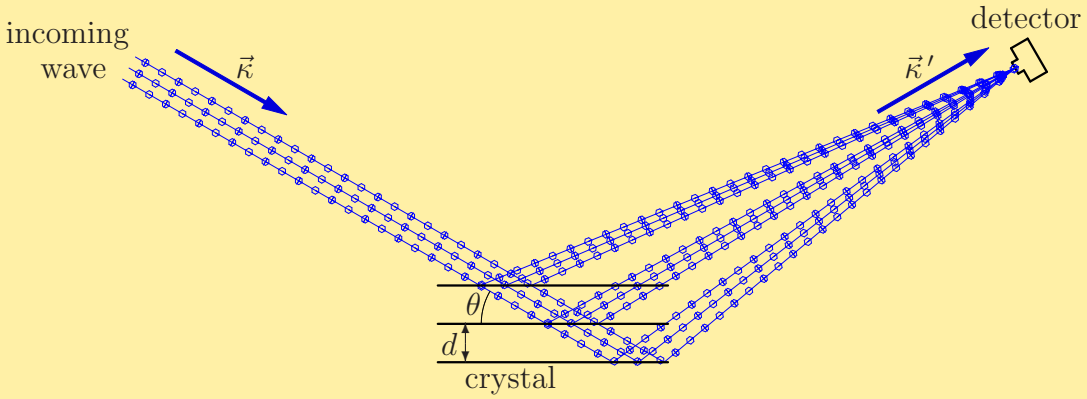


Figure 10.29: Scattering from multiple "planes of atoms."

(Also note that whatever variations in phase there are in the signals arriving at the detector in figure 10.29 are artifacts: for graphical reasons the detector is much closer to the specimen than it should be. The spacing between planes should be on the order of Å, while the detector should be a macroscopic distance away from the specimen.)

The spacing between planes needed to get a decent combined signal strength at the detector is known to satisfy the Bragg law:

$$\boxed{2d\sin\theta = n\lambda} \tag{10.16}$$

where n is a natural number. A derivation will be given below. One immediate consequence is that to get X-ray diffraction, the wave length λ of the X-ray cannot be more than twice the spacing between the planes of atoms. That requires wave lengths no longer than of the order of Ångstroms. Visible light does not qualify.

The above story is, of course, not very satisfactory. For one, layers of atoms are not flat planes on the scale of the required X-ray wave lengths. And how come that in one direction the atoms have continuous positions and in another discrete? Furthermore, it is not obvious what to make of the results. Observing a refracted X-ray at some angular location may suggest that there is some reflecting plane in the crystal at an angle deducible from the law of reflection, but many different planes of atoms exist in a crystal. If a large number of measurements are done, typically by surrounding the specimen by detectors and rotating it while shining an X-ray on it, how is the crystal structure to be deduced from that overwhelming amount of information?

Clearly, a mathematical analysis is needed, and actually it is not very complicated. First a mathematical expression is needed for the signal along the ray; it can be taken to be a complex exponential

$$e^{\mathrm{i}\kappa(s-ct)},$$

where s is the distance traveled along the ray from a suitable chosen starting position, t the time, and c the speed of light. The real part of the exponential can be taken as the electric field, with a suitable constant, and the imaginary part as the magnetic field, with another constant. The only important point here is that if there is a difference in travel distance Δs between two rays, their signals at the detector will be out of phase by a factor $e^{\mathrm{i}\kappa\Delta s}$. Unless this factor is one, which requires $\kappa\Delta s$ to be zero or a whole multiple of 2π, there will be at least some cancelation of signals at the detector.

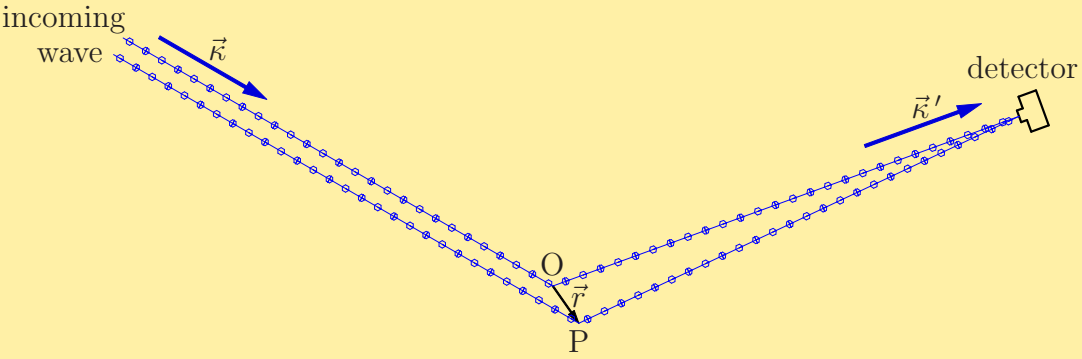


Figure 10.30: Difference in travel distance when scattered from P rather than O.

So, how much is the phase factor $e^{\mathrm{i}\kappa\Delta s}$? Figure 10.30 shows one ray that is scattered at a chosen reference point O in the crystal, and another ray that is scattered at another point P. The position vector of P relative to origin O is $\vec{r}$. Now the difference in travel distance for the second ray to reach P versus the first one to reach O is given by the component of vector $\vec{r}$ in the direction of the incoming wave vector $\vec{\kappa}$. This component can be found as a dot product with the unit vector in the direction of $\vec{\kappa}$:

$$\Delta s_1 = \vec{r} \cdot \frac{\vec{\kappa}}{\kappa} \qquad \text{so} \qquad e^{\mathrm{i}\kappa\Delta s_1} = e^{\mathrm{i}\vec{\kappa}\cdot\vec{r}}.$$

The difference in travel distance for the second ray to reach the detector from point P versus the first from O is similarly given as

$$\Delta s_2 = -\vec{r} \cdot \frac{\vec{\kappa}'}{\kappa} \qquad \text{so} \qquad e^{\mathrm{i}\kappa\Delta s_2} = e^{-\mathrm{i}\vec{\kappa}'\cdot\vec{r}}$$

assuming that the detector is sufficiently far away from the crystal that the rays can be assumed to travel to the detector in parallel.

The net result is then that the phase factor with which the ray from P arrives at the detector compared to the ray from O is

$$e^{\mathrm{i}(\vec{\kappa}-\vec{\kappa}')\cdot\vec{r}}.$$

This result may be used to check the law of reflection and Bragg's law above.

First of all, for the law of reflection of figure 10.28, the positions of the scattering points P vary continuously through the horizontal plane. That means that the phase factor of the rays received at the detector will normally also vary continuously from positive to negative back to positive etcetera, leading to large-scale cancelation of the net signal. The one exception is when $\vec{\kappa}-\vec{\kappa}'$ happens to be normal to the reflecting plane, since a dot product with a normal vector is always zero. For $\vec{\kappa}-\vec{\kappa}'$ to be normal to the plane, its horizontal component must be zero, meaning that the horizontal components of $\vec{\kappa}$ and $\vec{\kappa}'$ must be equal, and for that to be true, their angles with the horizontal plane must be equal, since the vectors have the same length. So the law of reflection is obtained.

Next for Bragg's law of figure 10.29, the issue is the phase difference between successive crystal planes. So the vector $\vec{r}$ in this case can be assumed to point from one crystal plane to the next. Since from the law of reflection, it is already known that $\vec{\kappa}-\vec{\kappa}'$ is normal to the planes, the only component of $\vec{r}$ of importance is the vertical one, and that is the crystal plane spacing d. It must be multiplied by the vertical component of $\vec{\kappa}-\vec{\kappa}'$, (its only component), which is according to basic trig is equal to $-2\kappa\sin\theta$. The phase factor between successive planes is therefore $e^{-\mathrm{i}d2\kappa\sin\theta}$. The argument of the exponential is obviously negative, and then the only possibility for the phase factor to be one is if the argument is a whole multiple n times $-\mathrm{i}2\pi$. So for signals from different crystal planes to arrive at the detector in phase,

$$d2\kappa\sin\theta = n2\pi.$$

Substitute $\kappa = 2\pi/\lambda$ and you have Bragg's law.

Now how about diffraction from a *real* crystal? Well, assume that every location in the crystal elastically scatters the incoming wave by a small amount that is proportional to the electron density n at that point. (This n not to be confused with the n in Bragg's law.) Then the total signal D received by the detector can be written as

$$D = C\int_{\text{all }\vec{r}} n(\vec{r})e^{\mathrm{i}(\vec{\kappa}-\vec{\kappa}')\cdot\vec{r}}\,\mathrm{d}^3\vec{r}$$

where C is some constant. Now the electron density is periodic on crystal lattice scale, so according to section 10.3.10 it can be written as a Fourier series, giving the signal as

$$D = C\sum_{\text{all }\vec{k}_{\vec{n}}}\int_{\text{all }\vec{r}} n_{\vec{k}_{\vec{n}}}e^{\mathrm{i}(\vec{k}_{\vec{n}}+\vec{\kappa}-\vec{\kappa}')\cdot\vec{r}}\,\mathrm{d}^3\vec{r}$$

where the $\vec{k}_{\vec{n}}$ wave number vectors form the reciprocal lattice and the numbers $n_{\vec{k}_{\vec{n}}}$ are constants. Because the volume integration above extends over countless lattice cells, there will be massive cancelation of signal unless the exponential is

constant, which requires that the factor multiplying the position coordinate is zero:

$$\boxed{\vec{k}_{\vec{n}} = \vec{\kappa}' - \vec{\kappa}} \tag{10.17}$$

So the changes in the x-ray wave number vector $\vec{\kappa}$ for which there is a detectable signal tell you the reciprocal lattice vectors. (Or at least the ones for which $n_{\vec{k}_{\vec{n}}}$ is not zero because of some symmetry.) After you infer the reciprocal lattice vectors it is easy to figure out the primitive vectors of the physical crystal you are analyzing. Furthermore, the relative strength of the received signal tells you the magnitude of the Fourier coefficient $n_{\vec{k}_{\vec{n}}}$ of the electron density. Obviously, all of this is very specific and powerful information, far above trying to make some sense out of mere collections of flat planes and their spacings.

One interesting additional issue has to do with what incoming wave vectors $\vec{\kappa}$ are diffracted, regardless of where the diffracted wave ends up. To answer it, just eliminate $\vec{\kappa}'$ from the above equation by finding its square and noting that $\vec{\kappa}' \cdot \vec{\kappa}'$ is κ^2 since the magnitude of the wave number does not change in elastic scattering. It produces

$$\boxed{\vec{\kappa} \cdot \vec{k}_{\vec{n}} = -\tfrac{1}{2}\vec{k}_{\vec{n}} \cdot \vec{k}_{\vec{n}}} \tag{10.18}$$

For this equation to be satisfied, the X-ray wave number vector $\vec{\kappa}$ must be in the Bragg plane between $-\vec{k}_{\vec{n}}$ and the origin. For example, for a simple cubic crystal, $\vec{\kappa}$ must be in one of the Bragg planes shown in cross section in figure 10.24. One general consequence is that the wave number vector κ must at least be long enough to reach the surface of the first Brillouin zone for any Bragg diffraction to occur. That determines the maximum wave length of usable X-rays according to $\lambda = 2\pi/\kappa$. You may recall that the Bragg planes are also the surfaces of the Brillouin zone segments and the surfaces along which the electron energy states develop discontinuities if there is a lattice potential. They sure get around.

Historically, Bragg diffraction was important to show that particles are indeed associated with wave functions, as de Broglie had surmised. When Davisson and Germer bombarded a crystal with a beam of single-momentum electrons, they observed Bragg diffraction just like for electromagnetic waves. Assuming for simplicity that the momentum of the electrons is in the z-direction and that uncertainty in momentum can be ignored, the eigenfunctions of the momentum operator $\widehat{p}_z = \hbar\partial/\mathrm{i}\partial z$ are proportional to $e^{\mathrm{i}\kappa z}$, where $\hbar\kappa$ is the z-momentum eigenvalue. From the known momentum of the electrons, Davisson and Germer could compute the wave number κ and verify that the electrons suffered Bragg diffraction according to that wave number. (The value of $\hbar$ was already known from Planck's blackbody spectrum, and from the Planck-Einstein relation that the energy of the photons of electromagnetic radiation equals $\hbar\omega$ with ω the angular frequency.)

Chapter 11

Basic and Quantum Thermodynamics

Chapter 6 mentioned the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein energy distributions of systems of weakly interacting particles. This chapter explains these results and then goes on to put quantum mechanics and thermodynamics in context.

It is assumed that you have had a course in basic thermodynamics. If not, rejoice, you are going to get one now. The exposition depends relatively strongly upon the material in chapter 5.7–5.9 and chapter 6.1–6.16.

This chapter will be restricted to systems of particles that are all the same. Such a system is called a "pure substance." Water would be a pure substance, but air not really; air is mostly nitrogen, but the 20% oxygen can probably not be ignored. That would be particularly important under cryogenic conditions in which the oxygen condenses out first.

The primary quantum system to be studied in detail will be a macroscopic number of weakly interacting particles, especially particles in a box. Nontrivial interactions between even a few particles are very hard to account for correctly, and for a macroscopic system, that becomes much more so: just a millimol has well over 10^{20} particles. By ignoring particle interactions, the system can be described in terms of single-particle energy eigenstates, allowing some real analysis to be done.

However, a system of strictly noninteracting unperturbed particles would be stuck into the initial energy eigenstate, or the initial combination of such states, according to the Schrödinger equation. To get such a system to settle down into a physically realistic configuration, it is necessary to include the effects of the unavoidable real life perturbations, (molecular motion of the containing box, ambient electromagnetic field, cosmic rays, whatever.) The effects of such small random perturbations will be accounted for using reasonable assumptions. In particular, it will be assumed that they tend to randomly stir up things a bit over time, taking the system out of any physically unlikely state it may be

stuck in and making it settle down into the macroscopically stable one, called "thermal equilibrium."

11.1 Temperature

This book frequently uses the word "temperature," but what does that really mean? It is often said that temperature is some measure of the kinetic energy of the molecules, but that is a dubious statement. It is OK for a thin noble gas, where the kinetic energy per atom is $\frac{3}{2}k_{\rm B}T$ with $k_{\rm B} = 1.380\,65\ 10^{-23}$ J/K the Boltzmann constant and T the (absolute) temperature in degrees Kelvin. But the valence electrons in an metal typically have kinetic energies many times greater than $\frac{3}{2}k_{\rm B}T$. And when the absolute temperature becomes zero, the kinetic energy of a system of particles does not normally become zero, since the uncertainty principle does not allow that.

In reality, the temperature of a system is not a measure of its thermal kinetic energy, but of its "hotness." So, to understand temperature, you first have to understand hotness. A system A is hotter than a system B, (and B is colder than A,) if heat energy flows from A to B if they are brought into thermal contact. If no heat flows, A and B are equally hot. Temperature is a numerical value defined so that, if two systems A and B are equally hot, they have the same value for the temperature.

The so-called "zeroth law of thermodynamics" ensures that this definition makes sense. It says that if systems A and B have the same temperature, and systems B and C have the same temperature, then systems A and C have the same temperature. Otherwise system B would have two temperatures: A and C would have different temperatures, and B would have the same temperature as each of them.

The systems are supposed to be in thermal equilibrium. For example, a solid chunk of matter that is hotter on its inside than its outside simply does not have a (single) temperature, so there is no point in talking about it.

The requirement that systems that are equally hot must have the *same* value of the temperature does not say anything about *what* that value must be. Definitions of the actual values have historically varied. A good one is to compute the temperature of a system A using an ideal gas B at equal temperature as system A. Then $\frac{3}{2}k_{\rm B}T$ can simply be *defined* to be the mean translational kinetic energy of the molecules of ideal gas B. That kinetic energy, in turn, can be computed from the pressure and density of the gas. With this definition of the temperature scale, the temperature is zero in the ground state of ideal gas B. The reason is that a highly accurate ideal gas means very few atoms or molecules in a very roomy box. With the vast uncertainty in position that the roomy box provides to the ground-state, the uncertainty-demanded kinetic energy is vanishingly small. So $k_{\rm B}T$ will be zero.

It then follows that *all* ground states are at absolute zero temperature, regardless how large their kinetic energy. The reason is that all ground states must have the same temperature: if two systems in their ground states are brought in thermal contact, no heat can flow: neither ground state can sacrifice any more energy, the ground state energy cannot be reduced.

However, the "ideal gas thermometer" is limited by the fact that the temperatures it can describe must be positive. There are some unstable systems that in a technical and approximate, but meaningful, sense have *negative* absolute temperatures [4]. Unlike what you might expect, (aren't negative numbers less than positive ones?) such systems are *hotter* than any normal system. Systems of negative temperature will give off heat regardless of how searingly hot the normal system that they are in contact with is.

In this chapter a definition of temperature scale will be given based on the quantum treatment. Various equivalent definitions will pop up. Eventually, section 11.14.4 will establish it is the same as the ideal gas temperature scale.

You might wonder why the laws of thermodynamics are numbered from zero. The reason is historical; the first, second, and third laws were already firmly established before in the early twentieth century it was belatedly recognized that an explicit statement of the zeroth law was really needed. If you are already familiar with the second law, you might think it implies the zeroth, but things are not quite that simple.

What about these other laws? The "first law of thermodynamics" is simply stolen from general physics; it states that energy is conserved. The second and third laws will be described in sections 11.8 through 11.10.

11.2 Single-Particle versus System States

The purpose of this section is to describe the generic form of the energy eigenfunctions of a system of weakly interacting particles.

The total number of particles will be indicated by I. If the interactions between the I particles are ignored, any energy eigenfunction of the complete system of I particles can be written in terms of *single-particle* energy eigenfunctions $\psi^{\rm p}_1(\vec{r}, S_z), \psi^{\rm p}_2(\vec{r}, S_z), \ldots$.

The basic case is that of noninteracting particles in a box, like discussed in chapter 6.2. For such particles the single-particle eigenfunctions take the spatial form

$$\psi^{\rm p}_n = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin(k_x x) \sin(k_y y) \sin(k_z z)$$

where k_x, k_y, and k_z are constants, called the "wave number components." Different values for these constants correspond to different single-particle eigen-

functions, with single-particle energy

$$E_n^{\rm p} = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2) = \frac{\hbar^2}{2m}k^2$$

The single-particle energy eigenfunctions will in this chapter be numbered as $n = 1, 2, 3, \ldots, N$. Higher values of index n correspond to eigenfunctions of equal or higher energy $E_n^{\rm p}$.

The single-particle eigenfunctions do not always correspond to a particle in a box. For example, particles caught in a magnetic trap, like in the Bose-Einstein condensation experiments of 1995, might be better described using harmonic oscillator eigenfunctions. Or the particles might be restricted to move in a lower-dimensional space. But a lot of the formulae you can find in literature and in this chapter are in fact derived assuming the simplest case of noninteracting particles in a roomy box.

The details of the single-particle energy eigenfunctions are not really that important in this chapter. What is more interesting are the energy eigenfunctions $\psi_q^{\rm S}$ of complete systems of particles. It will be assumed that these system eigenfunctions are numbered using a counter q, but the way they are numbered also does not really make a difference to the analysis.

As long as the interactions between the particles are weak, energy eigenfunctions of the complete system can be found as products of the single-particle ones. As an important example, at absolute zero temperature, all particles will be in the single-particle ground state $\psi_1^{\rm p}$, and the system will be in its ground state

$$\psi_1^{\rm S} = \psi_1^{\rm p}(\vec{r}_1, S_{z1})\psi_1^{\rm p}(\vec{r}_2, S_{z2})\psi_1^{\rm p}(\vec{r}_3, S_{z3})\psi_1^{\rm p}(\vec{r}_4, S_{z4})\psi_1^{\rm p}(\vec{r}_5, S_{z5})\ldots\psi_1^{\rm p}(\vec{r}_I, S_{zI})$$

where I is the total number of particles in the system. This does assume that the single-particle ground state energy $E_1^{\rm p}$ is not degenerate. More importantly, it assumes that the I particles are not identical fermions. According to the exclusion principle, at most one fermion can go into a single-particle state. (For spin $\frac{1}{2}$ fermions like electrons, two can go into a single *spatial* state, one in the spin-up version, and the other in the spin-down one.)

Statistical thermodynamics, in any case, is much more interested in temperatures that are not zero. Then the system will not be in the ground state, but in some combination of system eigenfunctions of higher energy. As a completely arbitrary example of such a system eigenfunction, take the following one, describing $I = 36$ different particles:

$$\psi_q^{\rm S} = \psi_{24}^{\rm p}(\vec{r}_1, S_{z1})\psi_4^{\rm p}(\vec{r}_2, S_{z2})\psi_7^{\rm p}(\vec{r}_3, S_{z3})\psi_1^{\rm p}(\vec{r}_4, S_{z4})\psi_6^{\rm p}(\vec{r}_5, S_{z5})\ldots\psi_{54}^{\rm p}(\vec{r}_{36}, S_{z36})$$

This system eigenfunction has an energy that is the sum of the 36 single-particle eigenstate energies involved:

$$E_q^{\rm S} = E_{24}^{\rm p} + E_4^{\rm p} + E_7^{\rm p} + E_1^{\rm p} + E_6^{\rm p} + \ldots + E_{54}^{\rm p}$$

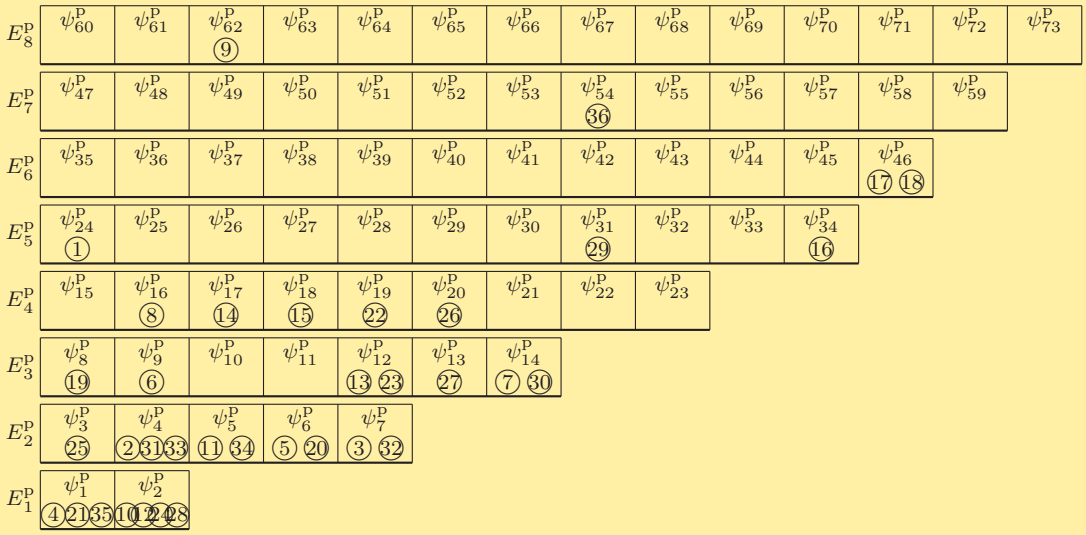


Figure 11.1: Graphical depiction of an arbitrary system energy eigenfunction for 36 distinguishable particles.

To understand the arguments in this chapter, it is essential to visualize the system energy eigenfunctions as in figure 11.1. In this figure the single-particle states are shown as boxes, and the particles that are in those particular single-particle states are shown inside the boxes. In the example, particle 1 is inside the $\psi^{\rm p}_{24}$ box, particle 2 is inside the $\psi^{\rm p}_4$ one, etcetera. It is just the reverse from the mathematical expression above: the mathematical expression shows for each particle in turn what the single-particle eigenstate of that particle is. The figure shows for each type of single-particle eigenstate in turn what particles are in that eigenstate.

To simplify the analysis, in the figure single-particle eigenstates of about the same energy have been grouped together on "shelves." (As a consequence, a subscript to a single-particle energy $E^{\rm p}$ may refer to either a single-particle eigenfunction number n or to a shelf number s, depending on context.) The number of single-particle states on a shelf is intended to roughly simulate the density of states of the particles in a box as described in chapter 6.3. The larger the energy, the more single-particle states there are at that energy; it increases like the square root of the energy. This may not be true for other situations, such as when the particles are confined to a lower-dimensional space, compare chapter 6.12. Various formulae given here and in literature may need to be adjusted then.

Of course, in normal nonnano applications, the number of particles will be astronomically larger than 36 particles; the example is just a small illustration. Even a millimol of particles means on the order of 10^{20} particles. And unless the temperature is incredibly low, those particles will extend to many more single-particle states than the few shown in the figure.

Next, note that you are not going to have something like 10^{20} different types of particles. Instead they are more likely to all be helium atoms, or all electrons or so. If their wave functions overlap nontrivially, that makes a big difference because of the symmetrization requirements of the system wave function.

Consider first the case that the I particles are all identical bosons, like plain helium atoms. In that case the wave function must be symmetric, unchanged, under the exchange of any two of the bosons, and the example wave function above is not. If, for example, particles 2 and 5 are exchanged, it turns the example wave function from

$$\psi^{\mathrm{S}}_q = \psi^{\mathrm{p}}_{24}(\vec{r}_1, S_{z1})\psi^{\mathrm{p}}_4(\vec{r}_2, S_{z2})\psi^{\mathrm{p}}_7(\vec{r}_3, S_{z3})\psi^{\mathrm{p}}_1(\vec{r}_4, S_{z4})\psi^{\mathrm{p}}_6(\vec{r}_5, S_{z5})\ldots\psi^{\mathrm{p}}_{54}(\vec{r}_{36}, S_{z36})$$

into

$$\psi^{\mathrm{S}}_{\underline{q}} = \psi^{\mathrm{p}}_{24}(\vec{r}_1, S_{z1})\psi^{\mathrm{p}}_6(\vec{r}_2, S_{z2})\psi^{\mathrm{p}}_7(\vec{r}_3, S_{z3})\psi^{\mathrm{p}}_1(\vec{r}_4, S_{z4})\psi^{\mathrm{p}}_4(\vec{r}_5, S_{z5})\ldots\psi^{\mathrm{p}}_{54}(\vec{r}_{36}, S_{z36})$$

and that is simply a different wave function, because the states are different, independent functions. In terms of the pictorial representation figure 11.1, swapping the numbers "2" and "5" in the particles changes the picture.

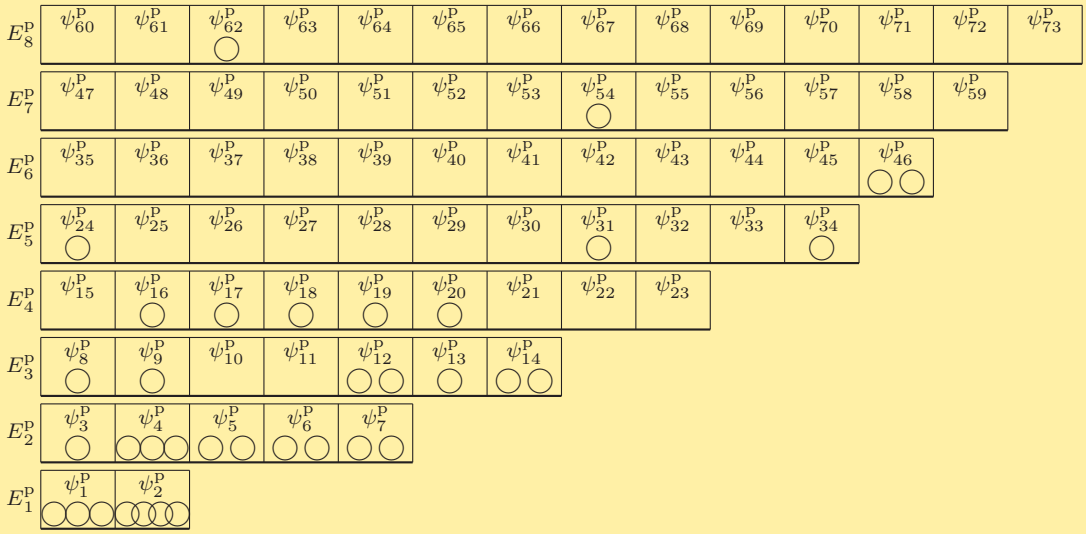


Figure 11.2: Graphical depiction of an arbitrary system energy eigenfunction for 36 identical bosons.

As chapter 5.7 explained, to eliminate the problem that exchanging particles 2 and 5 changes the wave function, the original and exchanged wave functions must be combined together. And to eliminate the problem for *any* two particles, *all* wave functions that can be obtained by merely swapping numbers must be combined together equally into a *single* wave function multiplied by a *single* undetermined coefficient. In terms of figure 11.1, we need to combine the wave functions with all possible permutations of the numbers inside the particles into

one. And if all permutations of the numbers are equally included, then those numbers no longer add any nontrivial additional information; they may as well be left out. That makes the pictorial representation of an example system wave function for identical bosons as shown in figure 11.2.

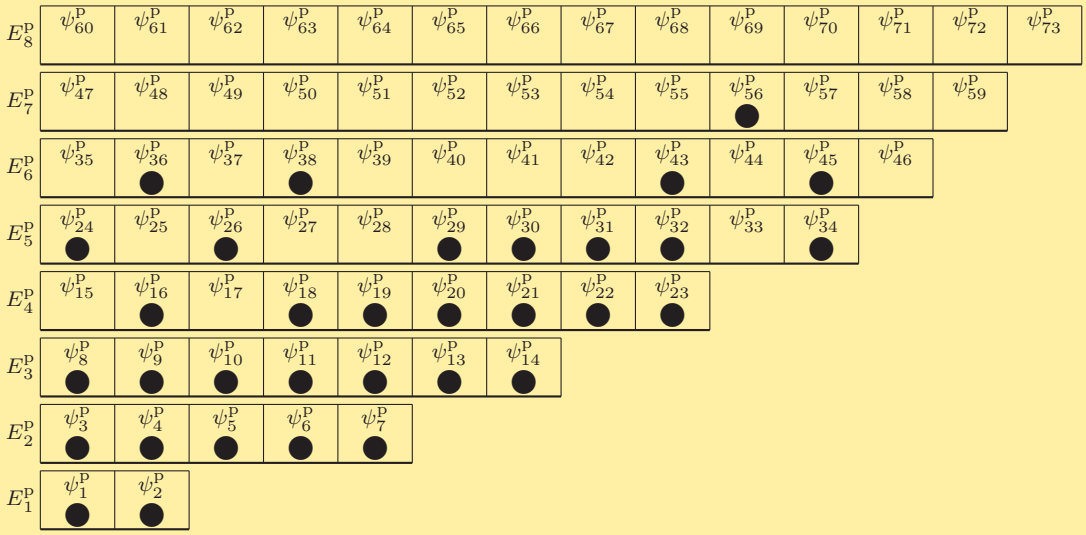


Figure 11.3: Graphical depiction of an arbitrary system energy eigenfunction for 33 identical fermions.

For identical fermions, the situation is similar, except that the different wave functions must be combined with equal or opposite sign, depending on whether it takes an odd or even number of particle swaps to turn one into the other. And such wave functions only exist if the I single-particle wave functions involved are all different. That is the Pauli exclusion principle. The pictorial representation figure 11.2 for bosons is totally unacceptable for fermions since it uses many of the single-particle states for more than one particle. There can be at most one fermion in each type of single-particle state. An example of a wave function that is acceptable for a system of identical fermions is shown in figure 11.3.

Looking at the example pictorial representations for systems of bosons and fermions, it may not be surprising that such particles are often called "indistinguishable." Of course, in classical quantum mechanics, there is still an electron 1, an electron 2, etcetera; they are mathematically distinguished. Still, it is convenient to use the term "distinguishable" for particles for which the symmetrization requirements can be ignored.

The prime example is the atoms of an ideal gas in a box; almost by definition, the interactions between such atoms are negligible. And that allows the quantum results to be referred back to the well-understood properties of ideal gases obtained in classical physics. Probably you would like to see all results follow naturally from quantum mechanics, not classical physics, and that would be very nice indeed. But it would be very hard to follow up on. As Baierlein [4,

p. 109] notes, real-life physics adopts whichever theoretical approach offers the easiest calculation or the most insight. This book's approach really is to formulate as much as possible in terms of the quantum-mechanical ideas discussed here. But do be aware that it is a much more messy world when you go out there.

11.3 How Many System Eigenfunctions?

The fundamental question from which all of quantum statistics springs is a very basic one: How many system energy eigenstates are there with given generic properties? This section will address that question.

Of course, by definition each system energy eigenfunction is unique. Figures 11.1–11.3 give examples of such unique energy eigenfunctions for systems of distinguishable particles, indistinguishable bosons, and indistinguishable fermions. But trying to get accurate data on each individual eigenfunction just does not work. That is much too big a challenge.

Quantum statistics must satisfy itself by figuring out the probabilities on groups of system eigenfunctions with similar properties. To do so, the single-particle energy eigenstates are best grouped together on shelves of similar energy, as illustrated in figures 11.1–11.3. Doing so allows for more answerable questions such as: "How many system energy eigenfunctions $\psi^{\rm S}_q$ have I_1 out of the I total particles on shelf 1, another I_2 on shelf 2, etcetera?" In other words, if $\vec{I}$ stands for a given set of shelf occupation numbers $(I_1, I_2, I_3, \ldots)$, then what is the number $Q_{\vec{I}}$ of system eigenfunctions $\psi^{\rm S}_q$ that have those shelf occupation numbers?

That question is answerable with some clever mathematics; it is a big thing in various textbooks. However, the suspicion is that this is more because of the "neat" mathematics than because of the actual physical insight that these derivations provide. In this book, the derivations are shoved away into {D.56}. But here are the results. (Drums please.) The system eigenfunction counts for distinguishable particles, bosons, and fermions are:

$$Q^{\rm d}_{\vec{I}} = I! \prod_{\text{all } s} \frac{N_s^{I_s}}{\Big(I_s\Big)!} \tag{11.1}$$

$$Q^{\rm b}_{\vec{I}} = \prod_{\text{all } s} \frac{\Big(I_s + N_s - 1\Big)!}{\Big(I_s\Big)!\Big(N_s - 1\Big)!} \tag{11.2}$$

$$Q^{\rm f}_{\vec{I}} = \prod_{\text{all } s} \frac{\Big(N_s\Big)!}{\Big(I_s\Big)!\Big(N_s - I_s\Big)!} \tag{11.3}$$

where Π means the product of all the terms of the form shown to its right that can be obtained by substituting in every possible value of the shelf number s. That is just like Σ would mean the sum of all these terms. For example, for distinguishable particles

$$Q^{\rm d}_{\vec{I}} = I! \frac{N_1^{I_1}}{\Big(I_1\Big)!} \frac{N_2^{I_2}}{\Big(I_2\Big)!} \frac{N_3^{I_3}}{\Big(I_3\Big)!} \frac{N_4^{I_4}}{\Big(I_4\Big)!} \ldots$$

where N_1 is the number of single-particle energy states on shelf 1 and I_1 the number of particles on that shelf, N_2 the number of single-particle energy states on shelf 2 and I_2 the number of particles on that shelf, etcetera. Also an exclamation mark indicates the factorial function, defined as

$$n! = \prod_{\underline{n}=1}^{n} \underline{n} = 1 \times 2 \times 3 \times \ldots \times n$$

For example, $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$. The eigenfunction counts may also involve 0!, which is defined to be 1, and $n!$ for negative n, which is defined to be infinity. The latter is essential to ensure that the eigenfunction count is zero as it should be for fermion eigenfunctions that try to put more particles on a shelf than there are states on it.

This section is mainly concerned with explaining qualitatively why these system eigenfunction counts matter *physically*. And to do so, a very simple model system having only three shelves will suffice.

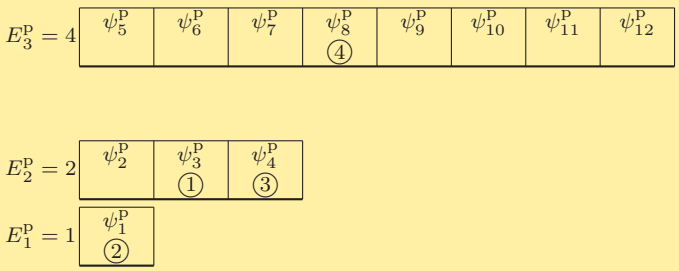


Figure 11.4: Illustrative small model system having 4 distinguishable particles. The particular eigenfunction shown is arbitrary.

The first example is illustrated in quantum-mechanical terms in figure 11.4. Like the other examples, it has only three shelves, and it has only $I = 4$ distinguishable particles. Shelf 1 has $N_1 = 1$ single-particle state with energy $E_1^{\rm p} = 1$ (arbitrary units), shelf 2 has $N_2 = 3$ single-particle states with energy $E_2^{\rm p} = 2$, (note that $3 \approx 2\sqrt{2}$), and shelf 3 has $N_3 = 4\sqrt{4} = 8$ single-particle states with energy $E_3^{\rm p} = 4$. One major deficiency of this model is the small number of particles and states, but that will be fixed in the later examples. More seriously is that there are no shelves with energies above $E_3^{\rm p} = 4$. To mitigate

that problem, for the time being the average energy per particle of the system eigenfunctions will be restricted to no more than 2.5. This will leave shelf 3 largely empty, reducing the effects of the missing shelves of still higher energy.

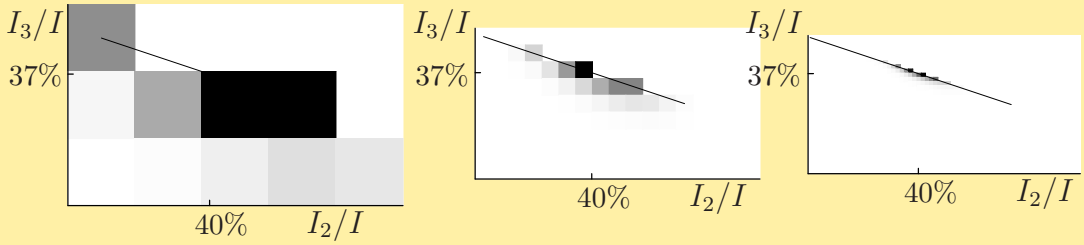


Figure 11.5: The number of system energy eigenfunctions for a simple model system with only three energy shelves. Positions of the squares indicate the numbers of particles on shelves 2 and 3; darkness of the squares indicates the relative number of eigenfunctions with those shelf numbers. Left: system with 4 distinguishable particles, middle: 16, right: 64.

Now the question is, how many energy eigenfunctions are there for a given set of shelf occupation numbers $\vec{I} = (I_1, I_2, I_3)$? The answer, as given by (11.1), is shown graphically in the left graph of figure 11.5. Darker squares indicate more eigenfunctions with those shelf occupation numbers. The oblique line in figure 11.5 is the line above which the average energy per particle exceeds the chosen limit of 2.5.

Some example observations about the figure may help to understand it. For example, there is only one system eigenfunction with all 4 particles on shelf 1, i.e. with $I_1 = 4$ and $I_2 = I_3 = 0$; it is

$$\psi_1^{\rm S} = \psi_1^{\rm p}(\vec{r}_1, S_{z1})\psi_1^{\rm p}(\vec{r}_2, S_{z2})\psi_1^{\rm p}(\vec{r}_3, S_{z3})\psi_1^{\rm p}(\vec{r}_4, S_{z4}).$$

This is represented by the white square at the origin in the left graph of figure 11.5.

As another example, the darkest square in the left graph of figure 11.5 represents system eigenfunctions that have shelf numbers $\vec{I} = (1, 2, 1)$, i.e. $I_1 = 1$, $I_2 = 2$, $I_3 = 1$: one particle on shelf 1, two particles on shelf 2, and one particle on shelf 3. A completely arbitrary example of such a system energy eigenfunction,

$$\psi_3^{\rm p}(\vec{r}_1, S_{z1})\psi_1^{\rm p}(\vec{r}_2, S_{z2})\psi_4^{\rm p}(\vec{r}_3, S_{z3})\psi_8^{\rm p}(\vec{r}_4, S_{z4}),$$

is the one depicted in figure 11.4. It has particle 1 in single-particle state $\psi_3^{\rm p}$, which is on shelf 2, particle 2 in $\psi_1^{\rm p}$, which is on shelf 1, particle 3 in $\psi_4^{\rm p}$ which is on shelf 2, and particle 4 in $\psi_8^{\rm p}$, which is on shelf 3. But there are a lot more system eigenfunctions with the same shelf occupation numbers; in fact, there are

$$4 \times 3 \times 8 \times 3 \times 3 = 864$$

such eigenfunctions, since there are 4 possible choices for the particle that goes on shelf 1, times a remaining 3 possible choices for the particle that goes on shelf 3, times 8 possible choices $\psi^{\rm p}_5$ through $\psi^{\rm p}_{12}$ for the single-particle eigenfunction on shelf 3 that that particle can go into, times 3 possible choices $\psi^{\rm p}_2$ through $\psi^{\rm p}_4$ that each of the remaining two particles on shelf 2 can go into.

Next, consider a system four times as big. That means that there are four times as many particles, so $I = 16$ particles, in a box that has four times the volume. If the volume of the box becomes 4 times as large, there are four times as many single-particle states on each shelf, since the number of states per unit volume at a given single-particle energy is constant, compare (6.6). Shelf 1 now has 4 states, shelf 2 has 12, and shelf 3 has 32. The number of energy states for given shelf occupation numbers is shown as grey tones in the middle graph of figure 11.5. Now the number of system energy eigenfunctions that have all particles on shelf 1 is not one, but 4^{16} or 4 294 967 296, since there are 4 different states on shelf 1 that each of the 16 particles can go into. That is obviously quite lot of system eigenfunctions, but it is dwarfed by the darkest square, states with shelf occupation numbers $\vec{I}$ = (4,6,6). There are about 1.4 10^{24} system energy eigenfunctions with those shelf occupation numbers. So the $\vec{I}$ = (16,0,0) square at the origin stays lily-white despite having over 4 billion energy eigenfunctions.

If the system size is increased by another factor 4, to 64 particles, the number of states with occupation numbers $\vec{I}$ = (64,0,0), all particles on shelf 1, is 1.2 10^{77}, a tremendous number, but totally humiliated by the 2.7 10^{138} eigenfunctions that have occupation numbers $\vec{I}$ = (14,27,23). Taking the ratio of these two numbers shows that there are 2.3 10^{61} energy eigenfunctions with shelf numbers $(14, 27, 23)$ for each eigenfunction with shelf numbers $(64, 0, 0)$. By the time the system reaches, say, 10^{20} particles, still less than a millimol, the number of system energy eigenstates for each set of occupation numbers is astronomical, but so are the differences between the shelf numbers that have the most and those that have less. The tick marks in figure 11.5 indicate that for large systems, the darkest square will have 40% of the particles on shelf 2, 37% on shelf 3, and the remaining 23% on shelf 1.

These general trends do not just apply to this simple model system; they are typical:

> *The number of system energy eigenfunctions for a macroscopic system is astronomical, and so are the differences in numbers.*

Another trend illustrated by figure 11.5 has to do with the effect of system energy. The system energy of an energy eigenfunction is given in terms of its shelf numbers by

$$E^{\rm S} = I_1 E^{\rm p}_1 + I_2 E^{\rm p}_2 + I_3 E^{\rm p}_3$$

so all eigenfunctions with the same shelf numbers have the same system energy. In particular, the squares just below the oblique cut-off line in figure 11.5 have

the highest system energy. It is seen that these shelf numbers also have by far the most energy eigenfunctions:

> *The number of system energy eigenfunctions with a higher energy typically dwarfs the number of system eigenfunctions with a lower energy.*

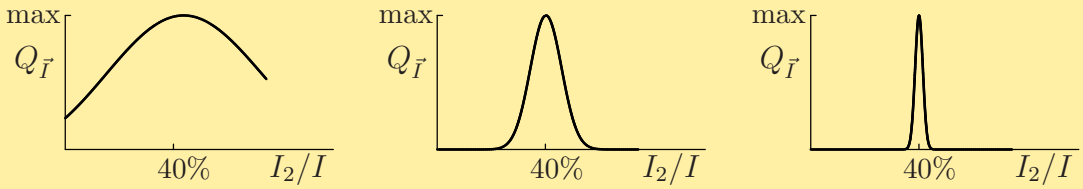


Figure 11.6: Number of energy eigenfunctions on the oblique energy line in the previous figure. (The curves are mathematically interpolated to allow a continuously varying fraction of particles on shelf 2.) Left: 4 particles, middle: 64, right: 1,024.

Next assume that the system has exactly the energy of the oblique cut-off line in figure 11.5, with *zero* uncertainty. The number of energy eigenstates $Q_{\vec{I}}$ on that oblique line is plotted in figure 11.6 as a function of the fraction of particles I_2/I on shelf 2. (To get a smooth continuous curve, the values have been mathematically interpolated in between the integer values of I_2. The continuous function that interpolates $n!$ is called the gamma function; see the notations section under "!" for details.) The maximum number of energy eigenstates occurs at about $I_2/I = 40\%$, corresponding to $I_3 = 37\%$ and $I_1 = 23\%$. This set of occupation numbers, $(I_1, I_2, I_3) = (0.23, 0.40, 0.37)I$, is called the "most probable set of occupation numbers." If you pick an eigenfunction at random, you have more chance of getting one with that set of occupation numbers than one with a different given set of occupation numbers.

To be sure, if the number of particles is large, the chances of picking any eigenfunction with an *exact* set of occupation numbers is small. But note how the "spike" in figure 11.6 becomes narrower with increasing number of particles. You may not pick an eigenfunction with *exactly* the most probable set of shelf numbers, but you are quite sure to pick one with shelf numbers very close to it. By the time the system size reaches, say, 10^{20} particles, the spike becomes for all practical purposes a mathematical line. Then essentially *all* eigenfunctions have very precisely 23% of their particles on shelf 1 at energy $E_1^{\rm p}$, 40% on shelf 2 at energy $E_2^{\rm p}$, and 37% on shelf 3 at energy $E_3^{\rm p}$.

Since there is only an incredibly small fraction of eigenfunctions that do not have very accurately the most probable occupation numbers, it seems intuitively obvious that in thermal equilibrium, the physical system must have the same distribution of particle energies. Why would nature prefer one of those extremely

rare eigenfunctions that do not have these occupation numbers, rather than one of the vast majority that do? In fact, {N.23},

> *It is a fundamental assumption of statistical mechanics that in thermal equilibrium, all system energy eigenfunctions with the same energy have the same probability.*

So the most probable set of shelf numbers, as found from the count of eigenfunctions, gives the distribution of particle energies in thermal equilibrium.

This then is the final conclusion: the particle energy distribution of a macroscopic system of weakly interacting particles at a given energy can be obtained by merely *counting the system energy eigenstates*. It can be done *without doing any physics*. Whatever physics may want to do, it is just not enough to offset the vast numerical superiority of the eigenfunctions with very accurately the most probable shelf numbers.

11.4 Particle-Energy Distribution Functions

The objective in this section is to relate the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac particle energy distributions of chapter 6 to the conclusions obtained in the previous section. The three distributions give the number of particles that have given single-particle energies.

In terms of the picture developed in the previous sections, they describe how many particles are on each energy shelf relative to the number of single-particle states on the shelf. The distributions also assume that the number of shelves is taken large enough that their energy can be assumed to vary continuously.

According to the conclusion of the previous section, for a system with given energy it is sufficient to find the most probable set of energy shelf occupation numbers, the set that has the highest number of system energy eigenfunctions. That gives the number of particles on each energy shelf that is the most probable. As the previous section demonstrated by example, the fraction of eigenfunctions that have significantly different shelf occupation numbers than the most probable ones is so small for a macroscopic system that it can be ignored.

Therefore, the basic approach to find the three distribution functions is to first identify all sets of shelf occupation numbers $\vec{I}$ that have the given energy, and then among these pick out the set that has the most system eigenfunctions $Q_{\vec{I}}$. There are some technical issues with that, {N.24}, but they can be worked out, as in derivation {D.57}.

The final result is, of course, the particle energy distributions from chapter 6:

$$\iota^{\rm b} = \frac{1}{e^{(E^{\rm p}-\mu)/k_{\rm B}T}-1} \qquad \iota^{\rm d} = \frac{1}{e^{(E^{\rm p}-\mu)/k_{\rm B}T}} \qquad \iota^{\rm f} = \frac{1}{e^{(E^{\rm p}-\mu)/k_{\rm B}T}+1}.$$

Here ι indicates the number of particles per single-particle state, more precisely, $\iota = I_s/N_s$. This ratio is independent of the precise details of how the shelves are selected, as long as their energies are closely spaced. However, for identical bosons it does assume that the number of single-particle states on a shelf is large. If that assumption is problematic, the more accurate formulae in derivation {D.57} should be consulted. The main case for which there is a real problem is for the ground state in Bose-Einstein condensation.

It may be noted that "T" in the above distribution laws is a temperature, but the derivation in the note did not establish it is the same temperature scale that you would get with an ideal-gas thermometer. That will be shown in section 11.14.4. For now note that T will normally have to be positive. Otherwise the derived energy distributions would have the number of particles become infinity at infinite shelf energies. For some weird system for which there is an upper limit to the possible single-particle energies, this argument does not apply, and negative temperatures cannot be excluded. But for particles in a box, arbitrarily large energy levels do exist, see chapter 6.2, and the temperature must be positive.

The derivation also did not show that μ in the above distributions is the chemical potential as is defined in general thermodynamics. That will eventually be shown in derivation {D.61}. Note that for particles like photons that can be readily created or annihilated, there is no chemical potential; μ entered into the derivation {D.57} through the constraint that the number of particles of the system is a given. A look at the note shows that the formulae still apply for such transient particles if you simply put $\mu = 0$.

For permanent particles, increasingly large negative values of the chemical potential μ decrease the number of particles at all energies. Therefore large negative μ corresponds to systems of very low particle densities. If μ is sufficiently negative that $e^{(E^{\rm p}-\mu)/k_{\rm B}T}$ is large even for the single-particle ground state, the ± 1 that characterize the Fermi-Dirac and Bose-Einstein distributions can be ignored compared to the exponential, and the three distributions become equal:

> *The symmetrization requirements for bosons and fermions can be ignored under conditions of very low particle densities.*

These are ideal gas conditions, section 11.14.4

Decreasing the temperature will primarily thin out the particle numbers at high energies. In this sense, yes, temperature reductions are indeed to some extent associated with (kinetic) energy reductions.

11.5 The Canonical Probability Distribution

The particle energy distribution functions in the previous section were derived assuming that the energy is given. In quantum-mechanical terms, it was as-

sumed that the energy had a definite value. However, that cannot really be right, for one because of the energy-time uncertainty principle.

Assume for a second that a lot of boxes of particles are carefully prepared, all with a system energy as precise as it can be made. And that all these boxes are then stacked together into one big system. In the combined system of stacked boxes, the energy is presumably quite unambiguous, since the random errors are likely to cancel each other, rather than add up systematically. In fact, simplistic statistics would expect the relative error in the energy of the combined system to decrease like the square root of the number of boxes.

But for the carefully prepared individual boxes, the future of their lack of energy uncertainty is much bleaker. Surely a single box in the stack may randomly exchange a bit of energy with the other boxes. Of course, when a box acquires much more energy than the others, the exchange will no longer be random, but almost certainly go from the hotter box to the cooler ones. Still, it seems unavoidable that quite a lot of uncertainty in the energy of the individual boxes would result. The boxes still have a precise *temperature*, being in thermal equilibrium with the larger system, but no longer a precise *energy*.

Then the appropriate way to describe the individual boxes is no longer in terms of given energy, but in terms of probabilities. The proper expression for the probabilities is "deduced" in derivation {D.58}. It turns out that when the temperature T, but not the energy of a system is certain, the system energy eigenfunctions $\psi_q^{\rm S}$ can be assigned probabilities of the form

$$\boxed{P_q = \frac{1}{Z} e^{-E_q^{\rm S}/k_{\rm B}T}} \tag{11.4}$$

where $k_{\rm B} = 1.380\,65\ 10^{-23}$ J/K is the Boltzmann constant. This equation for the probabilities is called the Gibbs "canonical probability distribution." Feynman [18, p. 1] calls it the summit of statistical mechanics.

The exponential by itself is called the "Boltzmann factor." The normalization factor Z, which makes sure that the probabilities all together sum to one, is called the "partition function." It equals

$$\boxed{Z = \sum_{\text{all } q} e^{-E_q^{\rm S}/k_{\rm B}T}} \tag{11.5}$$

You might wonder why a mere normalization factor warrants its own name. It turns out that if an analytical expression for the partition function $Z(T, V, I)$ is available, various quantities of interest may be found from it by taking suitable partial derivatives. Examples will be given in subsequent sections.

The canonical probability distribution conforms to the fundamental assumption of quantum statistics that eigenfunctions of the same energy have the same probability. However, it adds that for system eigenfunctions with different energies, the higher energies are less likely. Massively less likely, to be sure, because

the system energy $E_q^{\rm S}$ is a macroscopic energy, while the energy $k_{\rm B}T$ is a microscopic energy level, roughly the kinetic energy of a single atom in an ideal gas at that temperature. So the Boltzmann factor decays extremely rapidly with energy.

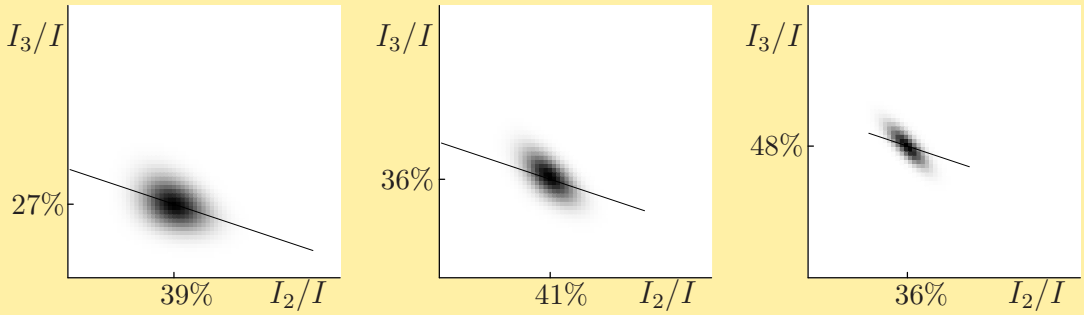


Figure 11.7: Probabilities of shelf-number sets for the simple 64 particle model system if there is uncertainty in energy. More probable shelf-number distributions are shown darker. Left: identical bosons, middle: distinguishable particles, right: identical fermions. The temperature is the same as in the previous two figures.

So, what happens to the simple model system from section 11.3 when the energy is no longer certain, and instead the probabilities are given by the canonical probability distribution? The answer is in the middle graphic of figure 11.7. Note that there is no longer a need to limit the displayed energies; the strong exponential decay of the Boltzmann factor takes care of killing off the high energy eigenfunctions. The rapid growth of the number of eigenfunctions does remain evident at lower energies where the Boltzmann factor has not yet reached enough strength.

There is still an oblique energy line in figure 11.7, but it is no longer limiting energy; it is merely the energy at the most probable shelf occupation numbers. Equivalently, it is the "expectation energy" of the system, defined following the ideas of chapter 4.4.1 as

$$\langle E \rangle \equiv \sum_{\text{all } q} P_q E_q^{\rm S} \equiv E$$

because for a macroscopic system size, the most probable and expectation values are the same. That is a direct result of the black blob collapsing towards a single point for increasing system size: in a macroscopic system, essentially all system eigenfunctions have the same macroscopic properties.

In thermodynamics, the expectation energy is called the "internal energy" and indicated by E or U. This book will use E, dropping the angular brackets. The difference in notation from the single-particle/shelf/system energies is that the internal energy is plain E with no subscripts or superscripts.

Figure 11.7 also shows the shelf occupation number probabilities if the example 64 particles are not distinguishable, but identical bosons or identical fermions. The most probable shelf numbers are not the same, since bosons and fermions have different numbers of eigenfunctions than distinguishable particles, but as the figure shows, the effects are not dramatic at the shown temperature, $k_{\rm B}T = 1.85$ in the arbitrary energy units.

11.6 Low Temperature Behavior

The three-shelf simple model used to illustrate the basic ideas of quantum statistics qualitatively can also be used to illustrate the low temperature behavior that was discussed in chapter 6. To do so, however, the first shelf must be taken to contain just a single, nondegenerate ground state.

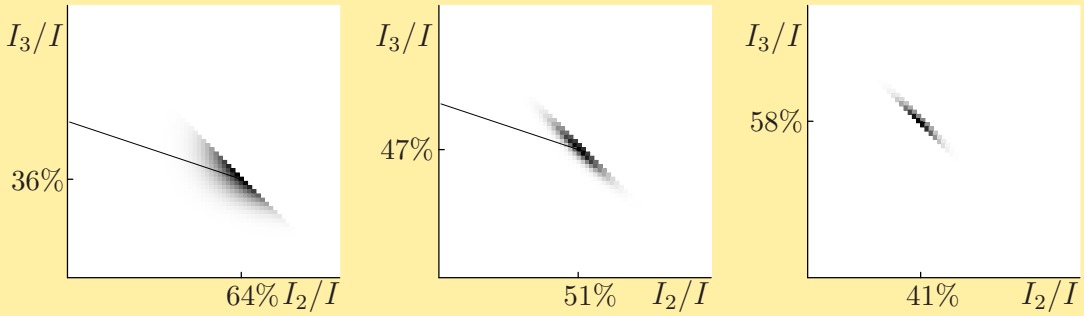


Figure 11.8: Probabilities of shelf-number sets for the simple 64 particle model system if shelf 1 is a nondegenerate ground state. Left: identical bosons, middle: distinguishable particles, right: identical fermions. The temperature is the same as in the previous figures.

In that case, figure 11.7 of the previous section turns into figure 11.8. Neither of the three systems sees much reason to put any measurable amount of particles in the first shelf. Why would they, it contains only one single-particle state out of 177? In particular, the most probable shelf numbers are right at the 45° limiting line through the points $I_2 = I$, $I_3 = 0$ and $I_2 = 0$, $I_3 = I$ on which $I_1 = 0$. Actually, the mathematics of the system of bosons would like to put a *negative* number of bosons on the first shelf, and must be constrained to put zero on it.

If the temperature is lowered however, as in figure 11.9 things change, especially for the system of bosons. Now the mathematics of the most probable state wants to put a positive number of bosons on shelf 1, and a large fraction of them to boot, considering that it is only one state out of 177. The most probable distribution drops way below the 45° limiting line. The mathematics for distinguishable particles and fermions does not yet see any reason to panic, and still leaves shelf 1 largely empty.

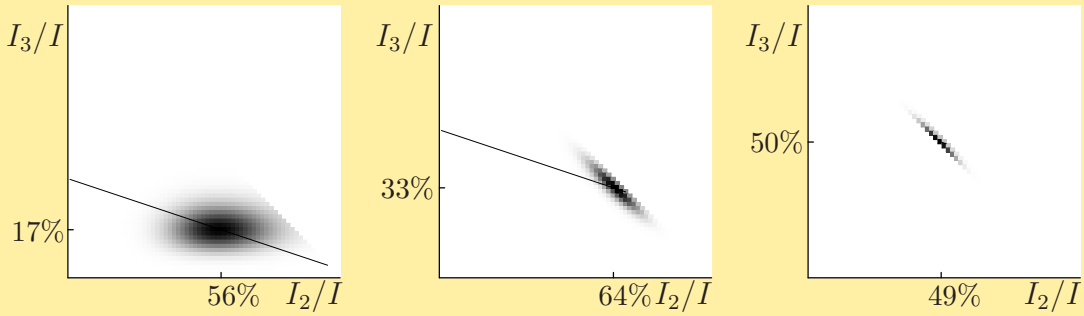


Figure 11.9: Like the previous figure, but at a lower temperature.

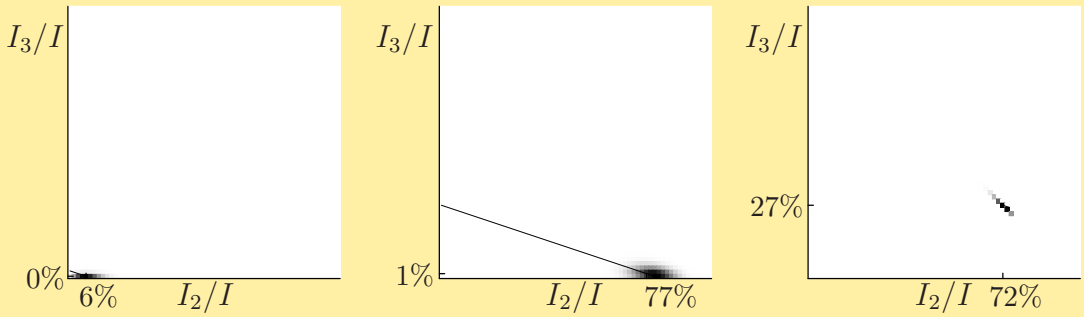


Figure 11.10: Like the previous figures, but at a still lower temperature.

When the temperature is lowered still much lower, as shown in figure 11.10, almost all bosons drop into the ground state and the most probable state is right next to the origin $I_2 = I_3 = 0$. In contrast, while the system of distinguishable particles does recognize that high-energy shelf 3 becomes quite unreachable with the available amount of thermal energy, it still has a quite significant fraction of the particles on shelf 2. And the system of fermions will never drop to shelf 1, however low the temperature. Because of the Pauli exclusion principle, only one fermion out of the 64 can ever go on shelf one, and only 48, 75%. can go on shelf 2. The remaining 23% will stay on the high-energy shelf however low the temperature goes.

If you still need convincing that temperature is a measure of hotness, and not of thermal kinetic energy, there it is. The three systems of figure 11.10 are all at the same temperature, but there are vast differences in their kinetic energy. In thermal contact at very low temperatures, the system of fermions runs off with almost all the energy, leaving a small morsel of energy for the system of distinguishable particles, and the system of bosons gets practically nothing.

It is really weird. Any distribution of shelf numbers that is valid for distinguishable particles is exactly as valid for bosons and vice/versa; it is just the *number* of eigenfunctions with those shelf numbers that is different. But when the two systems are brought into thermal contact at very low temperatures, the distinguishable particles get all the energy. It is just as possible from an energy

conservation and quantum mechanics point of view that all the energy goes to the bosons instead of to the distinguishable particles. But it becomes astronomically unlikely because there are so *few* eigenfunctions like that. (Do note that it is assumed here that the temperature is so low that almost all bosons have dropped in the ground state. As long as the temperatures do not become much smaller than the one of Bose-Einstein condensation, the energies of systems of bosons and distinguishable particles remain quite comparable, as in figure 11.9.)

11.7 The Basic Thermodynamic Variables

This section introduces the most important basic players in thermodynamics.

The primary thermodynamic property introduced so far is the temperature. Recall that temperature is a measure of the hotness of the substance, a measure of how eager it is to dump energy onto other systems. Temperature is called an "intensive variable;" it is the same for two systems that differ only in size.

The total number of particles I or the total volume of their box V are not intensive variables; they are "extensive variables," variables that increase in value proportional to the system size. Often, however, you are only interested in the properties of your substance, not the amount. In that case, intensive variables can be created by taking ratios of the extensive ones; in particular, I/V is an intensive variable called the "particle density." It is the number of particles per unit volume. If you restrict your attention to only one half of your box with particles, the particle density is still the same, with half the particles in half the volume.

Note that under equilibrium conditions, it suffices to know the temperature and particle density to fully fix the state that a given system is in. More generally, the rule is that:

> *Two intensive variables must be known to fully determine the intensive properties of a simple substance in thermal equilibrium.*

(To be precise, in a two-phase equilibrium like a liquid-vapor mixture, pressure and temperature are related, and would not be sufficient to determine something like *net* specific volume. They do still suffice to determine the specific volumes of the liquid and vapor parts individually, in any case.) If the amount of substance is also desired, knowledge of at least one extensive variable is required, making three variables that must be known in total.

Since the number of particles will have very large values, for macroscopic work the particle density is often not very convenient, and somewhat differently defined, but completely equivalent variables are used. The most common are the (mass) "density" ρ, found by multiplying the particle density with the single-particle mass m, $\rho \equiv mI/V$, or its reciprocal, the "specific volume" $v \equiv V/mI$.

The density is the system mass per unit system volume, and the specific volume is the system volume per unit system mass.

Alternatively, to keep the values for the number of particles in check, they may be expressed in "moles," multiples of Avogadro's number

$$I_{\rm A} \approx 6.022\,1\ 10^{23}$$

That produces the "molar density" $\bar{\rho} \equiv I/I_{\rm A}V$ and "molar specific volume" $\bar{v} \equiv VI_{\rm A}/I$. In thermodynamic textbooks, the use of kilo mol (kmol) instead of mol has become quite standard (but then, so has the use of kilo Newton instead of Newton.) The conversion factor between molar and nonmolar specific quantities is called the "molar mass" M; it is applied according to its units of kg/kmol. Note that thermo books for engineers may misname M to be the "molecular mass". The numerical value of the molar mass is roughly the total number of protons and neutrons in the nuclei of a single molecule; in fact, the weird number of particles given by Avogadro's number was chosen to achieve this.

So what else is there? Well, there is the energy of the system. In view of the uncertainty in energy, the appropriate system energy is defined as the expectation value,

$$\boxed{E = \sum_{\text{all } q} P_q E_q^{\rm S}} \tag{11.6}$$

where P_q is the canonical probability of (11.4), (11.5). Quantity E is called the "internal energy." In engineering thermodynamics books, it is usually indicated by U, but this is physics. The intensive equivalent e is found by dividing by the system mass; $e = E/mI$. Note the convention of indicating extensive variables by a capital and their intensive value per unit mass with the corresponding lower case letter. A specific quantity on a molar basis is lower case with a bar above it.

As a demonstration of the importance of the partition function mentioned in the previous section, if the partition function (11.5) is differentiated with respect to temperature, you get

$$\left(\frac{\partial Z}{\partial T}\right)_{V \text{ constant}} = \frac{1}{k_{\rm B}T^2} \sum_{\text{all } q} E_q^{\rm S} e^{-E_q^{\rm S}/k_{\rm B}T}.$$

(The volume of the system should be held constant in order that the energy eigenfunctions do not change.) Dividing both sides by Z turns the derivative in the left hand side into that of the logarithm of Z, and the sum in the right hand side into the internal energy E, and you get

$$\boxed{E = k_{\rm B}T^2 \left(\frac{\partial \ln Z}{\partial T}\right)_{V \text{ constant}}} \tag{11.7}$$

Next there is the "pressure" P, being the force with which the substance pushes on the surfaces of the box it is in per unit surface area. To identify P quantum mechanically, first consider a system in a single energy eigenfunction $E_q^{\rm S}$ for certain. If the volume of the box is slightly changed, there will be a corresponding slight change in the energy eigenfunction $E_q^{\rm S}$, (the boundary conditions of the Hamiltonian eigenvalue problem will change), and in particular its energy will slightly change. Energy conservation requires that the change in energy $\mathrm{d}E_q^{\rm S}$is offset by the work done by the containing walls on the substance. Now the work done by the wall pressure on the substance equals

$$-P\,\mathrm{d}V.$$

(The force is pressure times area and is normal to the area; the work is force times displacement in the direction of the force; combining the two, area times displacement normal to that area gives change in volume. The minus sign is because the displacement must be inwards for the pressure force on the substance to do positive work.) So for the system in a single eigenstate, the pressure equals $P = -\mathrm{d}E_q^{\rm S}/\mathrm{d}V$. For a real system with uncertainty in energy, the pressure is defined as the expectation value:

$$\boxed{P = -\sum_{\text{all } q} P_q \frac{\mathrm{d}E_q^{\rm S}}{\mathrm{d}V}} \tag{11.8}$$

It may be verified by simple substitution that this, too may be obtained from the partition function, now by differentiating with respect to volume keeping temperature constant:

$$\boxed{P = k_{\rm B}T \left(\frac{\partial \ln Z}{\partial V}\right)_{T \text{ constant}}} \tag{11.9}$$

While the final quantum mechanical *definition* of the pressure is quite sound, it should be pointed out that the original definition in terms of force was very artificial. And not just because force is a poor quantum variable. Even if a system in a single eigenfunction could be created, the walls of the system would have to be idealized to assume that the energy change equals the work $-P\,\mathrm{d}V$. For example, if the walls of the box would consist of molecules that were hotter than the particles inside, the walls too would add energy to the system, and take it out of its single energy eigenstate to boot. And even macroscopically, for pressure times area to be the force requires that the system is in thermal equilibrium. It would not be true for a system evolving in a violent way.

Often a particular combination of the variables defined above is very convenient; the"enthalpy" H is defined as

$$\boxed{H = E + PV} \tag{11.10}$$

Enthalpy is not a fundamentally new variable, just a combination of existing ones.

Assuming that the system evolves while staying at least approximately in thermal equilibrium, the "first law of thermodynamics" can be stated macroscopically as follows:

$$\boxed{\mathrm{d}E = \delta Q - P\,\mathrm{d}V} \tag{11.11}$$

In words, the internal energy of the system changes by the amount δQ of heat added plus the amount $-P\,\mathrm{d}V$ of work done on the system. It is just energy conservation expressed in thermodynamic terms. (And it assumes that other forms of energy than internal energy and work done while expanding can be ignored.)

Note the use of a straight d for the changes in internal energy E and volume V, but a δ for the heat energy added. It reflects that $\mathrm{d}E$ and $\mathrm{d}V$ are changes in properties of the system, but δQ is not; δQ is a small amount of energy exchanged between systems, not a property of any system. Also note that while popularly you might talk about the heat within a system, it is standard in thermodynamics to refer to the thermal energy within a system as internal energy, and reserve the term "heat" for *exchanged* thermal energy.

Just two more variables. The "specific heat at constant volume" C_v is defined as the heat that must be added to the substance for each degree temperature change, per unit mass and keeping the volume constant. In terms of the first law on a unit mass basis,

$$\mathrm{d}e = \delta q - P\,\mathrm{d}v,$$

it means that C_v is defined as $\delta q/\mathrm{d}T$ when $\mathrm{d}v = 0$. So C_v is the derivative of the specific internal energy e with respect to temperature. To be specific, since specifying e normally requires *two* intensive variables, C_v is the partial derivative of e keeping specific volume constant:

$$\boxed{C_v \equiv \left(\frac{\partial e}{\partial T}\right)_v} \tag{11.12}$$

Note that in thermodynamics the quantity being held constant while taking the partial derivative is shown as a subscript to parentheses enclosing the derivative. You did not see that in calculus, but that is because in mathematics, they tend to choose a couple of independent variables and stick with them. In thermodynamics, two independent variables are needed, (assuming the amount of substance is a given), but the choice of which two changes all the time. Therefore, listing what is held constant in the derivatives is crucial.

The specific heat at constant pressure C_p is defined similarly as C_v, except that pressure, instead of volume, is being held constant. According to the first law above, the heat added is now $de + P\,\mathrm{d}v$ and that is the change in enthalpy $h = e + Pv$. There is the first practical application of the enthalpy already! It

follows that

$$\boxed{C_p \equiv \left(\frac{\partial h}{\partial T}\right)_P} \tag{11.13}$$

11.8 Intro to the Second Law

Take a look around you. You are surrounded by air molecules. They are all over the place. Isn't that messy? Suppose there would be water all over the room, wouldn't you do something about it? Wouldn't it be much neater to compress all those air atoms together and put them in a glass? (You may want to wear a space suit while doing this.)

The reality, of course, is that if you put all the air atoms in a glass, the high pressure would cause the air to explode out of the glass and it would scatter all over the room again. All your efforts would be for naught. It is like the clothes of a ten-year old. Nature likes messiness. In fact, if messiness is properly defined, and it will be in section 11.10, nature will always increase messiness as much as circumstances and the laws of physics allow. The properly defined messiness is called "entropy." It is not to be confused with enthalpy, which is a completely different concept altogether.

Entropy provides an unrelenting arrow of time. If you take a movie and run it backwards, it simply does not look right, since you notice messiness getting smaller, rather than larger. The movie of a glass of water slipping out of your hand and breaking on the floor becomes, if run backwards, a spill of water and pieces of glass combining together and jumping into your hand. It does not happen. Messiness always increases. Even if you mop up the water and glue the pieces of broken glass back together, it does not work. While you reduce the messiness of the glass of water, you need to perform effort, and it turns out that this always increases messiness elsewhere more than the messiness of the glass of water is reduced.

It has big consequences. Would it not be nice if your car could run without using gas? After all, there is lots of random kinetic energy in the air molecules surrounding your car. Why not scope up some of that kinetic energy out of the air and use it to run your car? It does not work because it would decrease messiness in the universe, that's why. It would turn messy random molecular motion into organized motion of the engine of your car, and nature refuses to do it. And there you have it, the second law of thermodynamics, or at least the version of it given by Kelvin and Planck:

> *You cannot just take random thermal energy out of a substance and turn it into useful work.*

You expected a physical law to be a formula, instead of a verbal statement like that? Well, you are out of luck for now.

To be sure, if the air around your car is hotter than the ground below it, then it *is* possible with some ingenuity to set up a flow of heat from the air to the ground, and you can then divert *some* of this flow of heat and turn it into useful work. But that is not an unlimited free supply of energy; it stops as soon as the temperatures of air and ground have become equal. The temperature difference is an expendable energy source, much like oil in the ground is; you are not simply scooping up random thermal energy out of a substance. If that sounds like a feeble excuse, consider the following: after the temperature difference is gone, the air molecules still have almost exactly the same thermal energy as before, and the ground molecules have more. But you cannot get any of it out anymore as usable energy. Zero. (Practically speaking, the amount of energy you would get out of the temperature difference is not going to get you to work in time anyway, but that is another matter.)

Would it not be nice if your fridge would run without electricity? It would really save on the electricity bill. But it cannot be done; that is the Clausius statement of the second law:

> *You cannot move heat the wrong way, from cold to hot, without doing work.*

It is the same thing as the Kelvin-Planck statement, of course. If you could really have a fridge that ran for free, you could use it to create a temperature difference, and you could use that temperature difference to run your car. So your car would run for free. Conversely, if your car could run for free, you could use the cigarette lighter socket to run your fridge for free.

As patent offices all over the world can confirm, the second law has been solidly verified by countless masses of clever inventors all over the centuries doing everything possible to get around it. All have failed, however ingenious their tricks trying to fool nature. And don't forget about the most brilliant scientists of the last few centuries who have also tried wistfully and failed miserably, usually by trying to manipulate nature on the molecular level. The two verbal statements of the second law may not seem to have much mathematical precision, but they do. If you find a kink in either one's armor, however small, the fabric of current science and technology comes apart. Fabulous riches will be yours, and you will also be the most famous scientist of all time.

11.9 The Reversible Ideal

The statements of the previous section describing the second law are clearly common sense: yes, you still need to plug in your fridge, and no, you cannot skip the periodic stop at a gas station. What a surprise!

They seem to be fairly useless beyond that. For example, they say that it takes electricity to run our fridge, but they do not say it how much. It might be a megawatt, it might be a nanowatt.

Enter human ingenuity. With a some cleverness the two simple statements of the second law can be greatly leveraged, allowing an entire edifice to be constructed upon their basis.

A first insight is that if we are limited by nature's unrelenting arrow of time, then it should pay to study devices that almost ignore that arrow. If you make a movie of a device, and it looks almost exactly right when run backwards, the device is called (almost exactly) "reversible." An example is a mechanism that is carefully designed to move with almost no friction. If set into motion, the motion will slow down only a negligible amount during a short movie. When that movie is run backwards in time, at first glance it seems perfectly fine. If you look more carefully, you will see a slight problem: in the backward movie, the device is speeding up slightly, instead of slowing down due to friction as it should. But it is almost right: it would require only a very small amount of additional energy to speed up the actual device running backwards as it does in the reversed movie.

Dollar signs may come in front of your eyes upon reading that last sentence: it suggest that almost reversible devices may require very little energy to run. In context of the second law it suggests that it may be worthwhile to study refrigeration devices and engines that are almost reversible.

The second major insight is to look where there is light. Why not study, say, a refrigeration device that is simple enough that it can be analyzed in detail? At the very minimum it will give a standard against which other refrigeration devices can be compared. And so it will be done.

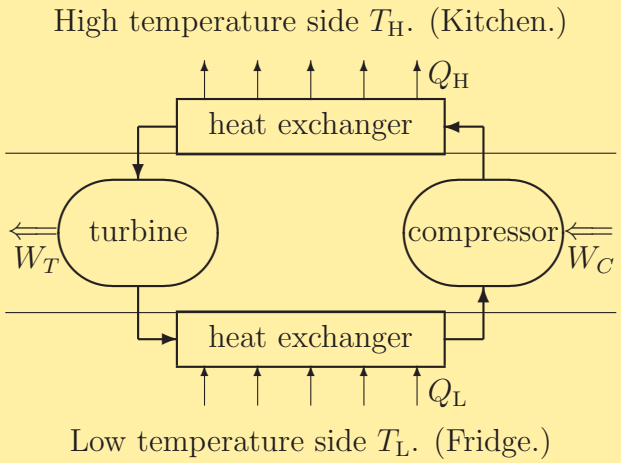


Figure 11.11: Schematic of the Carnot refrigeration cycle.

The theoretically simple refrigeration device is called a "Carnot cycle" refrigeration device, or Carnot heat pump. A schematic is shown in figure 11.11. A substance, the refrigerant, is circulating through four devices, with the objective of transporting heat out of the fridge, dumping it into the kitchen. In the discussed device, the refrigerant will be taken to be some ideal gas with a

constant specific heat like maybe helium. You would not really want to use an ideal gas as refrigerant in a real refrigerator, but the objective here is not to make a practical refrigerator that you can sell for a profit. The purpose here is to create a device that can be analyzed precisely, and an ideal gas is described by simple mathematical formulae discussed in basic physics classes.

Consider the details of the device. The refrigerant enters the fridge at a temperature colder than the inside of the fridge. It then moves through a long piping system, allowing heat to flow out of the fridge into the colder refrigerant inside the pipes. This piping system is called a heat exchanger. The first reversibility problem arises: heat flow is most definitely irreversible. Heat flow seen backwards would be flow from colder to hotter, and that is wrong. The only thing that can be done to minimize this problem as much as possible is to minimize the temperature differences. The refrigerant can be sent in just *slightly* colder than the inside of the fridge. Of course, if the temperature difference is small, the surface through which the heat flows into the refrigerant will have to be very large to take any decent amount of heat away. One impractical aspect of Carnot cycles is that they are huge; that piping system cannot be small. Be that as it may, the theoretical bottom line is that the heat exchange in the fridge can be approximated as (almost) isothermal.

After leaving the inside of the refrigerator, the refrigerant is compressed to increase its temperature to slightly above that of the kitchen. This requires an amount W_C of work to be done, indicating the need for electricity to run the fridge. To avoid irreversible heat conduction in the compression process, the compressor is thermally carefully insulated to eliminate any heat exchange with its surroundings. Also, the compressor is very carefully designed to be almost frictionless. It has expensive bearings that run with almost no friction. Additionally, the refrigerant itself has "viscosity;" it experiences internal friction if there are significant gradients in its velocity. That would make the work required to compress it greater than the ideal $-P\,\mathrm{d}V$, and to minimize that effect, the velocity gradients can be minimized by using lots of refrigerant. This also has the effect of minimizing any internal heat conduction within the refrigerant that may arise. Viscosity is also an issue in the heat exchangers, because the pressure differences cause velocity increases. With lots of refrigerant, the pressure changes over the heat exchangers are also minimized.

Now the refrigerant is sent to a heat exchanger open to the kitchen air. Since it enters slightly hotter than the kitchen, heat will flow out of the refrigerant into the kitchen. Again, the temperature difference must be small for the process to be almost reversible. Finally, the refrigerant is allowed to expand, which reduces its temperature to below that inside the fridge. The expansion occurs within a carefully designed turbine, because the substance does an amount of work W_T while expanding reversibly, and the turbine captures that work. It is used to run a high-quality generator and recover some of the electric power W_C needed to run the compressor. Then the refrigerant reenters the fridge and the

cycle repeats.

If this Carnot refrigerator is analyzed theoretically, {D.59}, a very simple result is found. The ratio of the heat $Q_{\rm H}$ dumped by the device into the kitchen to the heat $Q_{\rm L}$ removed from the refrigerator is exactly the same as the ratio of the temperature of the kitchen $T_{\rm H}$ to that of the fridge $T_{\rm L}$:

$$\boxed{\text{For an ideal cycle: } \frac{Q_{\rm H}}{Q_{\rm L}} = \frac{T_{\rm H}}{T_{\rm L}}} \tag{11.14}$$

That is a very useful result, because the net work $W = W_C - W_T$ that must go into the device is, by conservation of energy, the difference between $Q_{\rm H}$ and $Q_{\rm L}$. A "coefficient of performance" can be defined that is the ratio of the heat $Q_{\rm L}$ removed from the fridge to the required power input W:

$$\boxed{\text{For an ideal refrigeration cycle: } \beta \equiv \frac{Q_{\rm L}}{W} = \frac{T_{\rm L}}{T_{\rm H} - T_{\rm L}}} \tag{11.15}$$

Actually, some irreversibility is unavoidable in real life, and the true work required will be more. The formula above gives the required work if everything is truly ideal.

The same device can be used in winter to *heat* the inside of your house. Remember that heat was dumped *into* the kitchen. So, just cross out "kitchen" at the high temperature side in figure 11.11 and write in "house." And cross out "fridge" and write in "outside." The device removes heat from the outside and dumps it into your house. It is the exact same device, but it is used for a different *purpose*. That is the reason that it is no longer called a "refrigeration cycle" but a "heat pump." For an heat pump, the quantity of interest is the amount of heat dumped at the *high* temperature side, into your house. So an alternate coefficient of performance is now defined as

$$\boxed{\text{For an ideal heat pump: } \beta' \equiv \frac{Q_{\rm H}}{W} = \frac{T_{\rm H}}{T_{\rm H} - T_{\rm L}}} \tag{11.16}$$

The formula above is ideal. Real-life performance will be less, so the work required will be more.

It is interesting to note that if you take an amount W of electricity and dump it into a simple resistance heater, it adds exactly an amount W of heat to your house. If you dump that same amount of electricity into a Carnot heat pump that uses it to pump in heat from the outside, the amount of heat added to your house will be much larger than W. For example, if it is 300 K (27 °C) inside and 275 K (2 °C) outside, the amount of heat added is $300/25 = 12$ W, twelve times the amount you got from the resistance heater!

If you run the Carnot refrigeration cycle in reverse, as in figure 11.12, all arrows reverse and it turns into a "heat engine." The device now takes in

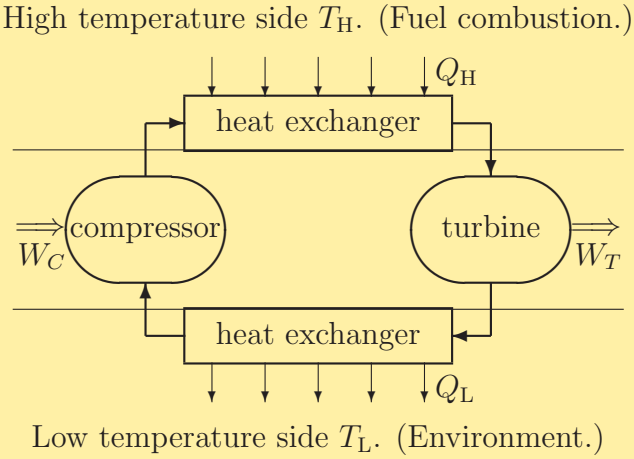


Figure 11.12: Schematic of the Carnot heat engine.

heat at the high temperature side and outputs a net amount of work. The high temperature side is the place where you are burning the fuel. The low temperature may be cooling water from the local river. The Kelvin-Planck statement says that the device will not run unless some of the heat from the combustion is dumped to a lower temperature. In a car engine, the exhaust and radiator are the ones that take much of the heat away. Since the device is almost reversible, the numbers for transferred heats and net work do not change much from the nonreversed version. But the purpose is now to create work, so the "thermal efficiency" of a heat engine is defined as

$$\boxed{\text{For an ideal heat engine: } \eta_{\rm th} \equiv \frac{W}{Q_{\rm H}} = \frac{T_{\rm H} - T_{\rm L}}{T_{\rm H}}} \tag{11.17}$$

Unfortunately, this is always less than one. And to get close to that, the engine must operate hot; the temperature at which the fuel is burned must be very hot.

(Note that slight corrections to the strictly reversed refrigeration process are needed; in particular, for the heat engine process to work, the substance must now be slightly colder than $T_{\rm H}$ at the high temperature side, and slightly hotter than $T_{\rm L}$ at the low temperature side. Heat cannot flow from colder to hotter. But since these are small changes, the mathematics is almost the same. In particular, the numerical values for $Q_{\rm H}$ and $Q_{\rm L}$ will be almost unchanged, though the heat now goes the opposite way.)

The final issue to be resolved is whether other devices could not be better than the Carnot ones. For example, could not a generic heat pump be more efficient than the reversible Carnot version in heating a house? Well, put them into different windows, and see. (The Carnot one will need the big window.) Assume that both devices are sized to produce the same heat flow into the house. On second thought, since the Carnot machine is reversible, run it in

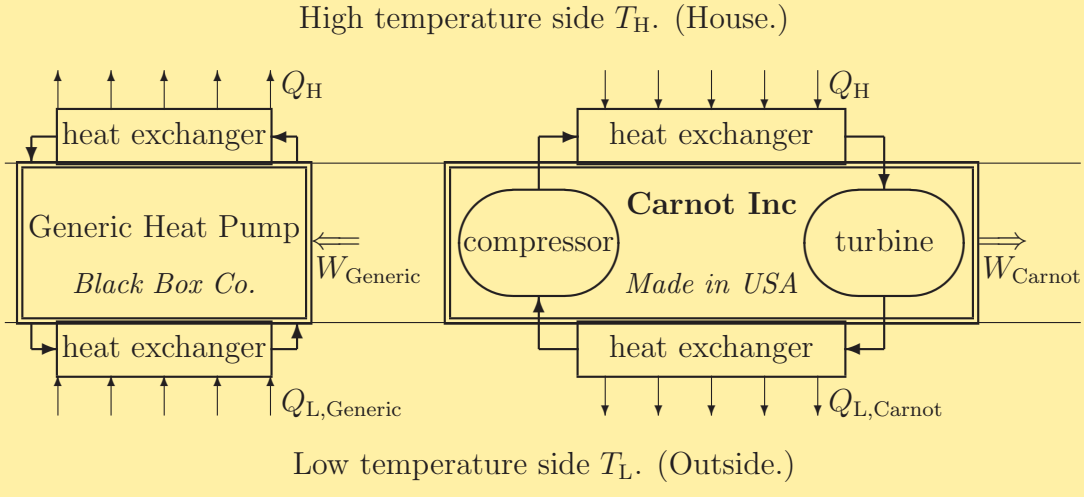


Figure 11.13: A generic heat pump next to a reversed Carnot one with the same heat delivery.

reverse; that can be done without changing its numbers for the heat fluxes and net work noticeably, and it will show up the *differences* between the devices.

The idea is shown in figure 11.13. Note that the net heat flow into the house is now zero, confirming that running the Carnot in reverse really shows the differences between the devices. Net heat is exchanged with the outside air and there is net work. Enter Kelvin-Planck. According to Kelvin-Planck, heat cannot simply be taken out of the outside air and converted into useful net work. The net work being taken out of the air will have to be negative. So the work required for the generic heat pump will need to be greater than that recovered by the reversed Carnot one, the excess ending up as heat in the outside air. So, the generic heat pump requires more work than a Carnot one running normally. No device can therefore be more efficient than the Carnot one. The best case is that the generic device, too, is reversible. In that case, neither device can win, because the generic device can be made to run in reverse instead of the Carnot one. That is the case where both devices are so perfectly constructed that whatever work goes into the generic device is almost 100% recovered by the reversed Carnot machine, with negligible amounts of work being turned into heat by friction or other irreversibility and ending up in the outside air.

The conclusion is that:

> *All reversible devices exchanging heat at a given high temperature T_H and low temperature T_L, (and nowhere else,) have the same efficiency. Irreversible devices have less.*

To see that it is true for refrigeration cycles too, just note that because of conservation of energy, $Q_L = Q_H - W$. It follows that, considered as a refrigeration cycle, not only does the generic heat pump above require more work, it also

removes less heat from the cold side. To see that it applies to heat engines too, just place a generic heat engine next to a reversed Carnot one producing the same power. The net work is then zero, and the heat flow $Q_{\rm H}$ of the generic device better be greater than that of the Carnot cycle, because otherwise net heat would flow from cold to hot, violating the Clausius statement. The heat flow $Q_{\rm H}$ is a measure of the amount of fuel burned, so the irreversible generic device uses more fuel.

Practical devices may exchange heat at more than two temperatures, and can be compared to a set of Carnot cycles doing the same. It is then seen that it is bad news; for maximum theoretical efficiency of a heat engine, you prefer to exchange heat at the highest available temperature and the lowest available temperature, and for heat pumps and refrigerators, at the lowest available high temperature and the highest available low temperature. But real-life and theory are of course not the same.

Since the efficiency of the Carnot cycle has a unique relation to the temperature ratio between the hot and cold sides, it is possible to *define* the temperature scale using the Carnot cycle. The only thing it takes is to select a single reference temperature to compare with, like water at its triple point. This was in fact proposed by Kelvin as a conceptual definition, to be contrasted with earlier definitions based on thermometers containing mercury or a similar fluid whose volume expansion is read-off. While a substance like mercury expands in volume very much linearly with the (Kelvin) temperature, it does not expand *exactly* linearly with it. So slight variations in temperature would occur based on which substance is arbitrarily selected for the reference thermometer. On the other hand, the second law requires that all substances used in the Carnot cycle will give the same Carnot temperature, with no deviation allowed. It may be noted that the definition of temperature used in this chapter is completely consistent with the Kelvin one, because "all" substances includes ideal gasses.

11.10 Entropy

With the cleverest inventors and the greatest scientists relentlessly trying to fool nature and circumvent the second law, how come nature never once gets confused, not even by the most complicated, convoluted, unusual, ingenious schemes? Nature does not outwit them by out-thinking them, but by maintaining an accounting system that cannot be fooled. Unlike human accounting systems, this accounting system does not assign a monetary value to each physical system, but a measure of messiness called "entropy." Then, in any transaction within or between systems, nature simply makes sure that this entropy is not being reduced; whatever entropy one system gives up must always be less than what the other system receives.

So what can this numerical grade of messiness called entropy be? Surely, it

must be related somehow to the second law as stated by Clausius and Kelvin and Planck, and to the resulting Carnot engines that cannot be beat. Note that the Carnot engines relate heat added to temperature. In particular an infinitesimally small Carnot engine would take in an infinitesimal amount $\delta Q_{\rm H}$ of heat at a temperature $T_{\rm H}$ and give up an infinitesimal amount $\delta Q_{\rm L}$ at a temperature $T_{\rm L}$. This is done so that $\delta Q_{\rm H}/\delta Q_{\rm L} = T_{\rm H}/T_{\rm L}$, or separating the two ends of the device, $\delta Q_{\rm H}/T_{\rm H} = \delta Q_{\rm L}/T_{\rm L}$. The quantity $\delta Q/T$ is the same at both sides, except that one is going in and the other out. Might this, then, be the change in messiness? After all, for the ideal reversible machine no messiness can be created, otherwise in the reversed process, messiness would be reduced. Whatever increase in messiness one side receives, the other side must give up, and $\delta Q/T$ fits the bill for that.

If $\delta Q/T$ gives the infinitesimal change in messiness, excuse, entropy, then it should be possible to find the entropy of a system by integration. In particular, choosing some arbitrary state of the system as reference, the entropy of a system in thermal equilibrium can be found as:

$$\boxed{S \equiv S_{\rm ref} + \int_{\text{reference state}}^{\text{desired state}} \frac{\delta Q}{T} \qquad \text{along any reversible path}} \tag{11.18}$$

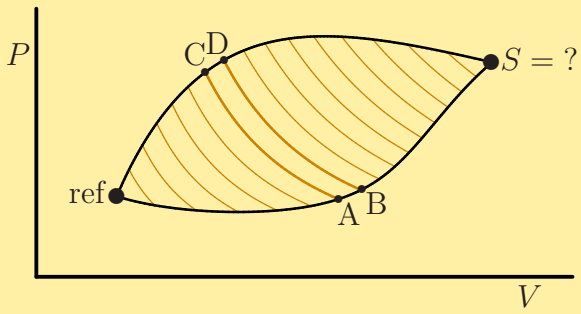


Figure 11.14: Comparison of two different integration paths for finding the entropy of a desired state. The two different integration paths are in black and the yellow lines are reversible adiabatic process lines.

The entropy as defined above is a specific number for a system in thermal equilibrium, just like its pressure, temperature, particle density, and internal energy are specific numbers. You might think that you could get a different value for the entropy by following a different process path from the reference state to the desired state. But the second law prevents that. To see why, consider the pressure-volume diagram in figure 11.14. Two different reversible processes are shown leading from the reference state to a desired state. A bundle of reversible adiabatic process lines is also shown; those are graphical representations of processes in which there is no heat exchange between the system and its surroundings. The bundle of adiabatic lines chops the two process

paths into small pieces, of almost constant temperature, that pairwise have the same value of $\delta Q/T$. For, if a piece like AB would have a lower value for $\delta Q/T$ than the corresponding piece CD, then a heat engine running the cycle CDBAC would lose less of the heat $\delta Q_{\rm H}$ at the low temperature side than the Carnot ideal, hence have a higher efficiency than Carnot and that is not possible. Conversely, if AB would have a higher value for $\delta Q/T$ than CD, then a refrigeration device running the cycle ABDCA would remove more heat from the low side than Carnot, again not possible. So all the little segments pairwise have the same value for $\delta Q/T$, which means the complete integrals must also be the same. It follows that the entropy for a system in thermal equilibrium is uniquely defined.

So what happens if the reference and final states are still the same, but there is a slight glitch for a single segment AB, making the process over that one segment irreversible? In that case, the heat engine argument no longer applies, since it runs through the segment AB in reversed order, and irreversible processes cannot be reversed. The refrigeration cycle argument says that the amount of heat δQ absorbed by the system will be less; more of the heat δQ going out at the high temperature side CD will come from the work done, and less from the heat removed at the cold side. The final entropy is still the same, because it only depends on the final state, not on the path to get there. So during the slight glitch, the entropy of the system increased more than $\delta Q/T$. In general:

$$\boxed{\mathrm{d}S \geqslant \frac{\delta Q}{T}} \tag{11.19}$$

where $=$ applies if the change is reversible and $>$ if it is not.

Note that the above formula is only valid if the system has an definite temperature, as in this particular example. Typically this is simply not true in irreversible processes; for example, the interior of the system might be hotter than the outside. The real importance of the above formula is to confirm that the defined entropy is indeed a measure of messiness and not of order; reversible processes merely shuffle entropy around from one system to the next, but irreversible processes *increase* the net entropy content in the universe.

So what about the entropy of a system that is not in thermal equilibrium? Equation (11.18) only applies for systems in thermal equilibrium. In order for nature not to become confused in its entropy accounting system, surely entropy must still have a numerical value for nonequilibrium systems. If the problem is merely temperature or pressure variations, where the system is still in approximate thermal equilibrium locally, you could just integrate the entropy per unit volume over the volume. But if the system is not in thermal equilibrium even on macroscopically small scales, it gets much more difficult. For example, air crossing a typical shock wave (sonic boom) experiences a significant increase in pressure over an extremely short distance. Better bring out the quantum

mechanics trick box. Or at least molecular dynamics.

Still, some important general observations can be made without running to a computer. An "isolated" system is a system that does not interact with its surroundings in any way. Remember the example where the air inside a room was collected and neatly put inside a glass? That was an example of an isolated system. Presumably, the doors of the room were hermetically sealed. The walls of the room are stationary, so they do not perform work on the air in the room. And the air comes rushing back out of the glass so quickly that there is really no time for any heat conduction through the walls. If there is no heat conduction with the outside, then there is no entropy exchange with the outside. So the entropy of the air can only increase due to irreversible effects. And that is exactly what happens: the air exploding out of the glass is highly irreversible, (no, it has no plans to go back in), and its entropy increases rapidly. Quite quickly however, the air spreads again out over the entire room and settles down. Beyond that point, the entropy remains further constant.

> *An isolated system evolves to the state of maximum possible entropy and then stays there.*

The state of maximum possible entropy is the thermodynamically stable state a system will assume if left alone.

A more general system is an "adiabatic" or "insulated" system. Work may be performed on such a system, but there is still no heat exchange with the surroundings. That means that the entropy of such a system can again only increase due to reversibility. A simple example is a thermos bottle with a cold drink inside. If you continue shaking this thermos bottle violently, the cold drink will heat up due to its viscosity, its internal friction, and it will not stay a cold drink for long. Its entropy will increase while you are shaking it.

> *The entropy of adiabatic systems can only increase.*

But, of course, that of an open system may not. It is the recipe of life, {N.25}.

You might wonder why this book on quantum mechanics included a concise, but still very lengthy classical description of the second law. It is because the evidence for the second law is so much more convincing based on the macroscopic evidence than on the microscopic one. Macroscopically, the most complex systems can be accurately observed, microscopically, the quantum mechanics of only the most simplistic systems can be rigorously solved. And whether we can observe the solution is still another matter.

However, given the macroscopic fact that there really is an accounting measure of messiness called entropy, the question becomes what is its actual microscopic nature? Surely, it must have a relatively simple explanation in terms of the basic microscopic physics? For one, nature never seems to get confused about what it is, and for another, you really would expect something that is

clearly so fundamental to nature to be relatively esthetic when expressed in terms of mathematics.

And that thought is all that is needed to *guess* the true microscopic nature of entropy. And guessing is good, because it gives a lot of insight why entropy is what it is. And to ensure that the final result is really correct, it can be cross checked against the macroscopic definition (11.18) and other known facts about entropy.

The first guess is about what physical microscopic quantity would be involved. Now microscopically, a simple system is described by energy eigenfunctions $\psi_q^{\rm S}$, and there is nothing messy about those. They are the systematic solutions of the Hamiltonian eigenvalue problem. But these eigenfunctions have probabilities P_q, being the square magnitudes of their coefficients, and they are a different story. A system of a given energy could in theory exist neatly as a single energy eigenfunction with that energy. But according to the fundamental assumption of quantum statistics, this simply does not happen. In thermal equilibrium, every single energy eigenfunction of the given energy achieves about the same probability. Instead of nature neatly leaving the system in the single eigenfunction it may have started out with, it gives every Johnny-come-lately state about the same probability, and it becomes a mess.

If the system is in a single eigenstate for sure, the probability P_q of that one eigenstate is one, and all others are zero. But if the probabilities are equally spread out over a large number, call it N, of eigenfunctions, then each eigenfunction receives a probability $P_q = 1/N$. So your simplest thought would be that maybe entropy is the average value of the probability. In particular, just like the average energy is $\sum P_q E_q^{\rm S}$, the average probability would be $\sum P_q^2$. It is always the sum of the values for which you want the average times their probability. You second thought would be that since $\sum P_q^2$ is one for the single eigenfunction case, and $1/N$ for the spread out case, maybe the entropy should be $-\sum P_q^2$ in order that the single eigenfunction case has the lower value of messiness. But macroscopically it is known that you can keep increasing entropy indefinitely by adding more and more heat, and the given expression starts at minus one and never gets above zero.

So try a slightly more general possibility, that the entropy is the average of some function of the probability, as in $S = \sum P_q f(P_q)$. The question is then, what function? Well, macroscopically it is also known that entropy is additive, the values of the entropies of two systems simply add up. It simplifies nature's task of maintaining a tight accounting system on messiness. For two systems with probabilities P_q and P_r,

$$S = \sum_q P_q f(P_q) + \sum_r P_r f(P_r)$$

This can be rewritten as

$$S = \sum_q \sum_r P_q P_r f(P_q) + \sum_q \sum_r P_q P_r f(P_r).$$

since probabilities by themselves must sum to one. On the other hand, if you combine two systems, the probabilities multiply, just like the probability of throwing a 3 with your red dice and a 4 with your black dice is $\frac{1}{6} \times \frac{1}{6}$. So the combined entropy should also be equal to

$$S = \sum_q \sum_r P_q P_r f(P_q P_r)$$

Comparing this with the previous equation, you see that $f(P_q P_r)$ must equal $f(P_q) + f(P_r)$. The function that does that is the logarithmic function. More precisely, you want minus the logarithmic function, since the logarithm of a small probability is a large negative number, and you need a large positive messiness if the probabilities are spread out over a large number of states. Also, you will need to throw in a factor to ensure that the units of the microscopically defined entropy are the same as the ones in the macroscopical definition. The appropriate factor turns out to be the Boltzmann constant $k_B = 1.380\,65\ 10^{-23}$ J/K; note that this factor has absolutely no effect on the physical meaning of entropy; it is just a matter of agreeing on units.

The microscopic definition of entropy has been guessed:

$$\boxed{S = -k_B \sum P_q \ln(P_q)} \tag{11.20}$$

That wasn't too bad, was it?

At absolute zero temperature, the system is in the ground state. That means that probability P_q of the ground state is 1 and all other probabilities are zero. Then the entropy is zero, because $\ln(1) = 0$. The fact that the entropy is zero at absolute zero is known as the "third law of thermodynamics," {A.35}.

At temperatures above absolute zero, many eigenfunctions will have nonzero probabilities. That makes the entropy positive, because logarithms of numbers less than one are negative. (It should be noted that $P_q \ln P_q$ becomes zero when P_q becomes zero; the blow up of $\ln P_q$ is no match for the reduction in magnitude of P_q. So highly improbable states will not contribute significantly to the entropy despite their relatively large values of the logarithm.)

To put the definition of entropy on a less abstract basis, assume that you schematize the system of interest into unimportant eigenfunctions that you give zero probability, and a remaining N important eigenfunctions that all have the same average probability $1/N$. Sure, it is crude, but it is just to get an idea. In this simple model, the entropy is $k_B \ln(N)$, proportional to the logarithm of the number of quantum states that have an important probability. The more

states, the higher the entropy. This is what you will find in popular expositions. And it would actually be correct for systems with zero indeterminacy in energy, if they existed.

The next step is to check the expression. Derivations are given in {D.60}, but here are the results. For systems in thermal equilibrium, is the entropy the same as the one given by the classical integration (11.18)? Check. Does the entropy exist even for systems that are not in thermal equilibrium? Check, quantum mechanics still applies. For a system of given energy, is the entropy smallest when the system is in a single energy eigenfunction? Check, it is zero then. For a system of given energy, is the entropy the largest when all eigenfunctions of that energy have the same probability, as the fundamental assumption of quantum statistics suggests? Check. For a system with given expectation energy but uncertainty in energy, is the entropy highest when the probabilities are given by the canonical probability distribution? Check. For two systems in thermal contact, is the entropy greatest when their temperatures have become equal? Check.

Feynman [18, p. 8] gives an argument to show that the entropy of an isolated system always increases with time. Taking the time derivative of (11.20),

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -k_{\mathrm{B}} \sum_q [\ln(P_q) + 1] \frac{\mathrm{d}P_q}{\mathrm{d}t} = -k_{\mathrm{B}} \sum_q \sum_r [\ln(P_q) + 1] R_{qr} [P_r - P_q],$$

the final equality being from time-dependent perturbation theory, with $R_{qr} = R_{rq} > 0$ the transition rate from state q to state p. In the double summation, a typical term with indices q and r combines with the term having the reversed indices as

$$k_{\mathrm{B}} [\ln(P_r) + 1 - \ln(P_q) - 1] R_{qr} [P_r - P_q]$$

and that is always greater that zero because the terms in the square brackets have the same sign: if P_q is greater/less than P_r then so is $\ln(P_q)$ greater/less than $\ln(P_r)$. However, given the dependence of time-dependent perturbation theory on linearization and worse, the "measurement" wild card, chapter 7.6 you might consider this more a validation of time dependent perturbation theory than of the expression for entropy. Then there is the problem of ensuring that a perturbed and measured system is adiabatic.

In any case, it may be noted that the checks on the expression for entropy, as given above, cut both ways. If you accept the expression for entropy, the canonical probability distribution follows. They are consistent, and in the end, it is just a matter of which of the two postulates you are more willing to accept as true.

11.11 The Big Lie of Distinguishable Particles

If you try to find the entropy of the system of distinguishable particles that produces the Maxwell-Boltzmann distribution, you are in for an unpleasant surprise. It just cannot be done. The problem is that the number of eigenfunctions for I distinguishable particles is typically roughly $I!$ larger than for I identical bosons or fermions. If the typical number of states becomes larger by a factor $I!$, the logarithm of the number of states increases by $I \ln I$, (using the Stirling formula), which is no longer proportional to the size of the system I, but much larger than that. The specific entropy would blow up with system size.

What gives? Now the truth must be revealed. The entire notion of distinguishable particles is a blatant lie. You are simply not going to have 10^{23} distinguishable particles in a box. Assume they would be 10^{23} different molecules. It would a take a chemistry handbook of 10^{21} pages to list them, one line for each. Make your system size 1 000 times as big, and the handbook gets 1 000 times thicker still. That would be really messy! When identical bosons or fermions are far enough apart that their wave functions do no longer overlap, the symmetrization requirements are no longer important for most practical purposes. But if you start counting energy eigenfunctions, as entropy does, it is a different story. Then there is no escaping the fact that the particles really are, after all, indistinguishable forever.

11.12 The New Variables

The new kid on the block is the entropy S. For an adiabatic system the entropy is always increasing. That is highly useful information, if you want to know what thermodynamically stable final state an adiabatic system will settle down into. No need to try to figure out the complicated time evolution leading to the final state. Just find the state that has the highest possible entropy S, that will be the stable final state.

But a lot of systems of interest are not well described as being adiabatic. A typical alternative case might be a system in a rigid box in an environment that is big enough, and conducts heat well enough, that it can at all times be taken to be at the same temperature T_{surr}. Also assume that initially the system itself is in some state 1 at the ambient temperature T_{surr}, and that it ends up in a state 2 again at that temperature. In the evolution from 1 to 2, however, the system temperature could be be different from the surroundings, or even undefined, no thermal equilibrium is assumed. The first law, energy conservation, says that the heat Q_{12} added to the system from the surroundings equals the change in internal energy $E_2 - E_1$ of the system. Also, the entropy change in the isothermal environment will be $-Q_{12}/T_{\text{surr}}$, so the system entropy

change $S_2 - S_1$ must be at least $Q_{12}/T_{\rm surr}$ in order for the net entropy in the universe not to decrease. From that it can be seen by simply writing it out that the "Helmholtz free energy"

$$\boxed{F = E - TS} \tag{11.21}$$

is smaller for the final system 2 than for the starting one 1. In particular, if the system ends up into a stable final state that can no longer change, it will be the state of smallest possible Helmholtz free energy. So, if you want to know what will be the final fate of a system in a rigid, heat conducting, box in an isothermal environment, just find the state of lowest possible Helmholtz energy. That will be the one.

A slightly different version occurs even more often in real applications. In these the system is not in a rigid box, but instead its surface is at all times exposed to ambient atmospheric pressure. Energy conservation now says that the heat added Q_{12} equals the change in internal energy $E_2 - E_1$ *plus* the work done expanding against the atmospheric pressure, which is $P_{\rm surr}(V_2 - V_1)$. Assuming that both the initial state 1 and final state 2 are at ambient atmospheric pressure, as well as at ambient temperature as before, then it is seen that the quantity that decreases is the "Gibbs free energy"

$$\boxed{G = H - TS} \tag{11.22}$$

in terms of the enthalpy H defined as $H = E + PV$. As an example, phase equilibria are at the same pressure and temperature. In order for them to be stable, the phases need to have the same specific Gibbs energy. Otherwise all particles would end up in whatever phase has the lower Gibbs energy. Similarly, chemical equilibria are often posed at an ambient pressure and temperature.

There are a number of differential expressions that are very useful in doing thermodynamics. The primary one is obtained by combining the differential first law (11.11) with the differential second law (11.19) for reversible processes:

$$\boxed{\mathrm{d}E = T\,\mathrm{d}S - P\,\mathrm{d}V} \tag{11.23}$$

This no longer involves the heat transferred from the surroundings, just state variables of the system itself. The equivalent one using the enthalpy H instead of the internal energy E is

$$\boxed{\mathrm{d}H = T\,\mathrm{d}S + V\,\mathrm{d}P} \tag{11.24}$$

The differentials of the Helmholtz and Gibbs free energies are, after cleaning up with the two expressions immediately above:

$$\boxed{\mathrm{d}F = -S\,\mathrm{d}T - P\,\mathrm{d}V} \tag{11.25}$$

and

$$\boxed{\mathrm{d}G = -S\,\mathrm{d}T + V\,\mathrm{d}P} \tag{11.26}$$

Expression (11.25) shows that the work obtainable in an isothermal reversible process is given by the decrease in Helmholtz free energy. That is why Helmholtz called it "free energy" in the first place. The Gibbs free energy is applicable to steady flow devices such as compressors and turbines; the first law for these devices must be corrected for the "flow work" done by the pressure forces on the substance entering and leaving the device. The effect is to turn $P\,\mathrm{d}V$ into $-V\,\mathrm{d}P$ as the differential for the actual work obtainable from the device. (This assumes that the kinetic and/or potential energy that the substance picks up while going through the device is a not a factor.)

Maxwell noted that, according to the total differential of calculus, the coefficients of the differentials in the right hand sides of (11.23) through (11.26) must be the partial derivatives of the quantity in the left hand side:

$$\left(\frac{\partial E}{\partial S}\right)_V = T \qquad \left(\frac{\partial E}{\partial V}\right)_S = -P \qquad \left(\frac{\partial T}{\partial V}\right)_S = -\left(\frac{\partial P}{\partial S}\right)_V \tag{11.27}$$

$$\left(\frac{\partial H}{\partial S}\right)_P = T \qquad \left(\frac{\partial H}{\partial P}\right)_S = V \qquad \left(\frac{\partial T}{\partial P}\right)_S = \left(\frac{\partial V}{\partial S}\right)_P \tag{11.28}$$

$$\left(\frac{\partial F}{\partial T}\right)_V = -S \qquad \left(\frac{\partial F}{\partial V}\right)_T = -P \qquad \left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial P}{\partial T}\right)_V \tag{11.29}$$

$$\left(\frac{\partial G}{\partial T}\right)_P = -S \qquad \left(\frac{\partial G}{\partial P}\right)_T = V \qquad \left(\frac{\partial S}{\partial P}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_P \tag{11.30}$$

The final equation in each line can be verified by substituting in the previous two and noting that the order of differentiation does not make a difference. Those are called the "Maxwell relations." They have a lot of practical uses. For example, either of the final equations in the last two lines allows the entropy to be found if the relationship between the "normal" variables P, V, and T is known, assuming that at least one data point at every temperature is already available. Even more important from an applied point of view, the Maxwell relations allow whatever data you find about a substance in literature to be stretched thin. Approximate the derivatives above with difference quotients, and you can compute a host of information not initially in your table or graph.

There are two even more remarkable relations along these lines. They follow from dividing (11.23) and (11.24) by T and rearranging so that S becomes the

quantity differentiated. That produces

$$\left(\frac{\partial S}{\partial T}\right)_V = \frac{1}{T}\left(\frac{\partial E}{\partial T}\right)_V \qquad \left(\frac{\partial S}{\partial V}\right)_T = \frac{1}{T}\left(\frac{\partial E}{\partial V}\right)_T + \frac{P}{T}$$
$$\left(\frac{\partial E}{\partial V}\right)_T = T^2\left(\frac{\partial P/T}{\partial T}\right)_V \tag{11.31}$$

$$\left(\frac{\partial S}{\partial T}\right)_P = \frac{1}{T}\left(\frac{\partial H}{\partial T}\right)_P \qquad \left(\frac{\partial S}{\partial P}\right)_T = \frac{1}{T}\left(\frac{\partial H}{\partial P}\right)_T - \frac{V}{T}$$
$$\left(\frac{\partial H}{\partial P}\right)_T = -T^2\left(\frac{\partial V/T}{\partial T}\right)_P \tag{11.32}$$

What is so remarkable is the final equation in each case: they do not involve entropy in any way, just the "normal" variables P, V, T, H, and E. Merely because entropy *exists*, there must be relationships between these variables which seemingly have absolutely nothing to do with the second law.

As an example, consider an ideal gas, more precisely, any substance that satisfies the ideal gas law

$$\boxed{Pv = RT \quad \text{with} \quad R = \frac{k_\mathrm{B}}{m} = \frac{R_\mathrm{u}}{M} \quad R_\mathrm{u} = 8.314\,472\ \mathrm{kJ/kmol\ K}} \tag{11.33}$$

The constant R is called the specific gas constant; it can be computed from the ratio of the Boltzmann constant k_B and the mass of a single molecule m. Alternatively, it can be computed from the "universal gas constant" $R_\mathrm{u} = I_\mathrm{A} k_\mathrm{B}$ and the molar mass $M = I_\mathrm{A} m$. For an ideal gas like that, the equations above show that the internal energy and enthalpy are functions of temperature only. And then so are the specific heats C_v and C_p, because those are their temperature derivatives:

$$\boxed{\text{For ideal gases:} \quad e, h, C_v, C_p = e, h, C_v, C_p(T) \quad C_P = C_v + R} \tag{11.34}$$

(The final relation is because $C_P = \mathrm{d}h/\mathrm{d}T = \mathrm{d}(e + Pv)/\mathrm{d}T$ with $\mathrm{d}e/\mathrm{d}T = C_v$ and $Pv = RT$.) Ideal gas tables can therefore be tabulated by temperature only, there is no need to include a second independent variable. You might think that entropy should be tabulated against both varying temperature and varying pressure, because it does depend on both pressure and temperature. However, the Maxwell equation (11.30) may be used to find the entropy at any pressure as long as it is listed for just one pressure, say for one bar.

There is a sleeper among the Maxwell equations; the very first one, in (11.27). Turned on its head, it says that

$$\boxed{\frac{1}{T} = \left(\frac{\partial S}{\partial E}\right)_{V \text{ and other external parameters fixed}}} \tag{11.35}$$

This can be used as a *definition* of temperature. Note that in taking the derivative, the volume of the box, the number of particles, and other external parameters, like maybe an external magnetic field, must be held constant. To understand qualitatively why the above derivative defines a temperature, consider two systems A and B for which A has the larger temperature according to the definition above. If these two systems are brought into thermal contact, then net messiness increases when energy flows from high temperature system A to low temperature system B, because system B, with the higher value of the derivative, increases its entropy more than A decreases its.

Of course, this new definition of temperature is completely consistent with the ideal gas one; it was derived from it. However, the new definition also works fine for negative temperatures. Assume a system A has a negative temperature according to he definition above. Then its messiness (entropy) increases if it *gives up* heat. That is in stark contrast to normal substances at positive temperatures that increase in messiness if they *take in* heat. So assume that system A is brought into thermal contact with a normal system B at a positive temperature. Then A will give off heat to B, and both systems increase their messiness, so everyone is happy. It follows that A will give off heat however hot is the normal system it is brought into contact with. While the temperature of A may be negative, it is hotter than any substance with a normal positive temperature!

And now the big question: what is that "chemical potential" you hear so much about? Nothing new, really. For a pure substance with a single constituent like this chapter is supposed to discuss, the chemical potential is just the specific Gibbs free energy on a molar basis, $\bar{\mu} = \bar{g}$. More generally, if there is more than one constituent the chemical potential $\bar{\mu}_c$ of each constituent c is best defined as

$$\boxed{\bar{\mu}_c \equiv \left(\frac{\partial G}{\partial \bar{\imath}_c}\right)_{P,T}} \tag{11.36}$$

(If there is only one constituent, then $G = \bar{\imath}\bar{g}$ and the derivative does indeed produce $\bar{g}$. Note that an intensive quantity like $\bar{g}$, when considered to be a function of P, T, and $\bar{\imath}$, only depends on the two intensive variables P and T, not on the amount of particles $\bar{\imath}$ present.) If there is more than one constituent, and assuming that their Gibbs free energies simply add up, as in

$$G = \bar{\imath}_1\bar{g}_1 + \bar{\imath}\bar{g}_2 + \ldots = \sum_c \bar{\imath}_c\bar{g}_c,$$

then the chemical potential $\bar{\mu}_c$ of each constituent is simply the molar specific Gibbs free energy $\bar{g}_c$ of that constituent,

The partial derivatives described by the chemical potentials are important for figuring out the stable equilibrium state a system will achieve in an isothermal, isobaric, environment, i.e. in an environment that is at constant temperature

and pressure. As noted earlier in this section, the Gibbs free energy must be as small as it can be in equilibrium at a given temperature and pressure. Now according to calculus, the full differential for a change in Gibbs free energy is

$$\mathrm{d}G(P,T,\bar{\imath}_1,\bar{\imath}_2,\ldots) = \frac{\partial G}{\partial T}\,\mathrm{d}T + \frac{\partial G}{\partial P}\,\mathrm{d}P + \frac{\partial G}{\partial \bar{\imath}_1}\,\mathrm{d}\bar{\imath}_1 + \frac{\partial G}{\partial \bar{\imath}_2}\,\mathrm{d}\bar{\imath}_2 + \ldots$$

The first two partial derivatives, which keep the number of particles fixed, were identified in the discussion of the Maxwell equations as $-S$ and V; also the partial derivatives with respect to the numbers of particles of the constituent have been defined as the chemical potentials $\bar{\mu}_c$. Therefore more shortly,

$$\boxed{\mathrm{d}G = -S\,\mathrm{d}T + V\,\mathrm{d}P + \bar{\mu}_1\,\mathrm{d}\bar{\imath}_1 + \bar{\mu}_2\,\mathrm{d}\bar{\imath}_2 + \ldots = -S\,\mathrm{d}T + V\,\mathrm{d}P + \sum_c \bar{\mu}_c\,\mathrm{d}\bar{\imath}_c} \tag{11.37}$$

This generalizes (11.26) to the case that the numbers of constituents change. At equilibrium at given temperature and pressure, the Gibbs energy must be minimal. It means that $\mathrm{d}G$ must be zero whenever $\mathrm{d}T = \mathrm{d}P = 0$, regardless of any infinitesimal changes in the amounts of the constituents. That gives a condition on the fractions of the constituents present.

Note that there are typically constraints on the changes $\mathrm{d}\bar{\imath}_c$ in the amounts of the constituents. For example, in a liquid-vapor "phase equilibrium," any additional amount of particles $\mathrm{d}\bar{\imath}_\mathrm{f}$ that condenses to liquid must equal the amount $-\mathrm{d}\bar{\imath}_\mathrm{g}$ of particles that disappears from the vapor phase. (The subscripts follow the unfortunate convention liquid=fluid=f and vapor=gas=g. Don't ask.) Putting this relation in (11.37) it can be seen that the liquid and vapor phase must have the same chemical potential, $\bar{\mu}_\mathrm{f} = \bar{\mu}_\mathrm{g}$. Otherwise the Gibbs free energy would get smaller when more particles enter whatever is the phase of lowest chemical potential and the system would collapse completely into that phase alone.

The equality of chemical potentials suffices to derive the famous Clausius-Clapeyron equation relating pressure changes under two-phase, or "saturated," conditions to the corresponding temperature changes. For, the changes in chemical potentials must be equal too, $\mathrm{d}\mu_\mathrm{f} = \mathrm{d}\mu_\mathrm{g}$, and substituting in the differential (11.26) for the Gibbs free energy, taking it on a molar basis since $\bar{\mu} = \bar{g}$,

$$-\bar{s}_\mathrm{f}\mathrm{d}T + \bar{v}_\mathrm{f}\mathrm{d}P = -\bar{s}_\mathrm{g}\mathrm{d}T + \bar{v}_\mathrm{g}\mathrm{d}P$$

and rearranging gives the Clausius-Clapeyron equation:

$$\frac{\mathrm{d}P}{\mathrm{d}T} = \frac{s_\mathrm{g} - s_\mathrm{f}}{v_\mathrm{g} - v_\mathrm{f}}$$

Note that since the right-hand side is a ratio, it does not make a difference whether you take the entropies and volumes on a molar basis or on a mass

basis. The mass basis is shown since that is how you will typically find the entropy and volume tabulated. Typical engineering thermodynamic textbooks will also tabulate $s_{\rm fg} = s_{\rm g} - s_{\rm f}$ and $v_{\rm fg} = v_{\rm g} - v_{\rm f}$, making the formula above very convenient.

In case your tables do not have the entropies of the liquid and vapor phases, they often still have the "latent heat of vaporization," also known as "enthalpy of vaporization" or similar, and in engineering thermodynamics books typically indicated by $h_{\rm fg}$. That is the difference between the enthalpy of the saturated liquid and vapor phases, $h_{\rm fg} = h_{\rm g} - h_{\rm f}$. If saturated liquid is turned into saturated vapor by adding heat under conditions of constant pressure and temperature, (11.24) shows that the change in enthalpy $h_{\rm g} - h_{\rm f}$ equals $T(s_{\rm g} - s_{\rm f})$. So the Clausius-Clapeyron equation can be rewritten as

$$\boxed{\frac{{\rm d}P}{{\rm d}T} = \frac{h_{\rm fg}}{T(v_{\rm g} - v_{\rm f})}} \tag{11.38}$$

Because $T\,{\rm d}s$ is the heat added, the physical meaning of the latent heat of vaporization is the heat needed to turn saturated liquid into saturated vapor while keeping the temperature and pressure constant.

For chemical reactions, like maybe

$$2H_2 + O_2 \Longleftrightarrow 2H_2O,$$

the changes in the amounts of the constituents are related as

$$\mathrm{d}\bar{\imath}_{\mathrm{H_2}} = -2\mathrm{d}\bar{r} \quad \mathrm{d}\bar{\imath}_{\mathrm{O_2}} = -1\mathrm{d}\bar{r} \quad \mathrm{d}\bar{\imath}_{\mathrm{H_2O}} = 2\mathrm{d}\bar{r}$$

where $\mathrm{d}\bar{r}$ is the additional number of times the forward reaction takes place from the starting state. The constants -2, -1, and 2 are called the "stoichiometric coefficients." They can be used when applying the condition that at equilibrium, the change in Gibbs energy due to an infinitesimal amount of further reactions $\mathrm{d}\bar{r}$ must be zero.

However, chemical reactions are often posed in a context of constant volume rather than constant pressure, for one because it simplifies the reaction kinematics. For constant volume, the Helmholtz free energy must be used instead of the Gibbs one. Does that mean that a second set of chemical potentials is needed to deal with those problems? Fortunately, the answer is no, the same chemical potentials will do for Helmholtz problems. To see why, note that by definition $F = G - PV$, so ${\rm d}F = {\rm d}G - P{\rm d}V - V{\rm d}P$, and substituting for ${\rm d}G$ from (11.37), that gives

$$\boxed{{\rm d}F = -S\,{\rm d}T - P\,{\rm d}V + \bar{\mu}_1\,\mathrm{d}\bar{\imath}_1 + \bar{\mu}_2\,\mathrm{d}\bar{\imath}_2 + \ldots = -S\,{\rm d}T - P\,{\rm d}V + \sum_c \bar{\mu}_c\,\mathrm{d}\bar{\imath}_c} \tag{11.39}$$

Under isothermal and constant volume conditions, the first two terms in the right hand side will be zero and F will be minimal when the differentials with respect to the amounts of particles add up to zero.

Does this mean that the chemical potentials are also specific Helmholtz free energies, just like they are specific Gibbs free energies? Of course the answer is no, and the reason is that the partial derivatives of F represented by the chemical potentials keep extensive volume V, instead of intensive molar specific volume $\bar{v}$ constant. A single-constituent molar specific Helmholtz energy $\bar{f}$ can be considered to be a function $\bar{f}(T,\bar{v})$ of temperature and molar specific volume, two intensive variables, and then $F = \bar{\imath}\bar{f}(T,\bar{v})$, but $\left(\partial\bar{\imath}\bar{f}(T,V/\bar{\imath})/\partial\bar{\imath}\right)_{TV}$ does not simply produce $\bar{f}$, even if $\left(\partial\bar{\imath}\bar{g}(T,P)/\partial\bar{\imath}\right)_{TP}$ produces $\bar{g}$.

11.13 Microscopic Meaning of the Variables

The new variables introduced in the previous section assume the temperature to be defined, hence there must be thermodynamic equilibrium in some meaningful sense. That is important for identifying their microscopic descriptions, since the canonical expression $P_q = e^{-E_q^{\mathrm{S}}/kt}/Z$ can be used for the probabilities of the energy eigenfunctions.

Consider first the Helmholtz free energy:

$$F = E - TS = \sum_q P_q E_q^{\mathrm{S}} + Tk_{\mathrm{B}} \sum_q P_q \ln\left(e^{-E_q^{\mathrm{S}}/k_{\mathrm{B}}T}/Z\right)$$

This can be simplified by taking apart the logarithm, and noting that the probabilities must sum to one, $\sum_q P_q = 1$, to give

$$\boxed{F = -k_{\mathrm{B}}T \ln Z} \tag{11.40}$$

That makes strike three for the partition function Z, since it already was able to produce the internal energy E, (11.7), and the pressure P, (11.9). Knowing Z as a function of volume V, temperature T, and number of particles I is all that is needed to figure out the other variables. Indeed, knowing F is just as good as knowing the entropy S, since $F = E - TS$. It illustrates why the partition function is much more valuable than you might expect from a mere normalization factor of the probabilities.

For the Gibbs free energy, add PV from (11.9):

$$\boxed{G = -k_{\mathrm{B}}T\left[\ln Z - V\left(\frac{\partial \ln Z}{\partial V}\right)_T\right]} \tag{11.41}$$

Dividing by the number of moles gives the molar specific Gibbs energy $\bar{g}$, equal to the chemical potential $\bar{\mu}$.

How about showing that this chemical potential is the same one as in the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein distribution functions for weakly interacting particles? It is surprisingly difficult to show it; in fact, it cannot be done for distinguishable particles for which the entropy does not exist. It further appears that the best way to get the result for bosons and fermions is to elaborately re-derive the two distributions from scratch, each separately, using a new approach. Note that they were already derived twice earlier, once for given system energy, and once for the canonical probability distribution. So the dual derivations in {D.61} make three. Please note that whatever this book tells you thrice is absolutely true.

11.14 Application to Particles in a Box

This section applies the ideas developed in the previous sections to weakly interacting particles in a box. This allows some of the details of the "shelves" in figures 11.1 through 11.3 to be filled in for a concrete case.

For particles in a macroscopic box, the single-particle energy levels $E^{\rm p}$ are so closely spaced that they can be taken to be continuously varying. The one exception is the ground state when Bose-Einstein condensation occurs; that will be ignored for now. In continuum approximation, the number of single-particle energy states in a macroscopically small energy range $\mathrm{d}E^{\rm p}$ is approximately, following (6.6),

$$\boxed{\mathrm{d}N = V n_s \mathcal{D}\,\mathrm{d}E^{\rm p} = V \frac{n_s}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E^{\rm p}}\,\mathrm{d}E^{\rm p}} \tag{11.42}$$

Here $n_s = 2s+1$ is the number of spin states.

Now according to the derived distributions, the number of particles in a single energy state at energy $E^{\rm p}$ is

$$\iota = \frac{1}{e^{(E^{\rm p}-\mu)/k_{\rm B}T} \pm 1}$$

where the plus sign applies for fermions and the minus sign for bosons. The term can be ignored completely for distinguishable particles.

To get the total number of particles, just integrate the particles per state ι over all states:

$$I = \int_{E^{\rm p}=0}^{\infty} \iota V n_s \mathcal{D}\,\mathrm{d}E^{\rm p} = V \frac{n_s}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \int_{E^{\rm p}=0}^{\infty} \frac{\sqrt{E^{\rm p}}}{e^{(E^{\rm p}-\mu)/k_{\rm B}T} \pm 1}\,\mathrm{d}E^{\rm p}$$

and to get the total energy, integrate the energy of each single-particle state times the number of particles in that state over all states:

$$E = \int_{E^{\rm p}=0}^{\infty} E^{\rm p} \iota n_s V \mathcal{D}\,\mathrm{d}E^{\rm p} = V \frac{n_s}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \int_{E^{\rm p}=0}^{\infty} \frac{E^{\rm p}\sqrt{E^{\rm p}}}{e^{(E^{\rm p}-\mu)/k_{\rm B}T} \pm 1}\,\mathrm{d}E^{\rm p}$$

The expression for the number of particles can be nondimensionalized by rearranging and taking a root to give

$$\boxed{\frac{\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}}{k_{\rm B}T} = \left(\frac{n_s}{4\pi^2}\int_{u=0}^{\infty}\frac{\sqrt{u}\,{\rm d}u}{e^{u-u_0}\pm 1}\right)^{2/3} \qquad u \equiv \frac{E^{\rm p}}{k_{\rm B}T} \quad u_0 \equiv \frac{\mu}{k_{\rm B}T}} \tag{11.43}$$

Note that the left hand side is a nondimensional ratio of a typical quantum microscopic energy, based on the average particle spacing $\sqrt[3]{V/I}$, to the typical classical microscopic energy $k_{\rm B}T$. This ratio is a key nondimensional number governing weakly interacting particles in a box. To put the typical quantum energy into context, a single particle in its own volume of size V/I would have a ground state energy $3\pi^2\hbar^2/2m(V/I)^{2/3}$.

Some references, [4], define a "thermal de Broglie wavelength" $\lambda_{\rm th}$ by writing the classical microscopic energy $k_{\rm B}T$ in a quantum-like way:

$$k_{\rm B}T \equiv 4\pi\frac{\hbar^2}{2m}\frac{1}{\lambda_{\rm th}^2}$$

In some simple cases, you can think of this as roughly the quantum wavelength corresponding to the momentum of the particles. It allows various results that depend on the nondimensional ratio of energies to be reformulated in terms of a nondimensional ratio of lengths, as in

$$\frac{\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}}{k_{\rm B}T} = \frac{1}{4\pi}\left[\frac{\lambda_{\rm th}}{(V/I)^{1/3}}\right]^2$$

Since the ratio of energies is fully equivalent, and has an unambiguous meaning, this book will refrain from making theory harder than needed by defining superfluous quantities. But in practice, thinking in terms of numerical values that are lengths is likely to be more intuitive than energies, and then the numerical value of the thermal wavelength would be the one to keep in mind.

Note that (11.43) provides a direct relationship between the ratio of typical quantum/classical energies on one side, and u_0, the ratio of atomic chemical potential μ to typical classical microscopic energy $k_{\rm B}T$ on the other side. While the two energy ratios are not the same, (11.43) makes them equivalent for systems of weakly interacting particles in boxes. Know one and you can in principle compute the other.

The expression for the system energy may be nondimensionalized in a similar way to get

$$\boxed{\frac{E}{Ik_{\rm B}T} = \int_{u=0}^{\infty}\frac{u\sqrt{u}\,{\rm d}u}{e^{u-u_0}\pm 1}\Bigg/\int_{u=0}^{\infty}\frac{\sqrt{u}\,{\rm d}u}{e^{u-u_0}\pm 1} \qquad u \equiv \frac{E^{\rm p}}{k_{\rm B}T} \quad u_0 \equiv \frac{\mu}{k_{\rm B}T}} \tag{11.44}$$

The integral in the bottom arises when getting rid of the ratio of energies that forms using (11.43).

The quantity in the left hand side is the nondimensional ratio of the actual system energy over the system energy if every particle had the typical classical energy $k_{\rm B}T$. It too is a unique function of u_0, and as a consequence, also of the ratio of typical microscopic quantum and classical energies.

11.14.1 Bose-Einstein condensation

Bose-Einstein condensation is said to have occurred when in a macroscopic system the number of bosons in the ground state becomes a finite fraction of the number of particles I. It happens when the temperature is lowered sufficiently or the particle density is increased sufficiently or both.

According to derivation {D.57}, the number of particles in the ground state is given by

$$I_1 = \frac{N_1 - 1}{e^{(E_1^{\rm p}-\mu)/k_{\rm B}T} - 1}. \tag{11.45}$$

In order for this to become a finite fraction of the large number of particles I of a macroscopic system, the denominator must become extremely small, hence the exponential must become extremely close to one, hence μ must come extremely close to the lowest energy level $E_1^{\rm p}$. To be precise, $E_1 - \mu$ must be small of order $k_{\rm B}T/I$; smaller than the classical *microscopic* energy by the humongous factor I. In addition, for a macroscopic system of weakly interacting particles in a box, $E_1^{\rm p}$ is extremely close to zero, (it is smaller than the microscopic quantum energy defined above by a factor $I^{2/3}$.) So condensation occurs when $\mu \approx E_1^{\rm p} \approx 0$, the approximations being extremely close. If the ground state is unique, $N_1 = 1$, Bose-Einstein condensation simply occurs when $\mu = E_1^{\rm p} \approx 0$.

You would therefore expect that you can simply put $u_0 = \mu/k_{\rm B}T$ to zero in the integrals (11.43) and (11.44). However, if you do so (11.43) fails to describe the number of particles in the ground state; it only gives the number of particles $I - I_1$ not in the ground state:

$$\frac{\dfrac{\hbar^2}{2m}\left(\dfrac{I-I_1}{V}\right)^{2/3}}{k_{\rm B}T} = \left(\frac{n_s}{4\pi^2}\int_{u=0}^{\infty}\frac{\sqrt{u}\,{\rm d}u}{e^u-1}\right)^{2/3} \qquad \text{for BEC} \tag{11.46}$$

To see that the number of particles in the ground state is indeed not included in the integral, note that while the integrand does become infinite when $u \downarrow 0$, it becomes infinite proportionally to $1/\sqrt{u}$, which integrates as proportional to $\sqrt{u}$, and $\sqrt{u_1} = \sqrt{E_1^{\rm p}/k_{\rm B}T}$ is vanishingly small, not finite. Arguments given in derivation {D.57} do show that the only significant error occurs for the ground state; the above integral does correctly approximate the number of particles not in the ground state when condensation has occurred.

The value of the integral can be found in mathematical handbooks, [41, p. 201, with typo], as $\frac{1}{2}!\zeta\left(\frac{3}{2}\right)$ with ζ the so-called Riemann zeta function, due to, who else, Euler. Euler showed that it is equal to a product of terms ranging over all prime numbers, but you do not want to know that. All you want to know is that $\zeta\left(\frac{3}{2}\right) \approx 2.612$ and that $\frac{1}{2}! = \frac{1}{2}\sqrt{\pi}$.

The Bose-Einstein temperature T_B is the temperature at which Bose-Einstein condensation starts. That means it is the temperature for which $I_1 = 0$ in the expression above, giving

$$\frac{\frac{\hbar^2}{2m}\left(\frac{I-I_1}{V}\right)^{2/3}}{k_{\rm B}T} = \frac{\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}}{k_{\rm B}T_B} = \left(\frac{n_s}{8\pi^{3/2}}\zeta\left(\tfrac{3}{2}\right)\right)^{2/3} \qquad T \leqslant T_B \tag{11.47}$$

It implies that for a given system of bosons, at Bose-Einstein condensation there is a fixed numerical ratio between the microscopic quantum energy based on particle density and the classical microscopic energy $k_{\rm B}T_B$. That also illustrates the point made at the beginning of this subsection that both changes in temperature and changes in particle density can produce Bose-Einstein condensation.

The first equality in the equation above can be cleaned up to give the fraction of bosons in the ground state as:

$$\frac{I_1}{I} = 1 - \left(\frac{T}{T_B}\right)^{3/2} \qquad T \leqslant T_B \tag{11.48}$$

11.14.2 Fermions at low temperatures

Another application of the integrals (11.43) and (11.44) is to find the Fermi energy $E_{\rm F}^{\rm p}$ and internal energy E of a system of weakly interacting fermions for vanishing temperature.

For low temperatures, the nondimensional energy ratio $u_0 = \mu/k_{\rm B}T$ blows up, since $k_{\rm B}T$ becomes zero and the chemical potential μ does not; μ becomes the Fermi energy $E_{\rm F}^{\rm p}$, chapter 6.10. To deal with the blow up, the integrals can be rephrased in terms of $u/u_0 = E^{\rm p}/\mu$, which does not blow up.

In particular, the ratio (11.43) involving the typical microscopic quantum energy can be rewritten by taking a factor $u_0^{3/2}$ out of the integral and root and to the other side to give:

$$\frac{\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}}{\mu} = \left(\frac{n_s}{4\pi^2}\int_{u/u_0=0}^{\infty}\frac{\sqrt{u/u_0}\,{\rm d}(u/u_0)}{e^{u_0[(u/u_0)-1]}+1}\right)^{2/3}$$

Now since u_0 is large, the exponential in the denominator becomes extremely large for $u/u_0 > 1$, making the integrand negligibly small. Therefore the upper limit of integration can be limited to $u/u_0 = 1$. In that range, the exponential

is vanishingly small, except for a negligibly small range around $u/u_0 = 1$, so it can be ignored. That gives

$$\frac{\dfrac{\hbar^2}{2m}\left(\dfrac{I}{V}\right)^{2/3}}{\mu} = \left(\frac{n_s}{4\pi^2}\int_{u/u_0=0}^{1}\sqrt{u/u_0}\,\mathrm{d}(u/u_0)\right)^{2/3} = \left(\frac{n_s}{6\pi^2}\right)^{2/3}$$

It follows that the Fermi energy is

$$E_{\mathrm{F}}^{\mathrm{p}} = \mu|_{T=0} = \left(\frac{6\pi^2}{n_s}\right)^{2/3}\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}$$

Physicist like to define a "Fermi temperature" as the temperature where the classical microscopic energy $k_{\mathrm{B}}T$ becomes equal to the Fermi energy. It is

$$T_{\mathrm{F}} = \frac{1}{k_{\mathrm{B}}}\left(\frac{6\pi^2}{n_s}\right)^{2/3}\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3} \tag{11.49}$$

It may be noted that except for the numerical factor, the expression for the Fermi temperature T_{F} is the same as that for the Bose-Einstein condensation temperature T_B given in the previous subsection.

Electrons have $n_s = 2$. For the valence electrons in typical metals, the Fermi temperatures are in the order of ten thousands of degrees Kelvin. The metal will melt before it is reached. The valence electrons are pretty much the same at room temperature as they are at absolute zero.

The integral (11.44) can be integrated in the same way and then shows that $E = \frac{3}{5}I\mu = \frac{3}{5}IE_{\mathrm{F}}^{\mathrm{p}}$. In short, at absolute zero, the average energy per particle is $\frac{3}{5}$ times $E_{\mathrm{F}}^{\mathrm{p}}$, the maximum single-particle energy.

It should be admitted that both of the results in this subsection have been obtained more simply in chapter 6.10. However, the analysis in this subsection can be used to find the corrected expressions when the temperature is fairly small but not zero, {D.62}, or for any temperature by brute-force numerical integration. One result is the specific heat at constant volume of the free-electron gas for low temperatures:

$$C_v = \frac{\pi^2}{2}\frac{k_{\mathrm{B}}T}{E_{\mathrm{F}}^{\mathrm{p}}}\frac{k_{\mathrm{B}}}{m}(1+\ldots) \tag{11.50}$$

where k_{B}/m is the gas constant R. All low-temperature expansions proceed in powers of $(k_{\mathrm{B}}T/E_{\mathrm{F}}^{\mathrm{p}})^2$, so the dots in the expression for C_v above are of that order. The specific heat vanishes at zero temperature and is typically small.

11.14.3 A generalized ideal gas law

While the previous subsections produced a lot of interesting information about weakly interacting particles near absolute zero, how about some info about

conditions that you can check in a T-shirt? And how about something mathematically simple, instead of elaborate integrals that produce weird functions?

Well, there is at least one. By definition, (11.8), the pressure is the expectation value of $-\mathrm{d}E_q^{\mathrm{S}}/\mathrm{d}V$ where the E_q^{S} are the system energy eigenvalues. For weakly interacting particles in a box, chapter 6.2 found that the single particle energies are inversely proportional to the squares of the linear dimensions of the box, which means proportional to $V^{-2/3}$. Then so are the system energy eigenfunctions, since they are sums of single-particle ones: $E_q^{\mathrm{S}} = \text{constant } V^{-2/3}$ Differentiating produces $\mathrm{d}E_q^{\mathrm{S}}/\mathrm{d}V = -\frac{2}{3}E_q^{\mathrm{S}}/V$ and taking the expectation value

$$\boxed{PV = \tfrac{2}{3}E} \tag{11.51}$$

This expression is valid for weakly interacting bosons and fermions even if the symmetrization requirements cannot be ignored.

11.14.4 The ideal gas

The weakly interacting particles in a box can be approximated as an ideal gas if the number of particles is so small, or the box so large, that the average number of particles in an energy state is much less than one.

Since the number of particles per energy state is given by

$$\iota = \frac{1}{e^{(E^{\mathrm{p}}-\mu)/k_{\mathrm{B}}T} \pm 1}$$

ideal gas conditions imply that the exponential must be much greater than one, and then the ± 1 can be ignored. That means that the difference between fermions and bosons, which accounts for the ± 1, can be ignored for an ideal gas. Both can be approximated by the distribution derived for distinguishable particles.

The energy integral (11.44) can now easily be done; the e^{u_0} factor divides away and an integration by parts in the numerator produces $E = \frac{3}{2}Ik_{\mathrm{B}}T$. Plug it into the generalized ideal gas law (11.51) to get the normal "ideal gas law"

$$\boxed{PV = Ik_{\mathrm{B}}T \qquad \Longleftrightarrow \qquad Pv = RT \quad R \equiv \frac{k_{\mathrm{B}}}{m}} \tag{11.52}$$

Also, following (11.34),

$$e = \tfrac{3}{2}\frac{k_{\mathrm{B}}}{m}T = C_v T \quad h = \tfrac{5}{2}\frac{k_{\mathrm{B}}}{m}T = C_p T \quad C_v = \tfrac{3}{2}R \quad C_p = \tfrac{5}{2}R$$

but note that these formulae are specific to the simplistic ideal gases described by the model, (like noble gases.) For ideal gases with more complex molecules,

like air, the specific heats are not constants, but vary with temperature, as discussed in section 11.15.

The ideal gas equation is identical to the one derived in classical physics. That is important since it establishes that what was defined to be the temperature in this chapter is in fact the ideal gas temperature that classical physics defines.

The integral (11.43) can be done using integration by parts and a result found in the notations under "!". It gives an expression for the single-particle chemical potential μ:

$$-\frac{\mu}{k_{\rm B}T} = \tfrac{3}{2}\ln\left[k_{\rm B}T \Big/ 4\pi n_s^{-2/3}\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}\right]$$

Note that the argument of the logarithm is essentially the ratio between the classical microscopic energy and the quantum microscopic energy based on average particle spacing. This ratio has to be big for an accurate ideal gas, to get the exponential in the particle energy distribution ι to be big.

Next is the specific entropy s. Recall that the chemical potential is just the Gibbs free energy. By the definition of the Gibbs free energy, the specific entropy s equals $(h-g)/T$. Now the specific Gibbs energy is just the Gibbs energy per unit mass, in other words, μ/m while $h/T = C_p$ as above. So

$$\boxed{s = C_v \ln\left[k_{\rm B}T \Big/ 4\pi n_s^{-2/3}\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3}\right] + C_p} \tag{11.53}$$

In terms of classical thermodynamics, V/I is m times the specific volume v. So classical thermodynamics takes the logarithm above apart as

$$s = C_v \ln(T) + R\ln(v) + \text{some combined constant}$$

and then promptly forgets about the constant, damn units.

11.14.5 Blackbody radiation

This section takes a closer look at blackbody radiation, discussed earlier in chapter 6.8. Blackbody radiation is the basic model for absorption and emission of electromagnetic radiation. Electromagnetic radiation includes light and a wide range of other radiation, like radio waves, microwaves, and X-rays. All surfaces absorb and emit radiation; otherwise we would not see anything. But "black" surfaces are the most easy to understand theoretically.

No, a black body need not look black. If its temperature is high enough, it could look like the sun. What defines an ideal black body is that it absorbs, (internalizes instead of reflects,) all radiation that hits it. But it may be emitting

its own radiation at the same time. And that makes a difference. If the black body is cool, you will need your infrared camera to see it; it would look really black to the eye. It is not reflecting any radiation, and it is not emitting any visible amount either. But if it is at the temperature of the sun, better take out your sunglasses. It is still absorbing all radiation that hits it, but it is emitting large amounts of its own too, and lots of it in the visible range.

So where do you get a nearly perfectly black surface? Matte black paint? A piece of blackboard? Soot? Actually, pretty much all materials will reflect in some range of wave lengths. You get the blackest surface by using no material at all. Take a big box and paint its interior the blackest you can. Close the box, then drill a very tiny hole in its side. From the outside, the area of the hole will be truly, absolutely black. Whatever radiation enters there is gone. Still, when you heat the box to very high temperatures, the hole will shine bright.

While any radiation entering the hole will most surely be absorbed somewhere inside, the inside of the box itself is filled with electromagnetic radiation, like a gas of photons, produced by the hot inside surface of the box. And some of those photons will manage to escape through the hole, making it shine.

The amount of photons in the box may be computed from the Bose-Einstein distribution with a few caveats. The first is that there is no limit on the number of photons; photons will be created or absorbed by the box surface to achieve thermal equilibrium at whatever level is most probable at the given temperature. This means the chemical potential μ of the photons is zero, as you can check from the derivations in notes {D.57} and {D.58}.

The second caveat is that the usual density of states (6.6) is nonrelativistic. It does not apply to photons, which move at the speed of light. For photons you must use the density of modes (6.7).

The third caveat is that there are only two independent spin states for a photon. As a spin-one particle you would expect that photons would have the spin values 0 and ± 1, but the zero value does not occur in the direction of propagation, addendum {A.21.6}. Therefore the number of independent states that exist is two, not three. A different way to understand this is classical: the electric field can only oscillate in the two independent directions normal to the direction of propagation, (13.10); oscillation in the direction of propagation itself is not allowed by Maxwell's laws because it would make the divergence of the electric field nonzero. The fact that there are only two independent states has already been accounted for in the density of modes (6.7).

The energy per unit box volume and unit frequency range found under the above caveats is Planck's blackbody spectrum already given in chapter 6.8:

$$\rho(\omega) \equiv \frac{\mathrm{d}(E/V)}{\mathrm{d}\omega} = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{e^{\hbar\omega/k_{\mathrm{B}}T} - 1} \tag{11.54}$$

The expression for the total internal energy per unit volume is called the "Stefan-Boltzmann formula." It is found by integration of Planck's spectrum

over all frequencies just like for the Stefan-Boltzmann law in chapter 6.8:

$$\boxed{\frac{E}{V} = \frac{\pi^2}{15\hbar^3 c^3}(k_{\mathrm{B}}T)^4} \tag{11.55}$$

The number of particles may be found similar to the energy, by dropping the $\hbar\omega$ energy per particle from the integral. It is, [41, 36.24, with typo]:

$$\frac{I}{V} = \frac{2\zeta(3)}{\pi^2\hbar^3 c^3}(k_{\mathrm{B}}T)^3 \qquad \zeta(3) \approx 1.202 \tag{11.56}$$

Taking the ratio with (11.55), the average energy per photon may be found:

$$\boxed{\frac{E}{I} = \frac{\pi^4}{30\zeta(3)}k_{\mathrm{B}}T \approx 2.7k_{\mathrm{B}}T} \tag{11.57}$$

The temperature has to be roughly 9 000 K for the average photon to become visible light. That is one reason a black body will look black at a room temperature of about 300 K. The solar surface has a temperature of about 6 000 K, so the visible light photons it emits are more energetic than average, but there are still plenty of them.

The entropy S of the photon gas follows from integrating $\int \mathrm{d}E/T$ using (11.55), starting from absolute zero and keeping the volume constant:

$$\boxed{\frac{S}{V} = \frac{4\pi^2}{45\hbar^3 c^3}k_{\mathrm{B}}(k_{\mathrm{B}}T)^3} \tag{11.58}$$

Dividing by (11.56) shows the average entropy per photon to be

$$\frac{S}{I} = \frac{2\pi^4}{45\zeta(3)}k_{\mathrm{B}} \tag{11.59}$$

independent of temperature.

The generalized ideal gas law (11.51) does not apply to the pressure exerted by the photon gas, because the energy of the photons is $\hbar ck$ and that is proportional to the wave number instead of its square. The corrected expression is:

$$\boxed{PV = \tfrac{1}{3}E} \tag{11.60}$$

11.14.6 The Debye model

To explain the heat capacity of simple solids, Debye modeled the energy in the crystal vibrations very much the same way as the photon gas of the previous subsection. This subsection briefly outlines the main ideas.

For electromagnetic waves propagating with the speed of light c, substitute acoustical waves propagating with the speed of sound c_s. For photons with energy $\hbar\omega$, substitute phonons with energy $\hbar\omega$. Since unlike electromagnetic waves, sound waves *can* vibrate in the direction of wave propagation, for the number of spin states substitute $n_s = 3$ instead of 2; in other words, just multiply the various expressions for photons by 1.5.

The critical difference for solids is that the number of modes, hence the frequencies, is not infinitely large. Since each individual atom has three degrees of freedom (it can move in three individual directions), there are $3I$ degrees of freedom, and reformulating the motion in terms of acoustic waves does not change the number of degrees of freedom. The shortest wave lengths will be comparable to the atom spacing, and no waves of shorter wave length will exist. As a result, there will be a highest frequency $\omega_{\max}$. The "Debye temperature" T_D is defined as the temperature at which the typical classical microscopic energy $k_B T$ becomes equal to the maximum quantum microscopic energy $\hbar\omega_{\max}$

$$\boxed{k_B T_D = \hbar\omega_{max}} \tag{11.61}$$

The expression for the internal energy becomes, from (6.11) times 1.5:

$$\boxed{\frac{E}{V} = \int_0^{\omega_{\max}} \frac{3\hbar}{2\pi^2 c_s^3} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1}\, d\omega} \tag{11.62}$$

If the temperatures are very low the exponential will make the integrand zero except for very small frequencies. Then the upper limit is essentially infinite compared to the range of integration. That makes the energy proportional to T^4 just like for the photon gas and the heat capacity is therefore proportional to T^3. At the other extreme, when the temperature is large, the exponential in the bottom can be expanded in a Taylor series and the energy becomes proportional to T, making the heat capacity constant.

The maximum frequency, hence the Debye temperature, can be found from the requirement that the number of modes is $3I$, to be applied by integrating (6.7), or an empirical value can be used to improve the approximation for whatever temperature range is of interest. Literature values are often chosen to approximate the low temperature range accurately, since the model works best for low temperatures. If integration of (6.7) is used at high temperatures, the law of Dulong and Petit results, as described in section 11.15.

More sophisticated versions of the analysis exist to account for some of the very nontrivial differences between crystal vibrations and electromagnetic waves. They will need to be left to literature.

11.15 Specific Heats

The specific heat of a substance describes its absorption of heat in terms of its temperature change. In particular, the specific heat at constant volume, C_v, of a substance is the thermal energy that gets stored internally in the substance per unit temperature rise and per unit amount of substance.

As a first example, consider simple monatomic ideal gases, and in particular noble gases. Basic physics, or section 11.14.4, shows that for an ideal gas, the molecules have $\frac{1}{2}k_{\rm B}T$ of translational kinetic energy in each of the three directions of a Cartesian coordinate system, where $k_{\rm B} = 1.38\ 10^{-23}$ J/K is Boltzmann's constant. So the specific heat per molecule is $\frac{3}{2}k_{\rm B}$ or $1.5k_{\rm B}$. For a kmol (i.e. $6.02\ 10^{26}$) of molecules instead of one, $k_{\rm B}$ becomes the "universal gas constant" $R_{\rm u} = 8.31$ kJ/kmol K. Hence for a

$$\text{monatomic ideal gas: } \bar{C}_v = 1.5R_{\rm u} \approx 12.5 \text{ kJ/kmol K} \tag{11.63}$$

on a kmol basis. As figure 11.15 shows, this is very accurate for noble gases, including helium. (To get the more usual specific heat C_v per kilogram instead of kmol, divide by the molar mass M. For example, for helium with two protons and two neutrons in its nucleus, the molar mass is about 4 kg/kmol, so divide by 4. In thermo books, you will probably find the molar mass values you need mislisted as "molecular mass," without units. Just use the values and ignore the name and the missing units of kg/kmol. See the notations for more.)

Many important ideal gases, such as hydrogen, as well as the oxygen and nitrogen that make up air, are diatomic. Now if we assume that the two atoms are point-size masses somehow rigidly connected to each other, we still have that the center of the entire molecule can move in three different directions, accounting for $\frac{3}{2}k_{\rm B}$ of kinetic translational energy. But at any given time, the molecule can also be conducting rotational motion around its center in two independent directions, both orthogonal to the connecting line between the atoms. (For point masses, rotation around the connecting axis would not do anything.) Classical physics, in particular the "equipartition theorem," would then predict that each of the two rotational motions has $\frac{1}{2}k_{\rm B}$ of kinetic energy too, raising the total specific heat to $\frac{5}{2}k_{\rm B}$ or $2.5k_{\rm B}$. Well, figure 11.15 shows that *at room temperature*, about 300 K, this is quite accurate for common diatomic gases like the nitrogen and oxygen in air.

But note that these experimental data show that there are problems, both at very low temperatures, and at very high ones. And there are major theoretical problems too. Surely the connection between the atoms is not going to be infinitely rigid. Allowing for that, we now have two individual atoms that can each move in three different directions independently of each other. That raises the kinetic energy to $\frac{6}{2}k_{\rm B}$. And assuming that the connecting force varies linearly with elongation, there would be another $\frac{1}{2}k_{\rm B}$ of potential energy, making the

total energy $\frac{7}{2}k_B$. But figure 11.15 shows that the common diatomic gasses only approach values like that well above room temperature.

It was all a big problem for classical physics. Not to mention that, as Maxwell noted, if you really take classical theory at face value, things get far, far, worse still, since the individual internal parts of the atoms, like the individual electrons and quarks in the nuclei, would each have to absorb their own thermal energy too. This should produce enormously high specific heats.

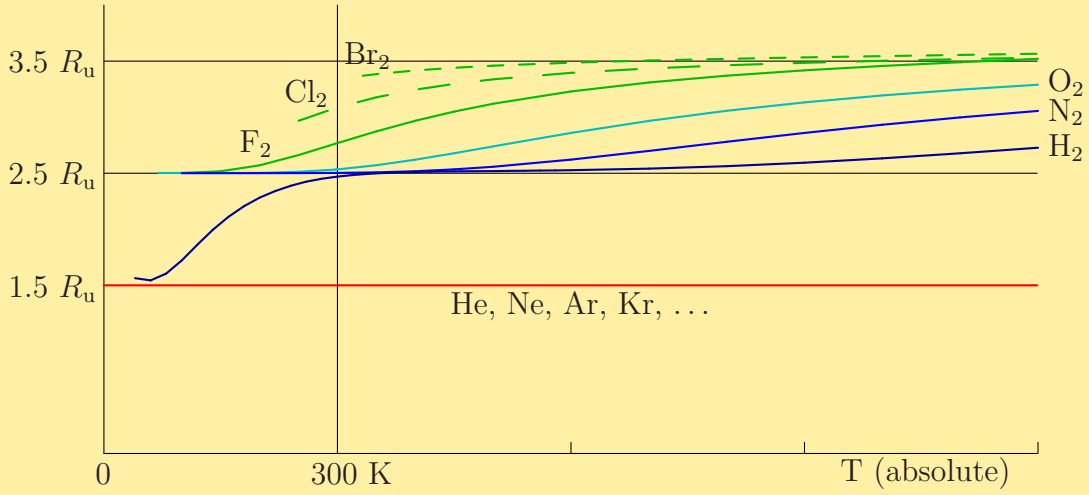


Figure 11.15: Specific heat at constant volume of gases. Temperatures from absolute zero to 1,200 K. Data from NIST-JANAF and AIP.

Hydrogen in particular was a mystery before the advent of quantum mechanics: at low temperatures it would behave as a *monatomic* gas, with a specific heat of $\frac{3}{2}k_B$ per molecule, figure 11.15. That meant that the molecule had to be translating *only*, like a monatomic gas. How could the random thermal motion not cause any angular rotation of the two atoms around their mutual center of gravity, nor vibration of the atoms towards and away from each other?

Quantum mechanics solved this problem. In quantum mechanics the angular momentum of the molecule, and so the corresponding kinetic energy, as well as the harmonic oscillation energy, are quantized. For hydrogen at low temperatures, the typical available thermal energy $\frac{1}{2}k_BT$ is not enough to reach even the first level above the ground state for either energy. No thermal energy can therefore be put into rotation of the molecule, nor into internal vibration. So hydrogen does indeed have the specific heat of monatomic gases at low temperatures, weird as it may seem. The rotational and vibrational motions are "frozen out."

At normal temperatures, there is enough thermal energy to reach the states where the molecule rotates normal to the line connecting the atoms, and the

specific heat becomes

$$\text{typical diatomic ideal gas: } \bar{C}_v = 2.5 R_{\text{u}} \approx 20.8 \text{ kJ/kmol K.} \tag{11.64}$$

Actual values for hydrogen, nitrogen and oxygen at room temperature are 2.47, 2.50, and 2.53 R_{u}.

For high enough temperature, the vibrational modes will start becoming active, and the specific heats will start inching up towards 3.5 R_{u} (and beyond), figure 11.15. But it takes to temperatures of 1 000 K (hydrogen), 600 K (nitrogen), or 400 K (oxygen) before there is a 5% deviation from the 2.5 R_{u} value.

These differences may be understood qualitatively if the motion is modeled as a simple harmonic oscillator as discussed in chapter 4.1. The energy levels of an harmonic oscillator are apart by an amount $\hbar\omega$, where ω is the angular frequency. And the frequency of a harmonic oscillator $\omega = \sqrt{c/m}$, where c is the effective stiffness and m the effective mass of the vibrational motion. So light atoms that are bound together tightly will require a lot of thermal energy to reach the first nontrivial vibrational state. Hydrogen is much lighter than nitrogen or oxygen, so the required energy $\hbar\omega$ should be quite large. This explains the high temperature before vibration become important for hydrogen. The molar masses of nitrogen and oxygen are similar, but nitrogen is bound with a triple bond, and oxygen only a double one. So nitrogen has the higher effective stiffness of the two and vibrates less readily.

Following this reasoning, you would expect fluorine, which is held together with only a single covalent bond, to have a higher specific heat still, and figure 11.15 confirms it. And chlorine and bromine, also held together by a single covalent bond, but heavier than fluorine, approach the classical value 3.5 R_{u} fairly closely at normal temperatures: Cl_2 has 3.08 R_{u} and Br_2 3.34 R_{u}.

For solids, the basic classical idea in terms of atomic motion would be that there would be $\frac{3}{2}R_{\text{u}}$ per atom in kinetic energy and $\frac{3}{2}R_{\text{u}}$ in potential energy:

$$\text{law of Dulong and Petit: } \bar{C}_v = 3 R_{\text{u}} \approx 25 \text{ kJ/kmol K.} \tag{11.65}$$

Not only is 3 a nice round number, it actually works well for a lot of relatively simple solids at room temperature. For example, aluminum is 2.91 R_{u}, copper 2.94, gold 3.05, iron 3.02.

Note that typically for solids $\bar{C}_p$, the heat added per unit temperature change at constant pressure is given instead of $\bar{C}_v$. However, unlike for gases, the difference between $\bar{C}_p$ and $\bar{C}_v$ is small for solids and most liquids and will be ignored here.

Dulong and Petit also works for liquid water if you take it per kmol of atoms, rather than kmol of molecules, but not for ice. Ice has 4.6 R_{u} per kmol of molecules and 1.5 R_{u} per kmol of atoms. For molecules, certainly there is an obvious problem in deciding how many pieces you need to count as independently moving units. A value of 900 R_{u} for paraffin wax (per molecule) found at

Wikipedia may sound astonishing, until you find elsewhere at Wikipedia that its chemical formula is $C_{25}H_{52}$. It is still quite capable of storing a lot of heat per unit weight too, in any case, but nowhere close to hydrogen. Putting $\frac{5}{2}k_B T$ in a molecule with the tiny molecular mass of just about two protons is the real way to get a high heat content per unit mass.

Complex molecules may be an understandable problem for the law of Dulong and Petit, but how come that diamond has about 0.73 R_u, and graphite 1.02 R_u, instead of 3 as it should? No molecules are involved there. The values of boron at 1.33 R_u and beryllium at 1.98 R_u are much too low too, though not as bad as diamond or graphite.

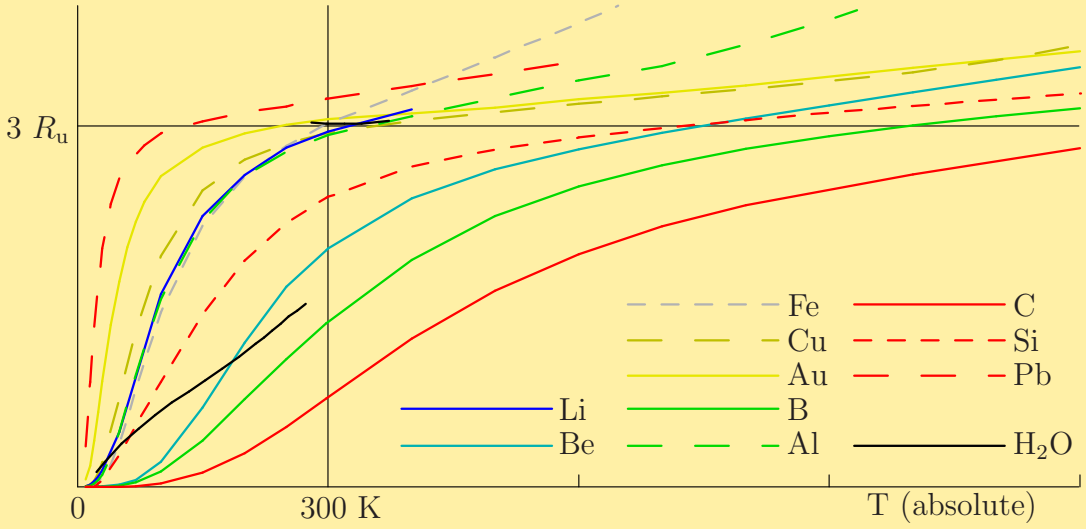


Figure 11.16: Specific heat at constant pressure of solids. Temperatures from absolute zero to 1,200 K. Carbon is diamond; graphite is similar. Water is ice and liquid. Data from NIST-JANAF, CRC, AIP, Rohsenow *et al.*

Actually, it turns out, figure 11.16, that at much higher temperatures diamond *does* agree nicely with the Dulong and Petit value. Conversely, if the elements that agree well with Dulong and Petit at room temperature are cooled to low temperatures, they too have a specific heat that is much lower than the Dulong and Petit value. For example, at 77 K, aluminum has 1.09 R_u, copper 1.5, and diamond 0.01.

It turns out that for all of them a characteristic temperature can by found above which the specific heat is about the Dulong and Petit value, but below which the specific heat starts dropping precariously. This characteristic temperature is called the Debye temperature. For example, aluminum, copper, gold, and iron have Debye temperatures of 394, 315, 170, and 460 K, all near or below room temperature, and their room temperature specific heats agree reasonably with the Dulong and Petit value. Conversely, diamond, boron, and beryllium

have Debye temperatures of 1 860, 1 250, and 1 000 K, and their specific heats are much too low at room temperature.

The lack of heat capacity below the Debye temperature is again a matter of "frozen out" vibrational modes, like the freezing out of the vibrational modes that gave common diatomic ideal gases a heat capacity of only $\frac{5}{2}R_{\rm u}$ instead of $\frac{7}{2}R_{\rm u}$. Note for example that carbon, boron and beryllium are light atoms, and that the diamond structure is particularly stiff, just the properties that froze out the vibrational modes in diatomic gas molecules too. However, the actual description is more complex than for a gas: if all vibrations were frozen out in a solid, there would be nothing left.

Atoms in a solid cannot be considered independent harmonic oscillators like the pairs of atoms in diatomic molecules. If an atom in a solid moves, its neighbors are affected. The proper way to describe the motion of the atoms is in terms of crystal-wide vibrations, such as those that in normal continuum mechanics describe acoustical waves. There are three variants of such waves, corresponding to the three independent directions the motion of the atoms can take with respect to the propagation direction of the wave. The atoms can move in the same direction, like in the acoustics of air in a pipe, or in a direction normal to it, like surface waves in water. Those are called longitudinal and transverse waves respectively. If there is more than one atom in the basis from which the solid crystal is formed, the atoms in a basis can also vibrate relative to each other's position in high-frequency vibrations called optical modes. However, after such details are accounted for, the classical internal energy of a solid is still the Dulong and Petit value.

Enter quantum mechanics. Just like quantum mechanics says that the energy of vibrating electromagnetic fields of frequency ω comes in discrete units called photons, with energy $\hbar\omega$, it says that the energy of crystal vibrations comes in discrete units called "phonons" with energy $\hbar\omega$. As long as the typical amount of heat energy, $k_{\rm B}T$, is larger than the largest of such phonon energies, the fact that the energy levels are discrete make no real difference, and classical analysis works fine. But for lower temperatures, there is not enough energy to create the high-energy phonons and the specific heat will be less. The representative temperature T_D at which the heat energy $k_{\rm B}T_D$ becomes equal to the highest phonon energies $\hbar\omega$ is the Debye temperature. (The Debye analysis is not exact except for low energies, and the definitions of Debye temperature vary somewhat. See section 11.14.6 for more details.)

Quantum mechanics did not just solve the low temperature problems for heat capacity; it also solved the electron problem. That problem was that classically electrons in at least metals too should have $\frac{3}{2}k_{\rm B}T$ of kinetic energy, since electrical conduction meant that they moved independently of the atoms. But observations showed it was simply not there. The quantum mechanical explanation was the Fermi-Dirac distribution of figure 6.11: only a small fraction of the electrons have free energy states above them within a distance of order

$k_B T$, and only these can take on heat energy. Since so few electrons are involved, the amount of energy they absorb is negligible except at very low temperatures. At very low temperatures, the energy in the phonons becomes very small, and the conduction electrons in metals then do make a difference.

Also, when the heat capacity due to the atom vibrations levels off to the Dulong and Petit value, that of the valence electrons keeps growing. Furthermore, at higher temperatures the increased vibrations lead to increased deviations in potential from the harmonic oscillator relationship. Wikipedia, Debye model, says anharmonicity causes the heat capacity to rise further; apparently authoritative other sources say that it can either increase or decrease the heat capacity. In any case, typical solids do show an increase of the heat capacity above the Dulong and Petit value at higher temperatures, figure 11.16.

Chapter 12

Angular momentum

The quantum mechanics of angular momentum is fascinating. It is also very basic to much of quantum mechanics. It is a model for dealing with other systems of particles

In chapter 5.4, it was already mentioned that angular momentum of particles comes in two basic kinds. Orbital angular momentum is a result of the angular motion of particles, while spin is "built-in" angular momentum of the particles.

Orbital angular momentum is usually indicated by $\widehat{\vec{L}}$ and spin angular momentum by $\widehat{\vec{S}}$. A system of particles will normally involve both orbital and spin angular momentum. The combined angular momentum is typically indicated by

$$\widehat{\vec{J}} = \widehat{\vec{L}} + \widehat{\vec{S}}$$

However, this chapter will use $\widehat{\vec{J}}$ as a generic name for any angular momentum. So in this chapter $\widehat{\vec{J}}$ can indicate orbital angular momentum, spin angular momentum, or any combination of the two.

12.1 Introduction

The standard eigenfunctions of orbital angular momentum are the so called "spherical harmonics" of chapter 4.2. They show that the square orbital angular momentum has the possible values

$$L^2 \equiv L_x^2 + L_y^2 + L_z^2 = l(l+1)\hbar^2 \quad \text{where } l \text{ is one of } 0,\ 1,\ 2,\ 3,\ \ldots$$

The nonnegative integer l is called the azimuthal quantum number.

Further, the orbital angular momentum in any arbitrarily chosen direction, taken as the z-direction from now on, comes in multiples m of Planck's constant $\hbar$:

$$L_z = m_l\hbar \quad \text{where } m_l \text{ is one of } -l,\ -l{+}1,\ -l{+}2,\ \ldots,\ l{-}1,\ l.$$

The integer m_l is called the magnetic quantum number.

The possible values of the square spin angular momentum can be written as

$$S^2 \equiv S_x^2 + S_y^2 + S_z^2 = s(s+1)\hbar^2 \qquad \text{where } s \text{ is one of } 0,\ ½,\ 1,\ ⅜,\ \ldots$$

The "spin azimuthal quantum number" s is usually called the "spin" for short. Note that while the orbital azimuthal quantum number l had to be an integer, the spin can be half integer. But one important conclusion of this chapter will be that the spin cannot be anything more. A particle with, say, spin $\frac{1}{3}$ cannot not exist according to the theory.

For the spin angular momentum in the z-direction

$$S_z = m_s\hbar \quad \text{where } m_s \text{ is one of } -s,\ -s{+}1,\ -s{+}2,\ \ldots,\ s{-}1,\ s.$$

Note that if the spin s is half integer, then so are all the spin magnetic quantum numbers m_s. If the nature of the angular momentum is self-evident, the subscript l or s of the magnetic quantum numbers m will be omitted.

Particles with half-integer spin are called fermions. That includes electrons, as well as protons and neutrons and their constituent quarks. All of these critically important particles have spin ½. (Excited proton and neutron states can have spin ⅜.) Particles with integer spin are bosons. That includes the particles that act as carriers of fundamental forces; the photons, intermediate vector bosons, gluons, and gravitons. All of these have spin 1, except the graviton which supposedly has spin 2.

12.2 The fundamental commutation relations

Analyzing nonorbital angular momentum is a challenge. How can you say anything sensible about angular momentum, the dynamic motion of masses around a given point, without a mass moving around a point? For, while a particle like an electron has spin angular momentum, trying to explain it as angular motion of the electron about some internal axis leads to gross contradictions such as the electron exceeding the speed of light [25, p. 172]. Spin is definitely part of the law of conservation of angular momentum, but it does not seem to be associated with any familiar idea of some mass moving around some axis as far as is known.

There goes the Newtonian analogy, then. Something else than classical physics is needed to analyze spin.

Now, the complex discoveries of mathematics are routinely deduced from apparently self-evident simple axioms, such as that a straight line will cross each of a pair of parallel lines under the same angle. Actually, such axioms are not as obvious as they seem, and mathematicians have deduced very different answers from changing the axioms into different ones. Such answers may be just

as good or better than others depending on circumstances, and you can invent imaginary universes in which they are the norm.

Physics has no such latitude to invent its own universes; its mission is to describe *ours* as well as it can. But the idea of mathematics is still a good one: try to guess the simplest possible basic "law" that nature really seems to obey, and then reconstruct as much of the complexity of nature from it as you can. The more you can deduce from the law, the more ways you have to check it against a variety of facts, and the more confident you can become in it.

Physicist have found that the needed equations for angular momentum are given by the following "fundamental commutation relations:"

$$[\widehat{J}_x, \widehat{J}_y] = \mathrm{i}\hbar\widehat{J}_z \quad [\widehat{J}_y, \widehat{J}_z] = \mathrm{i}\hbar\widehat{J}_x \quad [\widehat{J}_z, \widehat{J}_x] = \mathrm{i}\hbar\widehat{J}_y \tag{12.1}$$

They can be derived for orbital angular momentum (see chapter 4.5.4), but must be *postulated* to also apply to spin angular momentum {N.26}.

At first glance, these commutation relations do not look like a promising starting point for much analysis. All they say on their face is that the angular momentum operators $\widehat{J}_x$, $\widehat{J}_y$, and $\widehat{J}_z$ do not commute, so that they cannot have a full set of eigenstates in common. That is hardly impressive.

But if you read the following sections, you will be astonished by what knowledge can be teased out of them. For starters, one thing that immediately follows is that the *only* eigenstates that $\widehat{J}_x$, $\widehat{J}_y$, and $\widehat{J}_z$ have in common are states $|0\,0\rangle$ of no angular momentum at all {D.63}. No other common eigenstates exist.

One assumption will be implicit in the use of the fundamental commutation relations, namely that they can be taken at face value. It is certainly possible to imagine that say $\widehat{J}_x$ would turn an eigenfunction of say $\widehat{J}_z$ into some singular object for which angular momentum would be ill-defined. That would of course make application of the fundamental commutation relations improper. It will be assumed that the operators are free of such pathological nastiness.

12.3 Ladders

This section starts the quest to figure out everything that the fundamental commutation relations mean for angular momentum. It will first be verified that any angular momentum can always be described using $|j\,m\rangle$ eigenstates with definite values of square angular momentum J^2 and z angular momentum J_z. Then it will be found that these angular momentum states occur in groups called "ladders".

To start with the first one, the mathematical condition for a complete set of eigenstates $|j\,m\rangle$ to exist is that the angular momentum operators $\widehat{J}^2$ and $\widehat{J}_z$ commute. They do; using the commutator manipulations of chapter 4.5.4), it is easily found that:

$$[\widehat{J}^2, \widehat{J}_x] = [\widehat{J}^2, \widehat{J}_y] = [\widehat{J}^2, \widehat{J}_z] = 0 \quad \text{where } \widehat{J}^2 = \widehat{J}_x^2 + \widehat{J}_y^2 + \widehat{J}_z^2$$

So mathematics says that eigenstates $|j\,m\rangle$ of $\widehat{J}_z$ and $\widehat{J}^2$ exist satisfying

$$\widehat{J}_z|j\,m\rangle = J_z|j\,m\rangle \qquad \text{where } \textit{by definition } J_z = m\hbar \tag{12.2}$$

$$\widehat{J}^2|j\,m\rangle = J^2|j\,m\rangle \qquad \text{where } \textit{by definition } J^2 = j(j+1)\hbar^2 \text{ and } j \geqslant 0 \tag{12.3}$$

and that are complete in the sense that any state can be described in terms of these $|j\,m\rangle$.

Unfortunately the eigenstates $|j\,m\rangle$, except for $|0\,0\rangle$ states, do not satisfy relations like (12.2) for $\widehat{J}_x$ or $\widehat{J}_y$. The problem is that $\widehat{J}_x$ and $\widehat{J}_y$ do not commute with $\widehat{J}_z$. But $\widehat{J}_x$ and $\widehat{J}_y$ *do* commute with $\widehat{J}^2$, and you might wonder if that is still worth something. To find out, multiply, say, the zero commutator $[\widehat{J}^2, \widehat{J}_x]$ by $|j\,m\rangle$:

$$[\widehat{J}^2, \widehat{J}_x]|j\,m\rangle = (\widehat{J}^2\widehat{J}_x - \widehat{J}_x\widehat{J}^2)|j\,m\rangle = 0$$

Now take the second term to the right hand side of the equation, noting that $\widehat{J}^2|j\,m\rangle = J^2|j\,m\rangle$ with J^2 just a number that can be moved up-front, to get:

$$\widehat{J}^2\left(\widehat{J}_x|j\,m\rangle\right) = J^2\left(\widehat{J}_x|j\,m\rangle\right)$$

Looking a bit closer at this equation, it shows that the combination $\widehat{J}_x|j\,m\rangle$ satisfies the same eigenvalue problem for $\widehat{J}^2$ as $|j\,m\rangle$ itself. In other words, the multiplication by $\widehat{J}_x$ does not affect the square angular momentum J^2 at all.

To be picky, that is not quite true if $\widehat{J}_x|j\,m\rangle$ would be zero, because zero is not an eigenstate of anything. However, such a thing only happens if there is no angular momentum; (it would make $|j\,m\rangle$ an eigenstate of $\widehat{J}_x$ with eigenvalue zero in addition to an eigenstate of $\widehat{J}_z$ {D.63}). Except for that trivial case, $\widehat{J}_x$ does not affect square angular momentum. And neither does $\widehat{J}_y$ or any combination of the two.

Angular momentum in the z-direction is affected by $\widehat{J}_x$ and by $\widehat{J}_y$, since they do not commute with $\widehat{J}_z$ like they do with $\widehat{J}^2$. Nor is it possible to find any linear combination of $\widehat{J}_x$ and $\widehat{J}_y$ that does commute with $\widehat{J}_z$. What is the next best thing? Well, it *is* possible to find two combinations, to wit

$$\widehat{J}^+ \equiv \widehat{J}_x + \mathrm{i}\widehat{J}_y \quad \text{and} \quad \widehat{J}^- \equiv \widehat{J}_x - \mathrm{i}\widehat{J}_y, \tag{12.4}$$

that satisfy the "commutator eigenvalue problems":

$$[\widehat{J}_z, \widehat{J}^+] = \hbar\widehat{J}^+ \quad \text{and} \quad [\widehat{J}_z, \widehat{J}^-] = -\hbar\widehat{J}^-.$$

These two turn out to be quite remarkable operators.

Like $\widehat{J}_x$ and $\widehat{J}_y$, their combinations $\widehat{J}^+$ and $\widehat{J}^-$ leave J^2 alone. To examine what the operator $\widehat{J}^+$ does with the linear momentum in the z-direction, multiply its commutator relation above by an eigenstate $|j\,m\rangle$:

$$(\widehat{J}_z\widehat{J}^+ - \widehat{J}^+\widehat{J}_z)|j\,m\rangle = \hbar\widehat{J}^+|j\,m\rangle$$

Or, taking the second term to the right hand side of the equation and noting that by definition $\widehat{J}_z|j\,m\rangle = m\hbar|j\,m\rangle$,

$$\widehat{J}_z\left(\widehat{J}^+|j\,m\rangle\right) = (m+1)\hbar\left(\widehat{J}^+|j\,m\rangle\right)$$

That is a stunning result, as it shows that $\widehat{J}^+|j\,m\rangle$ is an eigenstate with z angular momentum $J_z = (m+1)\hbar$ instead of $m\hbar$. In other words, $\widehat{J}^+$ adds exactly one unit $\hbar$ to the z angular momentum, turning an $|j\,m\rangle$ state into a $|j\,m{+}1\rangle$ one!

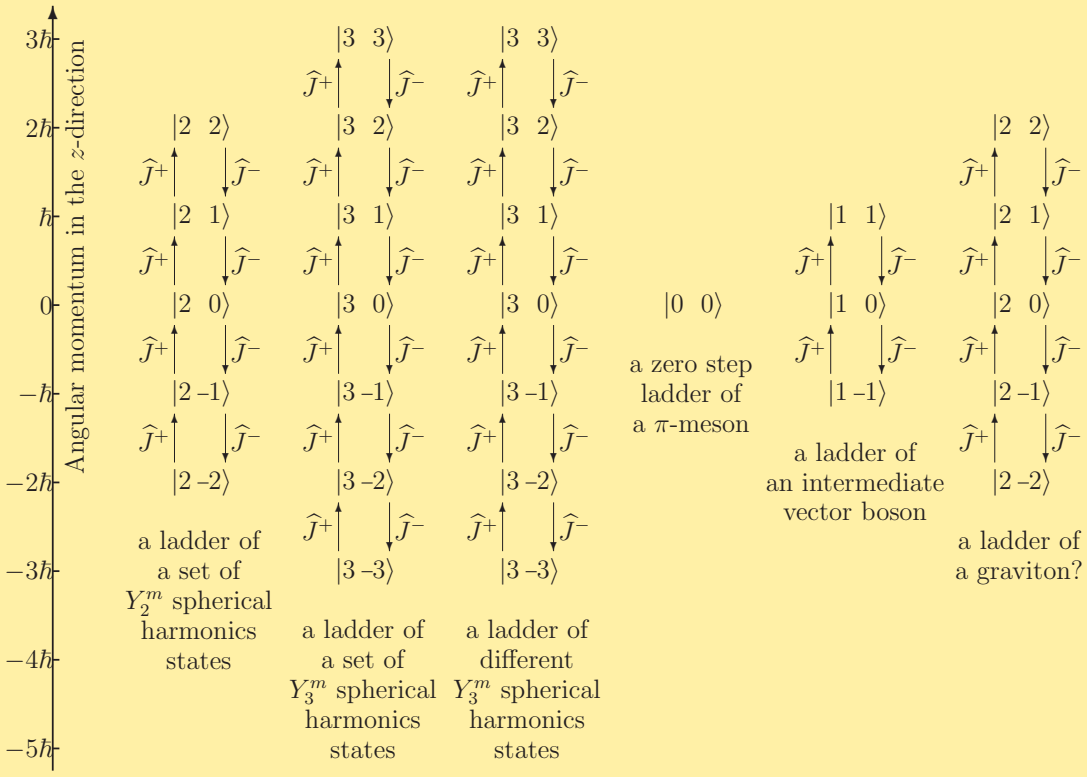


Figure 12.1: Example bosonic ladders.

If you apply $\widehat{J}^+$ another time, you get a state of still higher z angular momentum $|j\,m{+}2\rangle$, and so on, like the rungs on a ladder. This is graphically illustrated for some examples in figures 12.1 and 12.2. The process eventually comes to an halt at some top rung $m = m_{\max}$ where $\widehat{J}^+|j\,m_{\max}\rangle = 0$. It has to, because the angular momentum in the z-direction cannot just keep growing forever: the square angular momentum in the z-direction only must stay less than the total square angular momentum in all three directions {N.27}.

The second "ladder operator" $\widehat{J}^-$ works in much the same way, but it goes down the ladder; its deducts one unit $\hbar$ from the angular momentum in the z-direction at each application. $\widehat{J}^-$ provides the second stile to the ladders, and must terminate at some bottom rung $m_{\min}$.

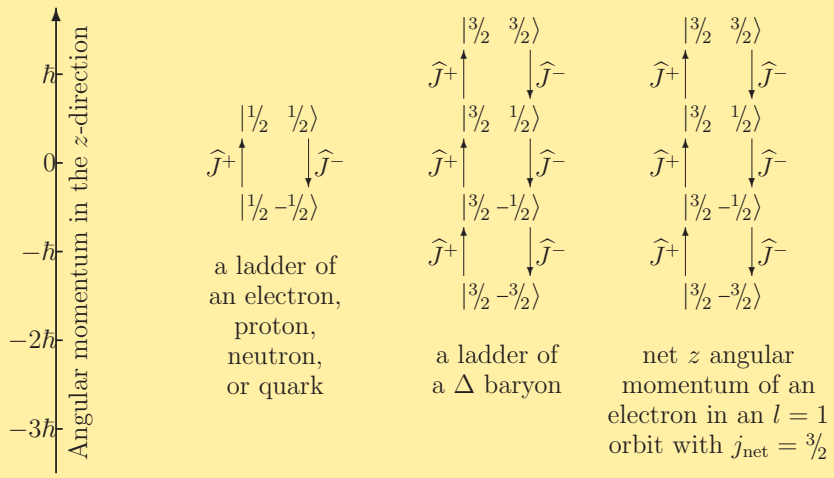


Figure 12.2: Example fermionic ladders.

12.4 Possible values of angular momentum

The fact that the angular momentum ladders of the previous section must have a top and a bottom rung restricts the possible values that angular momentum can take. This section will show that the azimuthal quantum number j can either be a nonnegative whole number or half of one, but nothing else. And it will show that the magnetic quantum number m must range from $-j$ to $+j$ in unit increments. In other words, the bosonic and fermionic example ladders in figures 12.1 and 12.2 are representative of all that is possible.

To start, in order for a ladder to end at a top rung $m_{\max}$, $\widehat{J}^+|l\,m\rangle$ has to be zero for $m = m_{\max}$. More specifically, its magnitude $\left|\widehat{J}^+|j\,m\rangle\right|$ must be zero. The square magnitude is given by the inner product with itself:

$$\left|\widehat{J}^+|j\,m\rangle\right|^2 = \left\langle \widehat{J}^+|j\,m\rangle \middle| \widehat{J}^+|j\,m\rangle \right\rangle = 0.$$

Now because of the complex conjugate that is used in the left hand side of an inner product, (see chapter 2.3), $\widehat{J}^+ = \widehat{J}_x + \mathrm{i}\widehat{J}_y$ goes to the other side of the product as $\widehat{J}^- = \widehat{J}_x - \mathrm{i}\widehat{J}_y$, and you must have

$$\left|\widehat{J}^+|j\,m\rangle\right|^2 = \left\langle |j\,m\rangle \middle| \widehat{J}^-\widehat{J}^+|j\,m\rangle \right\rangle$$

That operator product can be multiplied out:

$$\widehat{J}^-\widehat{J}^+ \equiv (\widehat{J}_x - \mathrm{i}\widehat{J}_y)(\widehat{J}_x + \mathrm{i}\widehat{J}_y) = \widehat{J}_x^2 + \widehat{J}_y^2 + \mathrm{i}(\widehat{J}_x\widehat{J}_y - \widehat{J}_y\widehat{J}_x),$$

but $\widehat{J}_x^2 + \widehat{J}_y^2$ is the square angular momentum $\widehat{J}^2$ except for $\widehat{J}_z^2$, and the term within the parentheses is the commutator $[\widehat{J}_x, \widehat{J}_y]$ which is according to the

fundamental commutation relations equal to $\mathrm{i}\hbar\widehat{J}_z$, so

$$\widehat{J}^-\widehat{J}^+ = \widehat{J}^2 - \widehat{J}_z^2 - \hbar\widehat{J}_z \tag{12.5}$$

The effect of each of the operators in the left hand side on a state $|j\,m\rangle$ is known and the inner product can be figured out:

$$\left|\widehat{J}^+|j\,m\rangle\right|^2 = j(j+1)\hbar^2 - m^2\hbar^2 - m\hbar^2 \tag{12.6}$$

The question where angular momentum ladders end can now be answered:

$$j(j+1)\hbar^2 - m_{\max}^2\hbar^2 - m_{\max}\hbar^2 = 0$$

There are two possible solutions to this quadratic equation for $m_{\max}$, to wit $m_{\max} = j$ or $-m_{\max} = j+1$. The second solution is impossible since it already would have the square z angular momentum exceed the total square angular momentum. So unavoidably,

$$m_{\max} = j$$

That is one of the things this section was supposed to show.

The lowest rung on the ladder goes the same way; you get

$$\widehat{J}^+\widehat{J}^- = \widehat{J}^2 - \widehat{J}_z^2 + \hbar\widehat{J}_z \tag{12.7}$$

and then

$$\left|\widehat{J}^-|j\,m\rangle\right|^2 = j(j+1)\hbar^2 - m^2\hbar^2 + m\hbar^2 \tag{12.8}$$

and the only acceptable solution for the lowest rung on the ladders is

$$m_{\min} = -j$$

It is nice and symmetric; ladders run from $m = -j$ up to $m = j$, as the examples in figures 12.1 and 12.2 already showed.

And in fact, it is more than that; it also limits what the quantum numbers j and m can be. For, since each step on a ladder increases the magnetic quantum number m by one unit, you have for the total number of steps up from bottom to top:

$$\text{total number of steps } = m_{\max} - m_{\min} = 2j$$

But the number of steps is a whole number, and so the azimuthal quantum j must either be a nonnegative integer, such as 0, 1, 2, ..., or half of one, such as $\frac{1}{2}$, $\frac{3}{2}$, ...

Integer j values occur, for example, for the spherical harmonics of orbital angular momentum and for the spin of bosons like photons. Half-integer values occur, for example, for the spin of fermions such as electrons, protons, neutrons, and Δ particles.

Note that if j is a half-integer, then so are the corresponding values of m, since m starts from $-j$ and increases in unit steps. See again figures 12.1 and 12.2 for some examples. Also note that ladders terminate just before z-momentum would exceed total momentum.

It may also be noted that ladders are distinct. It is not possible to go up one ladder, like the first Y_3^m one in figure 12.1 with $\widehat{J}^+$ and then come down the second one using $\widehat{J}^-$. The reason is that the states $|j\,m\rangle$ are eigenstates of the operators $\widehat{J}^-\widehat{J}^+$, (12.5), and $\widehat{J}^+\widehat{J}^-$, (12.7), so going up with $\widehat{J}^+$ and then down again with $\widehat{J}^-$, or vice-versa, returns to the same state. For similar reasons, if the tops of two ladders are orthonormal, then so is the rest of their rungs.

12.5 A warning about angular momentum

Normally, eigenstates are indeterminate by a complex number of magnitude one. If you so desire, you can multiply any normalized eigenstate by a number of unit magnitude of your own choosing, and it is still a normalized eigenstate. It is important to remember that in analytical expressions involving angular momentum, you are *not* allowed to do this.

As an example, consider a pair of spin 1/2 particles, call them a and b, in the "singlet state", in which their spins cancel and there is no net angular momentum. It was noted in chapter 5.5.6 that this state takes the form

$$|0\,0\rangle_{ab} = \frac{|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_b - |\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b}{\sqrt{2}}$$

(This section will use kets rather than arrows for spin states.) But if you were allowed to arbitrarily change the definition of say the spin state $|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_a$ by a minus sign, then the minus sign in the singlet state above would turn in a plus sign. The given expression for the singlet state, with its minus sign, is only correct if you use the right normalization factors for the individual states.

It all has to do with the ladder operators $\widehat{J}^+$ and $\widehat{J}^-$. They are very convenient for analysis, but to make that easiest, you would like to know *exactly* what they do to the angular momentum states $|j\,m\rangle$. What you have seen so far is that $\widehat{J}^+|j\,m\rangle$ produces a state with the same square angular momentum, and with angular momentum in the z-direction equal to $(m+1)\hbar$. In other words, $\widehat{J}^+|j\,m\rangle$ is some multiple of a suitably normalized eigenstate $|j\,m{+}1\rangle$;

$$\widehat{J}^+|j\,m\rangle = C|j\,m{+}1\rangle$$

where the number C is the multiple. What *is* that multiple? Well, from the magnitude of $\widehat{J}^+|j\,m\rangle$, derived earlier in (12.6) you know that its square magnitude is

$$|C|^2 = j(j+1)\hbar^2 - m^2\hbar^2 - m\hbar^2.$$

But that still leaves C indeterminate by a factor of unit magnitude. Which would be very inconvenient in the analysis of angular momentum.

To resolve this conundrum, restrictions are put on the normalization factors of the angular momentum states $|j\,m\rangle$ in ladders. It is required that the normalization factors are chosen such that the ladder operator constants are positive real numbers. That really leaves only *one* normalization factor in an entire ladder freely selectable, say the one of the top rung.

Most of the time, this is not a big deal. Only when you start trying to get too clever with angular momentum normalization factors, then you want to remember that you cannot really choose them to your own liking.

The good news is that in this convention, you know *precisely* what the ladder operators do {D.64}:

$$\widehat{J}^+|j\,m\rangle = \hbar\sqrt{j(j+1)-m(1+m)}\;|j\,m{+}1\rangle \tag{12.9}$$

$$\widehat{J}^-|j\,m\rangle = \hbar\sqrt{j(j+1)+m(1-m)}\;|j\,m{-}1\rangle \tag{12.10}$$

12.6 Triplet and singlet states

With the ladder operators, you can determine how different angular momenta add up to net angular momentum. As an example, this section will examine what net spin values can be produced by two particles, each with spin $\frac{1}{2}$. They may be the proton and electron in a hydrogen atom, or the two electrons in the hydrogen molecule, or whatever. The actual result will be to rederive the triplet and singlet states described in chapter 5.5.6, but it will also be an example for how more complex angular momentum states can be combined.

The particles involved will be denoted as a and b. Since each particle can have two different spin states $|\frac{1}{2}\,\frac{1}{2}\rangle$ and $|\frac{1}{2}\,{-\frac{1}{2}}\rangle$, there are four different combined "product" states:

$$|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b,\ |\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_b,\ |\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b,\ \text{and}\ |\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_b.$$

In these product states, each particle is in a single individual spin state. The question is, what is the combined angular momentum of these four product states? And what combination states have definite net values for square and z angular momentum?

The angular momentum in the z-direction is simple; it is just the sum of those of the individual particles. For example, the z-momentum of the $|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$ state follows from

$$\begin{aligned}\left(\widehat{J}_{za}+\widehat{J}_{zb}\right)|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b &= \tfrac{1}{2}\hbar|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b + |\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a\tfrac{1}{2}\hbar|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b \\ &= h|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b\end{aligned}$$

which makes the net angular momentum in the z-direction $\hbar$, or $\frac{1}{2}\hbar$ from each particle. Note that the z angular momentum operators of the two particles simply add up and that $\widehat{J}_{za}$ only acts on particle a, and $\widehat{J}_{zb}$ only on particle b {N.28}. In terms of quantum numbers, the magnetic quantum number m_{ab} is the sum of the individual quantum numbers m_a and m_b; $m_{ab} = m_a + m_b = 1$.

The net total angular momentum is not so obvious; you cannot just add total angular momenta. To figure out the total angular momentum of $|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$ anyway, there is a trick: multiply it with the combined step-up operator

$$\widehat{J}^+_{ab} = \widehat{J}^+_a + \widehat{J}^+_b$$

Each part returns zero: $\widehat{J}^+_a$ because particle a is at the top of its ladder and $\widehat{J}^+_b$ because particle b is. So the combined state $|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$ must be at the top of the ladder too; there is no higher rung. That must mean $j_{ab} = m_{ab} = 1$; the combined state must be a $|1\,1\rangle$ state. It can be *defined* it as *the* combination $|1\,1\rangle$ state:

$$|1\,1\rangle_{ab} \equiv |\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b \tag{12.11}$$

You could just as well have defined $|1\,1\rangle_{ab}$ as $-|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$ or $\mathrm{i}|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$, say. But why drag along a minus sign or i if you do not have to? The first triplet state has been found.

Here is another trick: multiply $|1\,1\rangle_{ab} = |\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$ by $\widehat{J}^-_{ab}$: that will go one step down the combined states ladder and produce a combination state $|1\,0\rangle_{ab}$:

$$\begin{aligned}\widehat{J}^-_{ab}|1\,1\rangle_{ab} &= \hbar\sqrt{1(1+1)+1(1-1)}|1\,0\rangle_{ab} \\ &= \widehat{J}^-_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b + \widehat{J}^-_b|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b\end{aligned}$$

or

$$\hbar\sqrt{2}|1\,0\rangle_{ab} = \hbar|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b + \hbar|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_b$$

where the effects of the ladder-down operators were taken from (12.10). (Note that this requires that the individual particle spin states are normalized consistent with the ladder operators.) The second triplet state is therefore:

$$|1\,0\rangle_{ab} \equiv \sqrt{\tfrac{1}{2}}\,|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_a|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_b + \sqrt{\tfrac{1}{2}}\,|\tfrac{1}{2}\,{-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\,\tfrac{1}{2}\rangle_b \tag{12.12}$$

But this gives only *one* $|j\,m\rangle$ combination state for the *two* product states $|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,{-\frac{1}{2}}\rangle_b$ and $|\frac{1}{2}\,{-\frac{1}{2}}\rangle_a|\frac{1}{2}\,\frac{1}{2}\rangle_b$ with zero net z-momentum. If you want to describe unequal combinations of them, like $|\frac{1}{2}\,\frac{1}{2}\rangle_a|\frac{1}{2}\,{-\frac{1}{2}}\rangle_b$ by itself, it cannot be just a multiple of $|1\,0\rangle_{ab}$. This suggests that there may be another $|j\,0\rangle_{ab}$ combination state involved here. How do you get this second state?

Well, you can reuse the first trick. If you construct a combination of the two product states that steps up to zero, it must be a state with zero z angular momentum that is at the end of its ladder, a $|0\,0\rangle_{ab}$ state. Consider an

arbitrary combination of the two product states with as yet unknown numerical coefficients C_1 and C_2:

$$C_1|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_a|\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_b + C_2|\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_b$$

For this combination to step up to zero,

$$\left(\widehat{J}_a^+ + \widehat{J}_b^+\right)\left(C_1|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_a|\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_b + C_2|\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_b\right)$$
$$= \hbar C_1|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_a|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_b + \hbar C_2|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_a|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_b$$

must be zero, which requires $C_2 = -C_1$, leaving C_1 undetermined. C_1 must be chosen such that the state is normalized, but that still leaves a constant of magnitude one undetermined. To fix it, C_1 is taken to be real and positive, and so the singlet state becomes

$$|0\ 0\rangle_{ab} = \sqrt{\tfrac{1}{2}}\ |\tfrac{1}{2}\ \tfrac{1}{2}\rangle_a|\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_b - \sqrt{\tfrac{1}{2}}\ |\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_b. \tag{12.13}$$

To find the remaining triplet state, just apply $\widehat{J}_{ab}^-$ once more, to $|1\ 0\rangle_{ab}$ above. It gives:

$$|1\ {-1}\rangle_{ab} = |\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_a|\tfrac{1}{2}\ {-\tfrac{1}{2}}\rangle_b \tag{12.14}$$

Of course, the normalization factor of this bottom state had to turn out to be one; all three step-down operators produce only positive real factors.

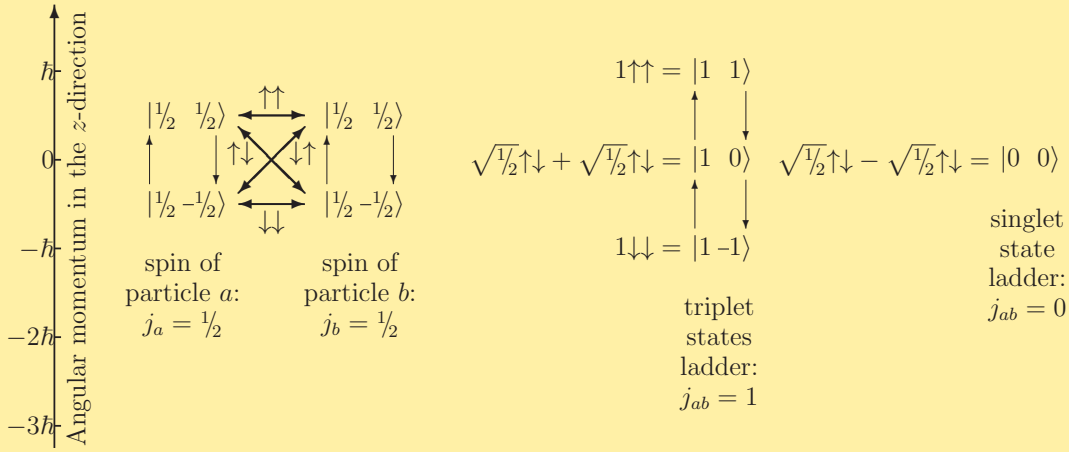


Figure 12.3: Triplet and singlet states in terms of ladders

Figure 12.3 shows the results graphically in terms of ladders. The two possible spin states of each of the two electrons produce 4 combined product states indicated using up and down arrows. These product states are then combined to produce triplet and singlet states that have definite values for both z and total net angular momentum, and can be shown as rungs on ladders.

Note that a product state like $|\frac{1}{2}\ \frac{1}{2}\rangle_a|\frac{1}{2}\ -\frac{1}{2}\rangle_b$ cannot be shown as a rung on a ladder. In fact, from adding (12.12) and (12.13) it is seen that

$$|\tfrac{1}{2}\ \tfrac{1}{2}\rangle_a|\tfrac{1}{2}\ -\tfrac{1}{2}\rangle_b = \sqrt{\tfrac{1}{2}}\,|1\ 0\rangle_{ab} + \sqrt{\tfrac{1}{2}}\,|0\ 0\rangle_{ab}$$

which makes it a combination of the middle rungs of the triplet and singlet ladders, rather than a single rung.

12.7 Clebsch-Gordan coefficients

In classical physics, combining angular momentum from different sources is easy; the net components in the x, y, and z directions are simply the sum of the individual components. In quantum mechanics, things are trickier, because if the component in the z-direction exists, those in the x and y directions do not. But the previous subsection showed how to the spin angular momenta of two spin $\frac{1}{2}$ particles could be combined. In similar ways, the angular momentum states of any two ladders, whatever their origin, can be combined into net angular momentum ladders. And then those ladders can in turn be combined with still other ladders, allowing net angular momentum states to be found for systems of arbitrary complexity.

The key is to be able to combine the angular momentum ladders from two different sources into net angular momentum ladders. To do so, the net angular momentum can in principle be described in terms of product states in which each source is on a single rung of its ladder. But as the example of the last section illustrated, such product states give incomplete information about the net angular momentum; they do not tell you what square net angular momentum is. You need to know what combinations of product states produce rungs on the ladders of the net angular momentum, like the ones illustrated in figure 12.3. In particular, you need to know the coefficients that multiply the product states in those combinations.

$\underline{\lvert 1\ -1\rangle_{ab}}$	$\underline{\lvert 1\ 0\rangle_{ab}}$	$\underline{\lvert 0\ 0\rangle_{ab}}$	$\underline{\lvert 1\ 1\rangle_{ab}}$	
			$\boxed{1}$	$\lvert\frac{1}{2}\ \frac{1}{2}\rangle_a\lvert\frac{1}{2}\ \frac{1}{2}\rangle_b$
	$\sqrt{\frac{1}{2}}$	$\sqrt{\frac{1}{2}}$		$\lvert\frac{1}{2}\ \frac{1}{2}\rangle_a\lvert\frac{1}{2}\ -\frac{1}{2}\rangle_b$
	$-\sqrt{\frac{1}{2}}$	$\sqrt{\frac{1}{2}}$		$\lvert\frac{1}{2}\ -\frac{1}{2}\rangle_a\lvert\frac{1}{2}\ \frac{1}{2}\rangle_b$
$\boxed{1}$				$\lvert\frac{1}{2}\ -\frac{1}{2}\rangle_a\lvert\frac{1}{2}\ -\frac{1}{2}\rangle_b$

Figure 12.4: Clebsch-Gordan coefficients of two spin one half particles.

These coefficients are called "Clebsch-Gordan" coefficients. The ones corresponding to figure 12.3 are tabulated in Figure 12.4. Note that there are really

three tables of numbers; one for each rung level. The top, single number, "table" says that the $|1\,1\rangle$ net momentum state is found in terms of product states as:

$$|1\,1\rangle_{ab} = 1 \times |{}^1\!/_2\ {}^1\!/_2\rangle_a |{}^1\!/_2\ {}^1\!/_2\rangle_b$$

The second table gives the states with zero net angular momentum in the z-direction. For example, the first column of the table says that the $|0\,0\rangle$ singlet state is found as:

$$|0\,0\rangle_{ab} = \sqrt{{}^1\!/_2}\ |{}^1\!/_2\ {}^1\!/_2\rangle_a |{}^1\!/_2\ {}^{-1}\!/_2\rangle_b - \sqrt{{}^1\!/_2}\ |{}^1\!/_2\ {}^{-1}\!/_2\rangle_a |{}^1\!/_2\ {}^1\!/_2\rangle_b$$

Similarly the second column gives the middle rung $|1\,0\rangle$ on the triplet ladder. The bottom "table" gives the bottom rung of the triplet ladder.

You can also read the tables horizontally {N.29}. For example, the first row of the middle table says that the $|{}^1\!/_2\ {}^1\!/_2\rangle_a |{}^1\!/_2\ {}^{-1}\!/_2\rangle_b$ product state equals

$$|{}^1\!/_2\ {}^1\!/_2\rangle_a |{}^1\!/_2\ {}^{-1}\!/_2\rangle_b = \sqrt{{}^1\!/_2}\ |0\,0\rangle_{ab} + \sqrt{{}^1\!/_2}\ |1\,0\rangle_{ab}$$

That in turn implies that if the net square angular momentum of this product state is measured, there is a 50/50 chance of it turning out to be either zero, or the $j = 1$ (i.e. $2\hbar^2$) value. The z-momentum will always be zero.

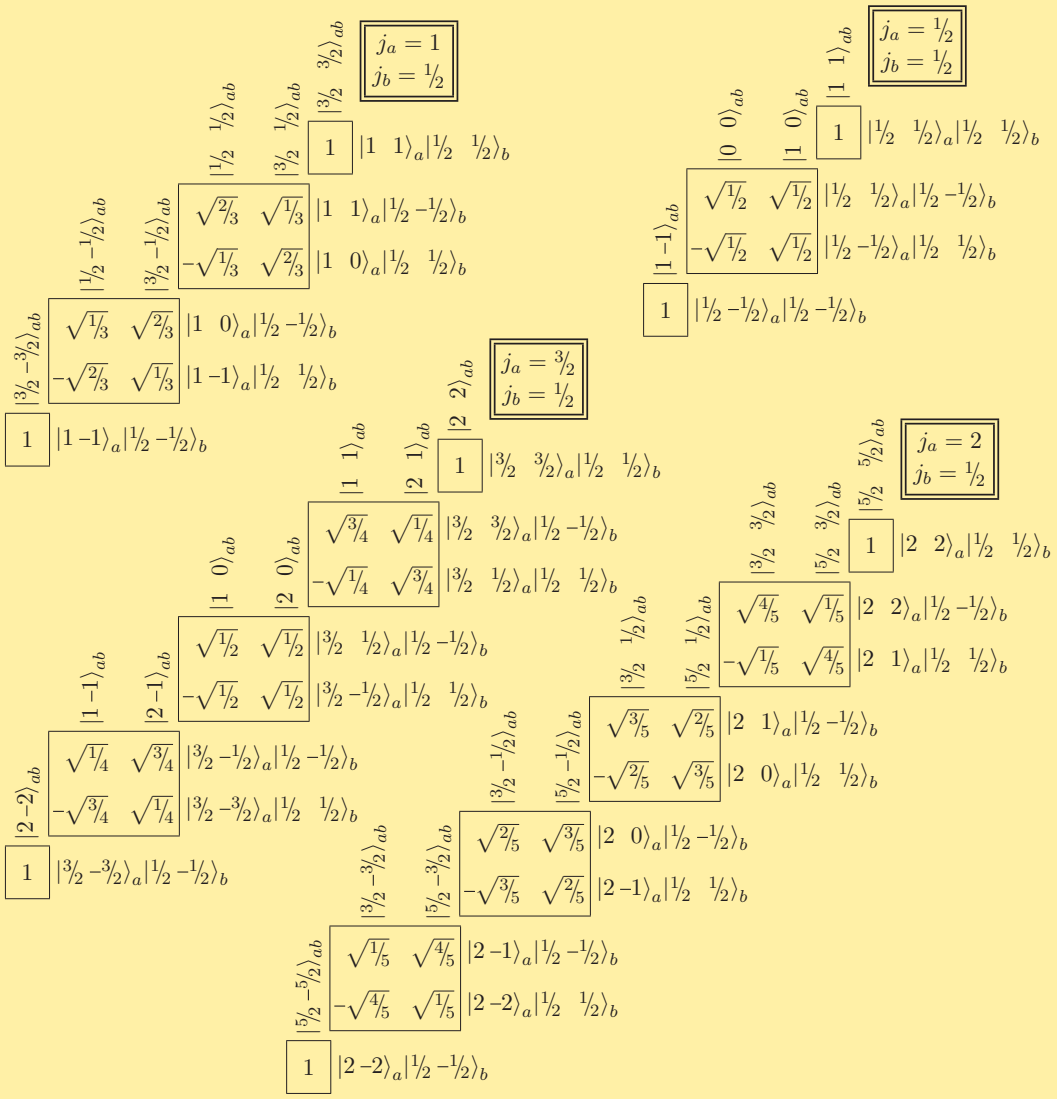


Figure 12.5: Clebsch-Gordan coefficients when the second angular momentum contribution has azimuthal quantum number $j_b = \frac{1}{2}$.

How about the Clebsch-Gordan coefficients to combine other ladders than the spins of two spin ½ particles? Well, the same procedures used in the previous section work just as well to combine the angular momenta of any two angular momentum ladders, whatever their size. Just the thing for a long winter night. Or, if you live in Florida, you just might want to write a little computer program that does it for you {D.65} and outputs the tables in human-readable form {N.30}, like figures 12.5 and 12.6.

From the figures you may note that when two states with total angular momentum quantum numbers j_a and j_b are combined, the combinations have total angular quantum numbers ranging from $j_a + j_b$ to $|j_a - j_b|$. This is similar to the fact that when in classical mechanics two angular momentum vectors are combined, the combined total angular momentum J_{ab} is at most $J_a + J_b$ and at least $|J_a - J_b|$. (The so-called "triangle inequality" for combining vectors.) But of course, j is not quite a proportional measure of J unless J is large; in fact, $J = \sqrt{j(j+1)}\hbar$ {D.66}.

12.8 Some important results

This section gives some results that are used frequently in quantum analysis, but usually not explicitly stated.

First a note on notations. It is fairly common to use the letter l for orbital angular momentum, s for spin, and j for combinations of orbital and angular momentum. This subsection will follow these conventions where appropriate.

1. If all possible angular momentum states are filled with a fermion, the resulting angular momentum is zero and the wave function is spherically symmetric. For example, consider the simplified case that there is one spinless fermion in each spherical harmonic at a given azimuthal quantum number l. Then it is easy to see from the form of the spherical harmonics that the combined wave function is independent of the angular position around the z-axis. And all spherical harmonics at that l are filled whatever you take to be the z-axis. This makes noble gasses into the equivalent of billiard balls. More generally, if there is one fermion for every possible "direction" of the angular momentum, by symmetry the net angular momentum can only be zero.
2. If a spin ½ fermion has orbital angular momentum quantum number l, net (orbital plus spin) angular momentum quantum number $j = l + \frac{1}{2}$, and net momentum in the z-direction quantum number m_j, its net state is given in terms of the individual orbital and spin states as:

$$j = l + \tfrac{1}{2}: \quad |j\, m_j\rangle = \sqrt{\frac{j+m_j}{2j}} Y_l^{m_j - \frac{1}{2}} \uparrow + \sqrt{\frac{j-m_j}{2j}} Y_l^{m_j + \frac{1}{2}} \downarrow$$

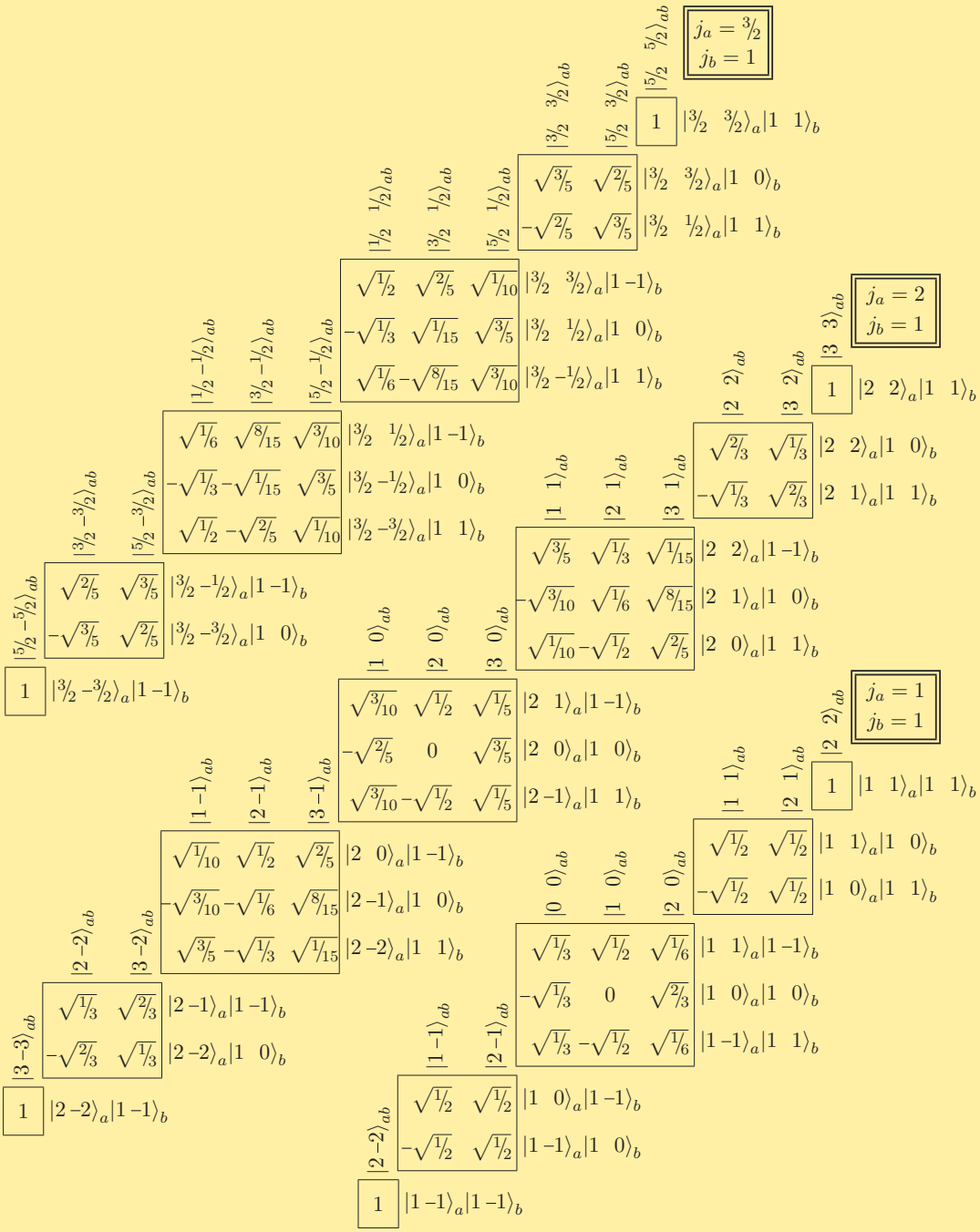


Figure 12.6: Clebsch-Gordan coefficients when the second angular momentum contribution has azimuthal quantum number $j_b = 1$.

If the net spin is $j = l - \frac{1}{2}$, assuming that $l > 0$, that becomes

$$j = l - \tfrac{1}{2}: \quad |j\, m_j\rangle = -\sqrt{\frac{j+1-m_j}{2j+2}}\, Y_l^{m_j-\frac{1}{2}}\uparrow + \sqrt{\frac{j+1+m_j}{2j+2}}\, Y_l^{m_j+\frac{1}{2}}\downarrow$$

Note that if the net angular momentum is unambiguous, the orbital and spin magnetic quantum numbers m and m_s are in general uncertain.

3. For identical particles, an important question is how the Clebsch-Gordan coefficients change under particle exchange:

$$\langle j_{ab} m_{ab} || j_a\, m_a\rangle |j_b\, m_b\rangle = (-1)^{j_a+j_b-j_{ab}} \langle j_{ab} m_{ab} || j_b\, m_b\rangle |j_a\, m_a\rangle$$

For $j_a = j_b = \frac{1}{2}$, this verifies that the triplet states $j_{ab} = 1$ are symmetric, and the singlet state $j_{ab} = 0$ is antisymmetric. More generally, states with the maximum net angular momentum $j_{ab} = j_a + j_b$ and whole multiples of 2 less are symmetric under particle exchange. States that are odd amounts less than the maximum are antisymmetric under particle exchange.

4. When the net angular momentum state is swapped with one of the component states, the relation is

$$\langle j_{ab} m_{ab} || j_a\, m_a\rangle |j_b\, m_b\rangle =$$
$$(-1)^{j_a-j_{ab}+m_b} \sqrt{\frac{2j_{ab}+1}{2j_a+1}} \langle j_a m_a || j_{ab}\, m_{ab}\rangle |j_b\, {-m_b}\rangle$$

This is of interest in figuring out what states produce zero net angular momentum, $j_{ab} = m_{ab} = 0$. In that case, the right hand side is zero unless $j_b = j_a$ and $m_b = -m_a$; and then $\langle j_a m_a || 0\, 0\rangle |j_a\, m_a\rangle = 1$. You can only create zero angular momentum from a pair of particles that have the same square angular momentum; also, only product states with zero net angular momentum in the z-direction are involved.

12.9 Momentum of partially filled shells

One very important case of combining angular momenta occurs for both electrons in atoms and nucleons in nuclei. In these problems there are a number of identical fermions in single-particle states that differ only in the net (orbital plus spin) momentum in the chosen z-direction. Loosely speaking, the single-particle states are the same, just at different angular orientations. Such a set of states is often called a "shell." The question is then: what combinations of the

		possible combined angular momentum j																	
j^{p}	I	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$	$\frac{9}{2}$	$\frac{11}{2}$	$\frac{13}{2}$	$\frac{15}{2}$	$\frac{17}{2}$	$\frac{19}{2}$	$\frac{21}{2}$	$\frac{23}{2}$	$\frac{25}{2}$	$\frac{27}{2}$	$\frac{29}{2}$	$\frac{31}{2}$	$\frac{33}{2}$	$\frac{35}{2}$
$\frac{1}{2}$	1	1																	
$\frac{3}{2}$	1		1																
$\frac{5}{2}$	1			1															
	3		1	1		1													
$\frac{7}{2}$	1				1														
	3		1	1	1	1	1		1										
$\frac{9}{2}$	1					1													
	3		1	1	1	2	1	1	1	1		1							
	5	1	1	2	2	3	2	2	2	2	1	1		1					
$\frac{11}{2}$	1						1												
	3		1	1	1	2	2	1	2	1	1	1	1		1				
	5	1	2	3	4	4	5	4	5	4	4	3	3	2	2	1	1		1

		possible combined angular momentum j																	
j^{p}	I	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\frac{3}{2}$	2	1		1																
$\frac{5}{2}$	2	1		1		1														
$\frac{7}{2}$	2	1		1		1		1												
	4	1		2		2	1	1		1										
$\frac{9}{2}$	2	1		1		1		1		1										
	4	2		2	1	3	1	3	1	2	1	1		1						
$\frac{11}{2}$	2	1		1		1		1		1		1								
	4	2		3	1	4	2	4	2	4	2	3	1	2	1	1		1		
	6	3		4	3	6	3	7	4	6	4	5	2	4	2	2	1	1		1

Table 12.1: Possible combined angular momentum of identical fermions in shells of single-particle states that differ in magnetic quantum number. The top shows odd numbers of particles, the bottom even numbers.

states are antisymmetric with respect to exchange of the fermions, and therefore allowed? More specifically, what is their *combined* net angular momentum?

The answer is given in table 12.1, {D.67}. In it, I is the number of fermions in the shell. Further j^{p} is the net angular momentum of the single-particle states that make up the shell. (Or the azimuthal quantum number of that angular momentum really.) Similarly the values of j indicate the possible net angular momentum quantum numbers of all i fermions combined. The main body of the table lists the multiplicity of sets with the given angular momentum. Note that the table is split into odd and even numbers of particles. That simplifies the presentation, because odd numbers of particles produce only half-integer net angular momentum, and even numbers only integer net angular momentum.

For example, consider a single particle, $I = 1$, in a set of single-particle states with angular momentum $j^{\mathrm{p}} = \frac{9}{2}$. For a single particle, the "combined" momentum j is simply the single particle momentum j^{p}, explaining the single 1 in the $\frac{9}{2}$ column. But note that the 1 stands for a set of states; the magnetic net quantum number m^{p} of the single particle could still be any one of $\frac{9}{2}$, $\frac{7}{2}$, $\ldots$, $-\frac{9}{2}$. All the ten states in this set have net angular momentum $j = j^{\mathrm{p}} = \frac{9}{2}$.

Next assume that there are two particles in the same $j^{\mathrm{p}} = \frac{9}{2}$ single-particle states. Then if both particles would be in the $m^{\mathrm{p}} = \frac{9}{2}$ single-particle state, their combined angular momentum in the z-direction m would be $2 \times \frac{9}{2} = 9$. Following the Clebsch-Gordan derivation shows that this state would have combined angular momentum $j = m = 9$. But the two identical fermions cannot be both in the $m^{\mathrm{p}} = \frac{9}{2}$ state; that violates the Pauli exclusion principle. That is why there is no entry in the $j = 9$ column. If the first particle is in the $m^{\mathrm{p}} = \frac{9}{2}$ state, the second one can at most be in the $m^{\mathrm{p}} = \frac{7}{2}$ state, for a total of $m = 8$. More precisely, the particles would have to be in the antisymmetric combination, or Slater determinant, of these two states. That antisymmetric combination can be seen to have combined angular momentum $j = 8$. There are other combinations of states that also have $j = 8$, but values of m equal to 7, 6, $\ldots$, -8, for a total of 17 states. That set of 17 states is indicated by the 1 in the $j = 8$ column.

It is also possible for the two $j^{\mathrm{p}} = \frac{9}{2}$ particles to combine their angular momentum into smaller even values of the total angular momentum j. In fact, it is possible for the particles to combine their angular momenta so that they exactly cancel one another; then the net angular momentum $j = 0$. That is indicated by the 1 in the $j = 0$ column. Classically you would say that the momentum vectors of the two particles are exactly opposite, producing a zero resultant. In quantum mechanics true angular momentum vectors do not exist due to uncertainty of the components, but complete cancelation is still possible.

The $j = 0$ set consists of just one state, because m can only be zero for a state with zero angular momentum. The entire table row for two $j^{\mathrm{p}} = \frac{9}{2}$ particles could in principle be derived by writing out the appropriate Clebsch-Gordan

coefficients. But that would be one very big table.

If there are five $j^{\rm p} = \frac{9}{2}$ particles. they can combine their angular momenta into quite a wide variety of net angular momentum values. For example, the 2 in the $j = \frac{5}{2}$ column indicates that there are two sets of states with combined angular momentum $j = \frac{5}{2}$. Each set has 6 members, because for each set m can be any one of $\frac{5}{2}$, $\frac{3}{2}$, …, $-\frac{5}{2}$. So there are a total of 12 independent combination states that have net angular momentum $j = \frac{5}{2}$.

Note that a shell has $2j^{\rm p} + 1$ different single-particle states, because the magnetic quantum number $m^{\rm p}$ can have the values $j^{\rm p}$, $j^{\rm p}-1$, …, $-j^{\rm p}$. Therefore a shell can accommodate up to $2j^{\rm p} + 1$ fermions according to the exclusion principle. However, the table only lists combined angular momentum values for up to $j^{\rm p} + \frac{1}{2}$ particles. The reason is that any more is unnecessary. A given number of "holes" in an otherwise filled shell produces the same combined angular momentum values as the same number of particles in an otherwise empty shell. For example, two fermions in a $j^{\rm p} = \frac{1}{2}$ shell, (zero holes), have the same combined angular momentum as zero particles: zero. Indeed, those two fermions must be in the antisymmetric singlet state with spin zero. In general, a completely filled shell has zero angular momentum and is spherically symmetric.

The same situation for identical bosons is shown in table 12.2. For identical bosons there is no limit to the number of particles that can go into a shell. The table was arbitrarily cut off at 9 particles and a maximum spin of 18.

12.10 Pauli spin matrices

This subsection returns to the simple two-rung spin ladder (doublet) of an electron, or any other spin $\frac{1}{2}$ particle for that matter, and tries to tease out some more information about the spin. While the analysis so far has made statements about the angular momentum in the arbitrarily chosen z-direction, you often also need information about the spin in the corresponding x and y directions. This subsection will find it.

But before getting at it, a matter of notations. It is customary to indicate angular momentum that is due to spin by a capital S. Similarly, the azimuthal quantum number of spin is indicated by s. This subsection will follow this convention.

Now, suppose you know that the particle is in the "spin-up" state with $S_z = \frac{1}{2}\hbar$ angular momentum in a chosen z direction; in other words that it is in the $|\frac{1}{2}\,\frac{1}{2}\rangle$, or $\uparrow$, state. You want the effect of the $\widehat{S}_x$ and $\widehat{S}_y$ operators on this state. In the absence of a physical model for the motion that gives rise to the spin, this may seem like a hard question indeed. But again the faithful ladder operators $\widehat{S}^+$ and $\widehat{S}^-$ clamber up and down to your rescue!

Assuming that the normalization factor of the $\downarrow$ state is chosen in terms of the one of the $\uparrow$ state consistent with the ladder relations (12.9) and (12.10),

		possible combined angular momentum j																	
j^{P}	I	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	2	1		1																
	3		1		1															
	4	1		1		1														
	5		1		1		1													
	6	1		1		1		1												
	7		1		1		1		1											
	8	1		1		1		1		1										
	9		1		1		1		1		1									
2	2	1		1		1														
	3	1		1	1	1		1												
	4	1		2		2	1	1		1										
	5	1		2	1	2	1	2	1	1		1								
	6	2		2	1	3	1	3	1	2	1	1		1						
	7	1		3	1	3	2	3	2	3	1	2	1	1		1				
	8	2		3	1	4	2	4	2	4	2	3	1	2	1	1		1		
	9	2		3	2	4	2	5	3	4	3	4	2	3	1	2	1	1		1
3	2	1		1		1		1												
	3		1		2	1	1	1	1		1									
	4	2		2	1	3	1	3	1	2	1	1		1						
	5		2	1	4	2	4	3	4	2	3	2	2	1	1		1			
	6	3		4	3	6	3	7	4	6	4	5	2	4	2	2	1	1		1
4	2	1		1		1		1		1										
	3	1		1	1	2	1	2	1	1	1	1		1						
	4	2		3	1	4	2	4	2	4	2	3	1	2	1	1		1		
5	2	1		1		1		1		1		1								
	3		1		2	1	2	2	2	1	2	1	1	1	1		1			
6	2	1		1		1		1		1		1		1						
	3	1		1	1	2	1	3	2	2	2	2	1	2	1	1	1	1		1
7	2	1		1		1		1		1		1		1		1				
8	2	1		1		1		1		1		1		1		1		1		
9	2	1		1		1		1		1		1		1		1		1		1

Table 12.2: Possible combined angular momentum of identical bosons.

you have:

$$\widehat{S}^+\uparrow = (\widehat{S}_x + \mathrm{i}\widehat{S}_y)\uparrow = 0 \qquad \widehat{S}^-\uparrow = (\widehat{S}_x - \mathrm{i}\widehat{S}_y)\uparrow = \hbar\downarrow$$

By adding or subtracting the two equations, you find the effects of $\widehat{S}_x$ and $\widehat{S}_y$ on the spin-up state:

$$\widehat{S}_x\uparrow = \tfrac{1}{2}\hbar\downarrow \qquad \widehat{S}_y\uparrow = \tfrac{1}{2}\mathrm{i}\hbar\downarrow$$

It works the same way for the spin-down state $\downarrow = |\frac{1}{2}\,{-\frac{1}{2}}\rangle$:

$$\widehat{S}_x\downarrow = \tfrac{1}{2}\hbar\uparrow \qquad \widehat{S}_y\downarrow = -\tfrac{1}{2}\mathrm{i}\hbar\uparrow$$

You now know the effect of the x and y angular momentum operators on the z-direction spin states. Chalk one up for the ladder operators.

Next, assume that you have some spin state that is an arbitrary combination of spin-up and spin-down:

$$a\uparrow + b\downarrow$$

Then, according to the expressions above, application of the x spin operator $\widehat{S}_x$ will turn it into:

$$\widehat{S}_x\,(a\uparrow + b\downarrow) = a\left(0\uparrow + \tfrac{1}{2}\hbar\downarrow\right) + b\left(\tfrac{1}{2}\hbar\uparrow + 0\downarrow\right)$$

while the operator $\widehat{S}_y$ turns it into

$$\widehat{S}_y\,(a\uparrow + b\downarrow) = a\left(0\uparrow + \tfrac{1}{2}\hbar\mathrm{i}\downarrow\right) + b\left(-\tfrac{1}{2}\hbar\mathrm{i}\uparrow + 0\downarrow\right)$$

And of course, since $\uparrow$ and $\downarrow$ are the eigenstates of $\widehat{S}_z$,

$$\widehat{S}_z\,(a\uparrow + b\downarrow) = a\left(\tfrac{1}{2}\hbar\uparrow + 0\downarrow\right) + b\left(0\uparrow - \tfrac{1}{2}\hbar\downarrow\right)$$

If you put the coefficients in the formula above, except for the common factor $\frac{1}{2}\hbar$, in little 2×2 tables, you get the so-called "Pauli spin matrices":

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma_y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{12.15}$$

where the convention is that a multiplies the first column of the matrices and b the second. Also, the top rows in the matrices produce the spin-up part of the result and the bottom rows the spin down part. In linear algebra, you also put the coefficients a and b together in a vector:

$$a\uparrow + b\downarrow \equiv \begin{pmatrix} a \\ b \end{pmatrix}$$

You can now go further and find the eigenstates of the $\widehat{S}_x$ and $\widehat{S}_y$ operators in terms of the eigenstates $\uparrow$ and $\downarrow$ of the $\widehat{S}_z$ operator. You can use the techniques of linear algebra, or you can guess. For example, if you guess $a = b = 1$,

$$\widehat{S}_x \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \tfrac{1}{2}\hbar\sigma_x \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \tfrac{1}{2}\hbar \begin{pmatrix} 0\times 1 + 1 \times 1 \\ 1 \times 1 + 0 \times 1 \end{pmatrix} = \tfrac{1}{2}\hbar \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

so $a = b = 1$ is an eigenstate of $\widehat{S}_x$ with eigenvalue $\frac{1}{2}\hbar$, call it a $\rightarrow$, "spin-right", state. To normalize the state, you still need to divide by $\sqrt{2}$:

$$\rightarrow = \frac{1}{\sqrt{2}}\uparrow + \frac{1}{\sqrt{2}}\downarrow$$

Similarly, you can guess the other eigenstates, and come up with:

$$\rightarrow = \frac{1}{\sqrt{2}}\uparrow + \frac{1}{\sqrt{2}}\downarrow \quad \leftarrow = -\frac{\mathrm{i}}{\sqrt{2}}\uparrow + \frac{\mathrm{i}}{\sqrt{2}}\downarrow \quad \otimes = \frac{1}{\sqrt{2}}\uparrow + \frac{\mathrm{i}}{\sqrt{2}}\downarrow \quad \odot = \frac{1}{\sqrt{2}}\uparrow - \frac{\mathrm{i}}{\sqrt{2}}\downarrow \tag{12.16}$$

Note that the square magnitudes of the coefficients of the states are all one half, giving a 50/50 chance of finding the z-momentum up or down. Since the choice of the axis system is arbitrary, this can be generalized to mean that if the spin in a given direction has an definite value, then there will be a 50/50 chance of the spin in any orthogonal direction turning out to be $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$.

You might wonder about the choice of normalization factors in the spin states (12.16). For example, why not leave out the common factor i in the $\leftarrow$, (negative x spin, or spin-left), state? The reason is to ensure that the x-direction ladder operator $\widehat{S}_y \pm \mathrm{i}\widehat{S}_z$ and the y-direction one $\widehat{S}_z \pm \mathrm{i}\widehat{S}_x$, as obtained by cyclic permutation of the ones for z, produce real, positive multiplication factors. This allows relations valid in the z-direction (like the expressions for triplet and singlet states) to also apply in the x and y directions. In addition, with this choice, if you do a simple change in the labeling of the axes, from xyz to yzx or zxy, the form of the Pauli spin matrices remains unchanged. The $\rightarrow$ and $\otimes$ states of positive x-, respectively y-momentum were chosen a different way: if you rotate the axis system 90° around the y or x axis, these are the spin-up states along the new z-axis, the x-axis or y-axis in the system you are looking at now, {D.68}.

12.11 General spin matrices

The arguments that produced the Pauli spin matrices for a system with spin ½ work equally well for systems with larger square angular momentum.

In particular, from the definition of the ladder operators

$$\widehat{J}^+ \equiv \widehat{J}_x + \mathrm{i}\widehat{J}_y \qquad \widehat{J}^- \equiv \widehat{J}_x - \mathrm{i}\widehat{J}_y$$

it follows by taking the sum, respectively difference, that

$$\widehat{J}_x = \tfrac{1}{2}\widehat{J}^+ + \tfrac{1}{2}\widehat{J}^- \qquad \widehat{J}_y = -\mathrm{i}\tfrac{1}{2}\widehat{J}^+ + \mathrm{i}\tfrac{1}{2}\widehat{J}^- \tag{12.17}$$

Therefore, the effect of either $\widehat{J}_x$ or $\widehat{J}_y$ is to produce multiples of the states with the next higher and the next lower magnetic quantum number. The multiples can be determined using (12.9) and (12.10).

If you put these multiples again in matrices, after ordering the states by magnetic quantum number, you get Hermitian tridiagonal matrices with nonzero sub and superdiagonals and zero main diagonal, where $\widehat{J}_x$ is real symmetric while $\widehat{J}_y$ is purely imaginary, equal to i times a real skew-symmetric matrix. Be sure to tell all you friends that you heard it here first. Do watch out for the well-informed friend who may be aware that forming such matrices is bad news anyway since they are almost all zeros. If you want to use canned matrix software, at least use the kind for tridiagonal matrices.

12.12 The Relativistic Dirac Equation

Relativity threw up some road blocks when quantum mechanics was first formulated, especially for the particles physicist wanted to look at most, electrons. This section explains some of the ideas.

You will need a good understanding of linear algebra to really follow the reasoning. A summary of the Dirac equation that is less heavy on the linear algebra can be found in {A.44}.

For zero spin particles, including relativity appears to be simple. The classical kinetic energy Hamiltonian for a particle in free space,

$$H = \frac{1}{2m}\sum_{i=1}^{3}\widehat{p}_i^{\,2} \qquad \widehat{p}_i = \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial r_i}$$

can be replaced by Einstein's relativistic expression

$$H = \sqrt{(mc^2)^2 + \sum_{i=1}^{3}(\widehat{p}_i c)^2}$$

where m is the rest mass of the particle and mc^2 is the energy this mass is equivalent to. You can again write $H\psi = E\psi$, or squaring the operators in both sides to get rid of the square root:

$$\left[\left(mc^2\right)^2 + \sum_{i=1}^{3}(\widehat{p}_i c)^2\right]\psi = E^2\psi$$

This is the "Klein-Gordon" relativistic version of the Hamiltonian eigenvalue problem. With a bit of knowledge of partial differential equations, you can check that the unsteady version, chapter 7.1, obeys the speed of light as the maximum propagation speed, as you would expect, chapter 8.6.

Unfortunately, throwing a dash of spin into this recipe simply does not seem to work in a convincing way. Apparently, that very problem led Schrödinger to limit himself to the nonrelativistic case. It is hard to formulate simple equations with an ugly square root in your way, and surely, you will agree, the relativistic equation for something so very fundamental as an electron in free space should be simple and beautiful like other fundamental equations in physics. (Can you be more concise than $\vec{F} = m\vec{a}$ or $E = mc^2$?).

So P.A.M. Dirac boldly proposed that for a particle like an electron, (and other spin ½ elementary particles like quarks, it turned out,) the square root produces a simple linear combination of the individual square root terms:

$$\sqrt{(mc^2)^2 + \sum_{i=1}^{3} (\widehat{p}_i c)^2} = \alpha_0 mc^2 + \sum_{i=1}^{3} \alpha_i \widehat{p}_i c \tag{12.18}$$

for suitable coefficients α_0, α_1, α_2 and α_3. Now, if you know a little bit of algebra, you will quickly recognize that there is absolutely no way this can be true. The teacher will have told you that, say, a function like $\sqrt{x^2+y^2}$ is definitely not the same as the function $\sqrt{x^2} + \sqrt{y^2} = x + y$, otherwise the Pythagorean theorem would look a lot different, and adding coefficients as in $\alpha_1 x + \alpha_2 y$ does not do any good at all.

But here is the key: while this does not work for plain numbers, Dirac showed it *is* possible if you are dealing with matrices, tables of numbers. In particular, it works if the coefficients are given by

$$\alpha_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \alpha_1 = \begin{pmatrix} 0 & \sigma_x \\ \sigma_x & 0 \end{pmatrix} \quad \alpha_2 = \begin{pmatrix} 0 & \sigma_y \\ \sigma_y & 0 \end{pmatrix} \quad \alpha_3 = \begin{pmatrix} 0 & \sigma_z \\ \sigma_z & 0 \end{pmatrix}$$

This looks like 2×2 size matrices, but actually they are 4×4 matrices since all elements are 2×2 matrices themselves: the ones stand for 2×2 unit matrices, the zeros for 2×2 zero matrices, and the σ_x, σ_y and σ_z are the so-called 2×2 Pauli spin matrices that also pop up in the theory of spin angular momentum, section 12.10. The square root cannot be eliminated with matrices smaller than 4×4 in actual size. (A derivation is in {D.70}. See also {A.36} for alternate forms of the equation.)

Now if the Hamiltonian is a 4×4 matrix, the wave function at any point must have four components. As you might guess from the appearance of the spin matrices, half of the explanation of the wave function splitting into four is the two spin states of the electron. How about the other half? It turns out that the Dirac equation brings with it states of negative total energy, in particular negative rest mass energy.

That was of course a curious thing. Consider an electron in what otherwise is an empty vacuum. What prevents the electron from spontaneously transitioning to the negative rest mass state, releasing twice its rest mass in energy? Dirac concluded that what is called empty vacuum should in the mathematics of quantum mechanics be taken to be a state in which all negative energy states are already filled with electrons. Clearly, that requires the Pauli exclusion principle to be valid for electrons, otherwise the electron could still transition into such a state. According to this idea, nature really does not have a free choice in whether to apply the exclusion principle to electrons if it wants to create a universe as we know it.

But now consider the vacuum without the electron. What prevents you from adding a big chunk of energy and lifting an electron out of a negative rest-mass state into a positive one? Nothing, really. It will produce a normal electron and a place in the vacuum where an electron is missing, a "hole". And here finally Dirac's boldness appears to have deserted him; he shrank from proposing that this hole would physically show up as the exact antithesis of the electron, its anti-particle, the positively charged positron. Instead Dirac weakly pointed the finger at the proton as a possibility. "Pure cowardice," he called it later. The positron that his theory really predicted was subsequently discovered anyway. (It had already been observed earlier, but was not recognized.)

The reverse of the production of an electron/positron pair is pair annihilation, in which a positron and an electron eliminate each other, creating two gamma-ray photons. There must be two, because viewed from the combined center of mass, the net momentum of the pair is zero, and momentum conservation says it must still be zero after the collision. A single photon would have nonzero momentum, you need two photons coming out in opposite directions. However, pairs can be created from a single photon with enough energy if it happens in the vicinity of, say, a heavy nucleus: a heavy nucleus can absorb the momentum of the photon without picking up much velocity, so without absorbing too much of the photon's energy.

The Dirac equation also gives a very accurate prediction of the magnetic moment of the electron, section 13.4, though the quantum electromagnetic field affects the electron and introduces a correction of about a tenth of a percent. But the importance of the Dirac equation was much more than that: it was the clue to our understanding how quantum mechanics can be reconciled with relativity, where particles are no longer absolute, but can be created out of nothing or destroyed according to the mass-energy relation $E = mc^2$, chapter 1.1.2.

Dirac was a theoretical physicist at Cambridge University, but he moved to Florida in his later life to be closer to his elder daughter, and was a professor of physics at the Florida State University when I got there. So it gives me some pleasure to include the Dirac equation in my text as the corner stone of relativistic quantum mechanics.

Chapter 13

Electromagnetism

This chapter explains how quantum mechanics deals with electromagnetic effects.

Some more advanced topics will be left to introductory addenda. That includes how the solution for the hydrogen atom may be corrected for relativistic effects, {A.39}, using perturbation theory, {A.38}. It also includes the quantization of the electromagnetic field, {A.23}, using quantum field theory, {A.15}.

Electromagnetics is closely tied to more advanced concepts in angular momentum and relativity. These have been discussed in chapters 1 and 12.

13.1 The Electromagnetic Hamiltonian

This section describes very basically how electromagnetism fits into quantum mechanics. However, electromagnetism is fundamentally relativistic; its carrier, the photon, readily emerges or disappears. To describe electromagnetic effects fully requires quantum electrodynamics, and that is far beyond the scope of this text. (However, see addenda {A.15} and {A.23} for some of the ideas.)

In classical electromagnetics, the force on a particle with charge q in a field with electric strength $\vec{\mathcal{E}}$ and magnetic strength $\vec{\mathcal{B}}$ is given by the Lorentz force law

$$\boxed{m\frac{\mathrm{d}\vec{v}}{\mathrm{d}t} = q\left(\vec{\mathcal{E}} + \vec{v}\times\vec{\mathcal{B}}\right)} \tag{13.1}$$

where $\vec{v}$ is the velocity of the particle and for an electron, the charge is $q = -e$.

Unfortunately, quantum mechanics uses neither forces nor velocities. In fact, the earlier analysis of atoms and molecules in this book used the fact that the electric field is described by the corresponding potential energy V, see for example the Hamiltonian of the hydrogen atom. The magnetic field must appear differently in the Hamiltonian; as the Lorentz force law shows, it couples with velocity. You would expect that still the Hamiltonian would be relatively simple, and the simplest idea is then that any potential corresponding to the

magnetic field moves in together with momentum. Since the momentum is a vector quantity, then so must be the magnetic potential. So, your simplest guess would be that the Hamiltonian takes the form

$$\boxed{H = \frac{1}{2m}\left(\widehat{\vec{p}} - q\vec{A}\right)^2 + q\varphi} \tag{13.2}$$

where $\varphi = V/q$ is the "electric potential" and $\vec{A}$ is the "magnetic vector potential." And this simplest guess is in fact right.

The relationship between the vector potential $\vec{A}$ and the magnetic field strength $\vec{\mathcal{B}}$ will now be found from requiring that the classical Lorentz force law is obtained in the classical limit that the quantum uncertainties in position and momentum are small. In that case, expectation values can be used to describe position and velocity, and the field strengths $\vec{\mathcal{E}}$ and $\vec{\mathcal{B}}$ will be constant on the small quantum scales. That means that the derivatives of φ will be constant, (since $\vec{\mathcal{E}}$ is the negative gradient of φ), and presumably the same for the derivatives of $\vec{A}$.

Now according to chapter 7.2, the evolution of the expectation value of position is found as

$$\frac{\mathrm{d}\langle\widehat{\vec{r}}\rangle}{\mathrm{d}t} = \left\langle \frac{\mathrm{i}}{\hbar}[H, \vec{r}] \right\rangle$$

Working out the commutator with the Hamiltonian above, {D.71}, you get,

$$\frac{\mathrm{d}\langle\widehat{\vec{r}}\rangle}{\mathrm{d}t} = \frac{1}{m}\left\langle \widehat{\vec{p}} - q\vec{A} \right\rangle$$

This is unexpected; it shows that $\widehat{\vec{p}}$, i.e. $\hbar\nabla/\mathrm{i}$, is no longer the operator of the normal momentum $m\vec{v}$ when there is a magnetic field; $\widehat{\vec{p}} - q\vec{A}$ gives the normal momentum. The momentum represented by $\widehat{\vec{p}}$ by itself is called "canonical" momentum to distinguish it from normal momentum:

> *The canonical momentum $\hbar\nabla/\mathrm{i}$ only corresponds to normal momentum if there is no magnetic field involved.*

(Actually, it was not that unexpected to physicists, since the same happens in the classical description of electromagnetics using the so-called Lagrangian approach, chapter 1.3.2.)

Next, Newton's second law says that the time derivative of the linear momentum $m\vec{v}$ is the force. Since according to the above, the linear momentum operator is $\widehat{\vec{p}} - q\vec{A}$, then

$$m\frac{\mathrm{d}\langle\vec{v}\rangle}{\mathrm{d}t} = \frac{\mathrm{d}\langle\widehat{\vec{p}} - q\vec{A}\rangle}{\mathrm{d}t} = \left\langle \frac{\mathrm{i}}{\hbar}[H, \widehat{\vec{p}} - q\vec{A}] \right\rangle - q\left\langle \frac{\partial\vec{A}}{\partial t} \right\rangle$$

The objective is now to ensure that the right hand side is the correct Lorentz force (13.1) for the assumed Hamiltonian, by a suitable definition of $\vec{\mathcal{B}}$ in terms of $\vec{A}$.

After a lot of grinding down commutators, {D.71}, it turns out that indeed the Lorentz force is obtained,

$$m\frac{\mathrm{d}\langle\vec{v}\rangle}{\mathrm{d}t} = q\left(\vec{\mathcal{E}} + \langle\vec{v}\rangle \times \vec{\mathcal{B}}\right)$$

provided that:

$$\boxed{\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial\vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla \times \vec{A}} \tag{13.3}$$

So the magnetic field is found as the curl of the vector potential $\vec{A}$. And the electric field is no longer just the negative gradient of the scalar potential φ if the vector potential varies with time.

These results are not new. The electric scalar potential φ and the magnetic vector potential $\vec{A}$ are the same in classical physics, though they are a lot less easy to guess than done here. Moreover, in classical physics they are just convenient mathematical quantities to simplify analysis. In quantum mechanics they appear as central to the formulation.

And it can make a difference. Suppose you do an experiment where you pass electron wave functions around both sides of a very thin magnet: you will get a wave interference pattern behind the magnet. The classical expectation is that this interference pattern will be independent of the magnet strength: the magnetic field $\vec{\mathcal{B}}$ outside a very thin and long ideal magnet is zero, so there is no force on the electron. But the magnetic vector potential $\vec{A}$ is *not* zero outside the magnet, and Aharonov and Bohm argued that the interference pattern would therefore change with magnet strength. So it turned out to be in experiments done subsequently. The conclusion is clear; nature really goes by the vector potential $\vec{A}$ and not the magnetic field $\vec{\mathcal{B}}$ in its actual workings.

13.2 Maxwell's Equations

Maxwell's equations are commonly not covered in a typical engineering program. While these laws are not directly related to quantum mechanics, they do tend to pop up in nanotechnology. This section intends to give you some of the ideas. The description is based on the divergence and curl spatial derivative operators, and the related Gauss and Stokes theorems commonly found in calculus courses (Calculus III in the US system.)

Skipping the first equation for now, the second of Maxwell's equations comes directly out of the quantum mechanical description of the previous section. Consider the expression for the magnetic field $\vec{\mathcal{B}}$ "derived" (guessed) there,

(13.3). If you take its divergence, (premultiply by $\nabla\cdot$), you get rid of the vector potential $\vec{A}$, since the divergence of any curl is always zero, so you get

$$\text{Maxwell's second equation:} \qquad \nabla \cdot \vec{\mathcal{B}} = 0 \tag{13.4}$$

and that is the second of Maxwell's four beautifully concise equations. (The compact modern notation using divergence and curl is really due to Heaviside and Gibbs, though.)

The first of Maxwell's equations is a similar expression for the electric field $\vec{\mathcal{E}}$, but its divergence is *not* zero:

$$\text{Maxwell's first equation:} \qquad \nabla \cdot \vec{\mathcal{E}} = \frac{\rho}{\epsilon_0} \tag{13.5}$$

where ρ is the electric charge per unit volume that is present and the constant $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is called the permittivity of space.

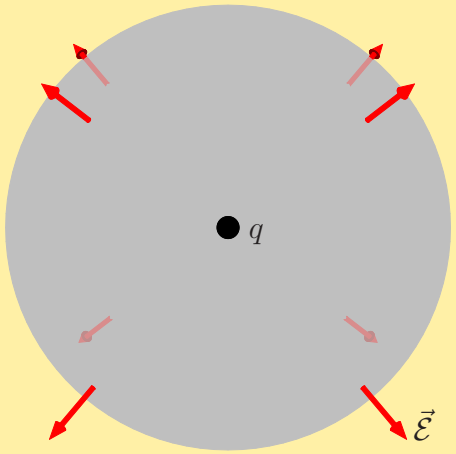


Figure 13.1: Relationship of Maxwell's first equation to Coulomb's law.

What does it all mean? Well, the first thing to verify is that Maxwell's first equation is just a very clever way to write Coulomb's law for the electric field of a point charge. Consider therefore an electric point charge of strength q, and imagine this charge surrounded by a translucent sphere of radius r, as shown in figure 13.1. By symmetry, the electric field at all points on the spherical surface is radial, and everywhere has the same magnitude $\mathcal{E} = |\vec{\mathcal{E}}|$; figure 13.1 shows it for eight selected points.

Now watch what happens if you integrate both sides of Maxwell's first equation (13.5) over the interior of this sphere. Starting with the right hand side, since the charge density is the charge per unit volume, by definition its integral over the volume is the charge q. So the right hand side integrates simply to q/ϵ_0. How about the left hand side? Well, the Gauss, or divergence, theorem of calculus says that the divergence of any vector, $\vec{\mathcal{E}}$ in this case, integrated over

the *volume* of the sphere, equals the radial electric field $\mathcal{E}$ integrated over the *surface* of the sphere. Since $\mathcal{E}$ is constant on the surface, and the surface of a sphere is just $4\pi r^2$, the right hand side integrates to $4\pi r^2\mathcal{E}$. So in total, you get for the integrated first Maxwell's equation that $4\pi r^2\mathcal{E} = q/\epsilon_0$. Take the $4\pi r^2$ to the other side and there you have the Coulomb electric field of a point charge:

$$\text{Coulomb's law:} \qquad \mathcal{E} = \frac{q}{4\pi r^2 \epsilon_0} \tag{13.6}$$

Multiply by $-e$ and you have the electrostatic force on an electron in that field according to the Lorentz equation (13.1). Integrate with respect to r and you have the potential energy $V = -qe/4\pi\epsilon_0 r$ that has been used earlier to analyze atoms and molecules.

Of course, all this raises the question, why bother? If Maxwell's first equation is just a rewrite of Coulomb's law, why not simply stick with Coulomb's law in the first place? Well, to describe the electric field at a given point using Coulomb's law requires you to consider every charge everywhere else. In contrast, Maxwell's equation only involves *local* quantities at the given point, to wit, the derivatives of the local electric field and the local charge per unit volume. It so happens that in numerical or analytical work, most of the time it is much more convenient to deal with local quantities, even if those are derivatives, than with global ones.

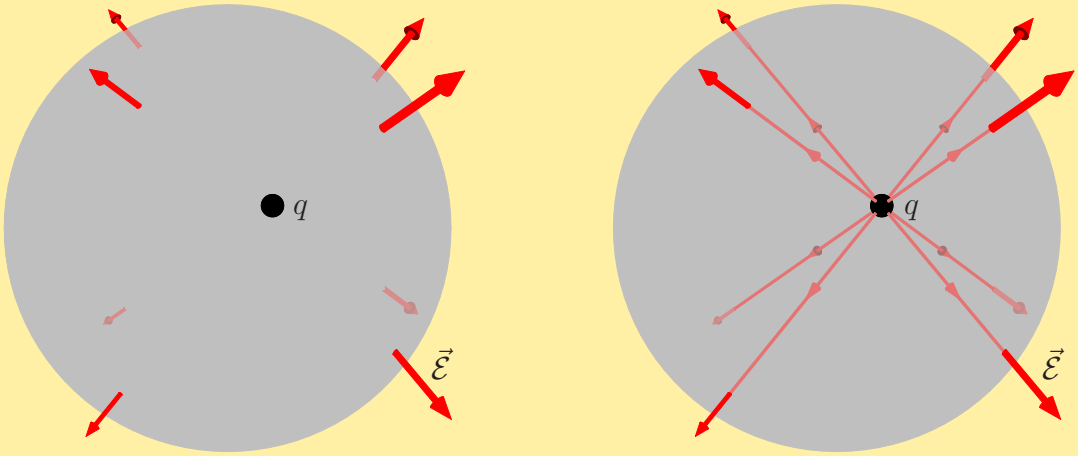


Figure 13.2: Maxwell's first equation for a more arbitrary region. The figure to the right includes the field lines through the selected points.

Of course, you can also integrate Maxwell's first equation over more general regions than a sphere centered around a charge. For example figure 13.2 shows a sphere with an off-center charge. But the electric field strength is no longer constant over the surface, and divergence theorem now requires you to integrate the component of the electric field normal to the surface over the surface. Clearly, that does not have much intuitive meaning. However, if you are willing

to loosen up a bit on mathematical preciseness, there is a better way to look at it. It is in terms of the "electric field lines", the lines that everywhere trace the direction of the electric field. The left figure in figure 13.2 shows the field lines through the selected points; a single charge has radial field lines.

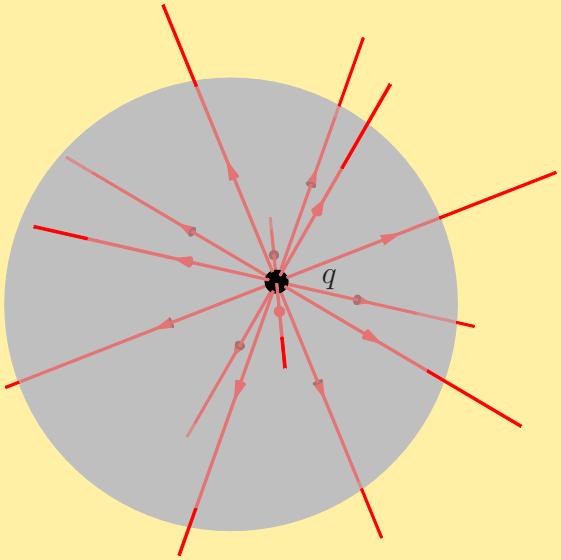


Figure 13.3: The net number of field lines leaving a region is a measure for the net charge inside that region.

Assume that you draw the field lines densely, more like figure 13.3 say, and moreover, that you make the number of field lines coming out of a charge proportional to the strength of that charge. In that case, the local density of field lines at a point becomes a measure of the strength of the electric field at that point, and in those terms, Maxwell's integrated first equation says that the net number of field lines *leaving* a region is proportional to the net charge *inside* that region. That remains true when you add more charges inside the region. In that case the field lines will no longer be straight, but the net number going out will still be a measure of the net charge inside.

Now consider the question why Maxwell's *second* equation says that the divergence of the magnetic field is zero. For the electric field you can shove, say, some electrons in the region to create a net negative charge, or you can shove in some ionized molecules to create a net positive charge. But the magnetic equivalents to such particles, called "magnetic monopoles", being separate magnetic north pole particles or magnetic south pole particles, simply do not exist, {N.31}. It might *appear* that your bar magnet has a north pole and a south pole, but if you take it apart into little pieces, you do not end up with north pole pieces and south pole pieces. Each little piece by itself is still a little magnet, with equally strong north and south poles. The only reason the com-

bined magnet *seems* to have a north pole is that all the microscopic magnets of which it consists have their north poles preferentially pointed in that direction.

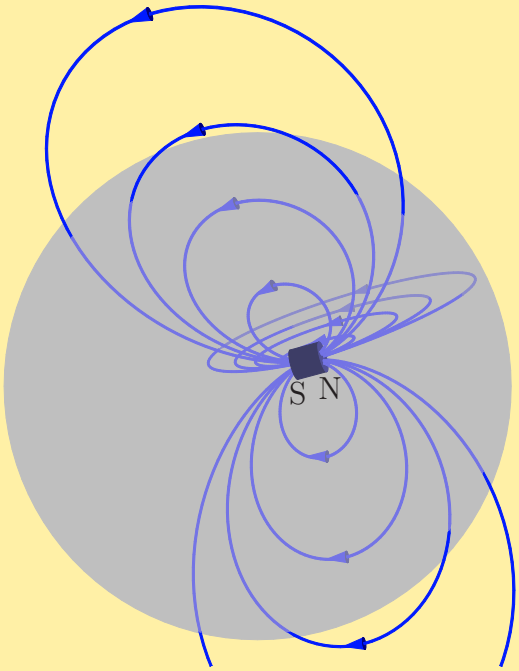


Figure 13.4: Since magnetic monopoles do not exist, the net number of magnetic field lines leaving a region is always zero.

If all microscopic magnets have equal strength north and south poles, then the same number of magnetic field lines that come out of the north poles go back into the south poles, as figure 13.4 illustrates. So the *net* magnetic field lines leaving a given region will be zero; whatever goes out comes back in. True, if you enclose the north pole of a long bar magnet by an imaginary sphere, you can get a pretty good magnetic approximation of the electrical case of figure 13.1. But even then, if you look *inside* the magnet where it sticks through the spherical surface, the field lines will be found to go *in* towards the north pole, instead of away from it. You see why Maxwell's second equation is also called "absence of magnetic monopoles." And why, say, electrons can have a net negative charge, but have zero magnetic pole strength; their spin and orbital angular momenta produce equally strong magnetic north and south poles, a magnetic "dipole" (di meaning two.)

You can get Maxwell's third equation from the electric field "derived" in the previous section. If you take its curl, (premultiply by $\nabla\times$), you get rid of the potential φ, since the curl of any gradient is always zero, and the curl of $\vec{A}$ is the magnetic field. So the third of Maxwell's equations is:

$$\text{Maxwell's third equation:} \qquad \nabla \times \vec{\mathcal{E}} = -\frac{\partial \vec{\mathcal{B}}}{\partial t} \tag{13.7}$$

The "curl", $\nabla\times$, is also often indicated as "rot".

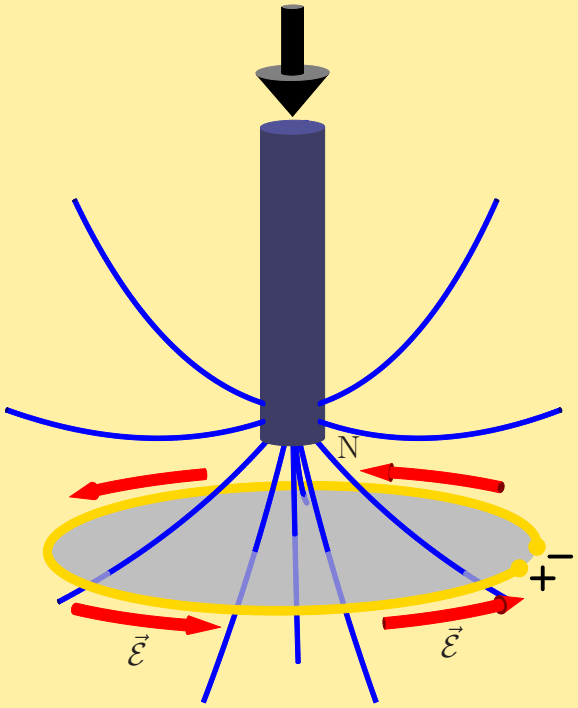


Figure 13.5: Electric power generation.

Now what does that one mean? Well, the first thing to verify in this case is that this is just a clever rewrite of Faraday's law of induction, governing electric power generation. Assume that you want to create a voltage to drive some load (a bulb or whatever, don't worry what the load is, just how to get the voltage for it.) Just take a piece of copper wire and bend it into a circle, as shown in figure 13.5. If you can create a voltage difference between the ends of the wire you are in business; just hook your bulb or whatever to the ends of the wire and it will light up. But to get such a voltage, you will need an electric field as shown in figure 13.5 because the voltage difference between the ends is the integral of the electric field strength along the length of the wire. Now Stokes' theorem of calculus says that the electric field strength along the wire integrated over the *length* of the wire equals the integral of the curl of the electric field strength integrated over the *inside* of the wire, in other words over the imaginary translucent circle in figure 13.5. So to get the voltage, you need a nonzero curl of the electric field on the translucent circle. And Maxwell's third equation above says that this means a time-varying magnetic field on the translucent circle. Moving the end of a strong magnet closer to the circle should do it, as suggested by figure 13.5. You better not make that a big bulb unless you you wrap the wire around a lot more times to form a spool, but anyway. {N.32}.

Maxwell's fourth and final equation is a similar expression for the curl of the magnetic field:

$$\text{Maxwell's fourth equation:} \qquad c^2 \nabla \times \vec{\mathcal{B}} = \frac{\vec{\jmath}}{\epsilon_0} + \frac{\partial \vec{\mathcal{E}}}{\partial t} \tag{13.8}$$

where $\vec{\jmath}$ is the "electric current density," the charge flowing per unit cross sectional area, and c is the speed of light. (It is possible to rescale $\vec{\mathcal{B}}$ by a factor c to get the speed of light to show up equally in the equations for the curl of $\vec{\mathcal{E}}$ and the curl of $\vec{\mathcal{B}}$, but then the Lorentz force law must be adjusted too.)

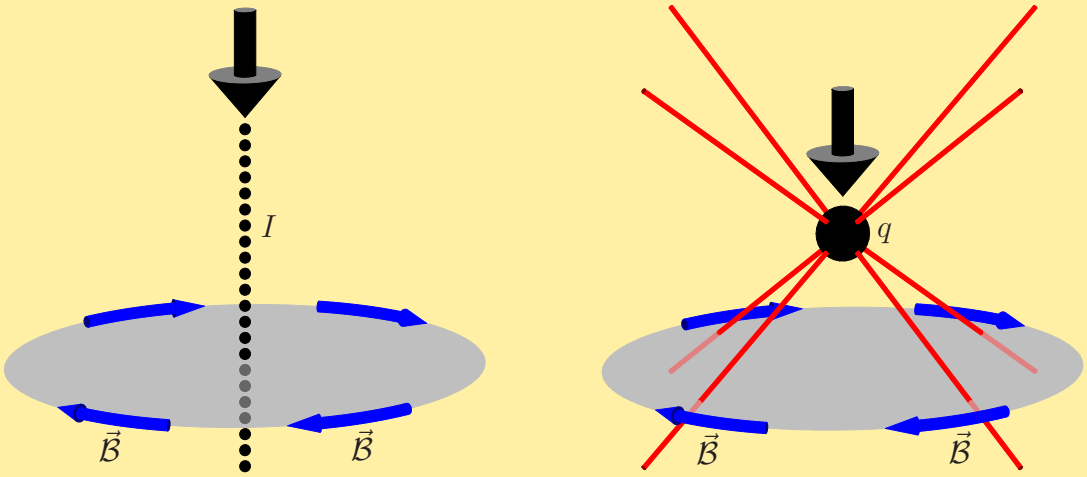


Figure 13.6: Two ways to generate a magnetic field: using a current (left) or using a varying electric field (right).

The big difference from the third equation is the appearance of the current density $\vec{\jmath}$. So, there are two ways to create a circulatory magnetic field, as shown in figure 13.6: (1) pass a current through the enclosed circle (the current density integrates over the area of the circle into the current through the circle), and (2) by creating a varying electric field over the circle, much like was done for the electric field in figure 13.5.

The fact that a current creates a surrounding magnetic field was already known as Ampere's law when Maxwell did his analysis. Maxwell himself however added the time derivative of the electric field to the equation to have the mathematics make sense. The problem was that the divergence of any curl must be zero, and by itself, the divergence of the current density in the right hand side of the fourth equation is *not* zero. Just like the divergence of the electric field is the net field lines coming out of a region per unit volume, the divergence of the current density is the net current coming out. And it is perfectly OK for a net charge to flow out of a region: it simply reduces the charge remaining within the region by that amount. This is expressed by the "continuity equation:"

$$\text{Maxwell's continuity equation:} \qquad \nabla \cdot \vec{\jmath} = -\frac{\partial \rho}{\partial t} \tag{13.9}$$

So Maxwell's fourth equation without the time derivative of the electric field is mathematically impossible. But after he added it, if you take the divergence of the total right hand side then you do indeed get zero as you should. To check that, use the continuity equation above and the first equation.

In empty space, Maxwell's equations simplify: there are no charges so both the charge density ρ and the current density $\vec{\jmath}$ will be zero. In that case, the solutions of Maxwell's equations are simply combinations of "traveling waves." A traveling wave takes the form

$$\vec{\mathcal{E}} = \hat{k}\mathcal{E}_0 \cos\Big(\omega(t - y/c) - \alpha\Big) \quad \vec{\mathcal{B}} = \hat{\imath}\frac{1}{c}\mathcal{E}_0 \cos\Big(\omega(t - y/c) - \alpha\Big) \tag{13.10}$$

where for simplicity, the y-axis of the coordinate system has been aligned with the direction in which the wave travels, and the z-axis with the amplitude $\hat{k}\mathcal{E}_0$ of the electric field of the wave. Such a wave is called "linearly polarized" in the z-direction. The constant ω is the angular frequency of the wave, equal to 2π times its frequency ν in cycles per second, and is related to its wave length λ by $\omega\lambda/c = 2\pi$. The constant α is just a phase angle. For these simple waves, the magnetic and electric field must be normal to each other, as well as to the direction of wave propagation.

You can plug the above wave solution into Maxwell's equations and so verify that it satisfies them all. With more effort and knowledge of Fourier analysis, you can show that they are the most general possible solutions that take this traveling wave form, and that any arbitrary solution is a combination of these waves (if all directions of the propagation direction and of the electric field relative to it, are included.)

The point is that the waves travel with the speed c. When Maxwell wrote down his equations, c was just a constant to him, but when the propagation speed of electromagnetic waves matched the experimentally measured speed of light, it was just too much of a coincidence and he correctly concluded that light must be traveling electromagnetic waves.

It was a great victory of mathematical analysis. Long ago, the Greeks had tried to use mathematics to make guesses about the physical world, and it was an abysmal failure. You do not want to hear about it. Only when the Renaissance started *measuring* how nature really works, the correct laws were discovered for people like Newton and others to put into mathematical form. But here, Maxwell successfully amends Ampere's *measured* law, just because *the mathematics did not make sense.* Moreover, by deriving how fast electromagnetic waves move, he discovers the very fundamental nature of the then mystifying *physical* phenomenon humans call light.

For those with a knowledge of partial differential equations, separate wave equations for the electric and magnetic fields and their potentials are derived in addendum {A.37}.

An electromagnetic field obviously contains energy; that is how the sun transports heat to our planet. The electromagnetic energy within an otherwise empty volume $\mathcal{V}$ can be found as

$$\boxed{E_{\mathcal{V}} = \tfrac{1}{2}\epsilon_0 \int_{\mathcal{V}} \left(\vec{\mathcal{E}}^2 + c^2 \vec{\mathcal{B}}^2 \right) \mathrm{d}^3\vec{r}} \tag{13.11}$$

This is typically derived by comparing the energy from discharging a condenser to the electric field that it initially holds, and from comparing the energy from discharging a coil to the magnetic field it initially holds. That is too much detail for this book.

But at least the result can be made plausible. First note that the time derivative of the energy above can be written as

$$\frac{\mathrm{d}E_{\mathcal{V}}}{\mathrm{d}t} = -\int_S \epsilon_0 c(\vec{\mathcal{E}} \times c\vec{\mathcal{B}}) \cdot \vec{n}\, \mathrm{d}S$$

Here S is the surface of volume $\mathcal{V}$, and $\vec{n}$ is the unit vector normal to the surface element $\mathrm{d}S$. To verify this expression, bring the time derivative inside the integral in (13.11), then get rid of the time derivatives using Maxwell's third and fourth laws, use the standard vector identity [41, 20.40], and finally the divergence theorem.

Now suppose you have a finite amount of radiation in otherwise empty space. If the amount of radiation is finite, the field should disappear at infinity. So, taking the volume to be all of space, the integral in the right hand side above will be zero. So $E_{\mathcal{V}}$ will be constant. That indicates that $E_{\mathcal{V}}$ should be at least a multiple of the energy. After all, what other scalar quantity than energy would be constant? And the factor ϵ_0 is needed because of units. That misses only the factor $\frac{1}{2}$ in the expression for the energy.

For an arbitrary volume $\mathcal{V}$, the surface integral must then be the energy outflow through the surface of the volume. That suggests that the energy flow rate per unit area is given by the so-called "Poynting vector"

$$\epsilon_0 c\, \vec{\mathcal{E}} \times c\vec{\mathcal{B}} \tag{13.12}$$

Unfortunately, this argument is flawed. You cannot deduce local values of the energy flow from its integral over an entire closed surface. In particular, you can find different vectors that describe the energy flow also without inconsistency. Just add an arbitrary solenoidal vector, a vector whose divergence is zero, to the Poynting vector. For example, adding a multiple of the magnetic field would do it. However, if you look at simple lightwaves like (13.10), the Poynting vector seems the intuitive choice. This paragraph was included because other books have Poynting vectors and you would be very disappointed if yours did not.

You will usually not find Maxwell's equations in the exact form described here. To explain what is going on inside materials, you would have to account

for the electric and magnetic fields of every electron and proton (and neutron!) of the material. That is just an impossible task, so physicists have developed ways to average away all those effects by messing with Maxwell's equations. But then the messed-up $\vec{\mathcal{E}}$ in one of Maxwell's equations is no longer the same as the messed-up $\vec{\mathcal{E}}$ in another, and the same for $\vec{\mathcal{B}}$. So physicists rename one messed-up $\vec{\mathcal{E}}$ as, maybe, the "electric flux density" $\vec{D}$, and a messed up magnetic field as, maybe, "the auxiliary field". And they define many other symbols, and even refer to the auxiliary field as being the magnetic field, all to keep engineers out of nanotechnology. Don't let them! When you need to understand the messed-up Maxwell's equations, Wikipedia has a list of the countless definitions.

13.3 Example Static Electromagnetic Fields

In this section, some basic solutions of Maxwell's equations are described. They will be of interest in addendum {A.39} for understanding relativistic effects on the hydrogen atom (though certainly not essential). They are also of considerable practical importance for a lot of nonquantum applications.

It is assumed throughout this subsection that the electric and magnetic fields do not change with time. All solutions also assume that the ambient medium is vacuum.

For easy reference, Maxwell's equations and various results to be obtained in this section are collected together in tables 13.1 and 13.2. While the existence of magnetic monopoles is unverified, it is often convenient to compute as if they do exist. It allows you to apply ideas from the electric field to the magnetic field and vice-versa. So, the tables include magnetic monopoles with strength q_m, in addition to electric charges with strength q, and a magnetic current density $\vec{\jmath}_m$ in addition to an electric current density $\vec{\jmath}$. The table uses the permittivity of space ϵ_0 and the speed of light c as basic physical constants; the permeability of space $\mu_0 = 1/\epsilon_0 c^2$ is just an annoyance in quantum mechanics and is avoided. The table has been written in terms of $c\vec{\mathcal{B}}$ and $\vec{\jmath}_m/c$ because in terms of those combinations Maxwell's equations have a very pleasing symmetry. It allows you to easily convert between expressions for the electric and magnetic fields. You wish that physicists would have defined the magnetic field as $c\vec{\mathcal{B}}$ instead of $\vec{\mathcal{B}}$ in SI units, but no such luck.

13.3.1 Point charge at the origin

A point charge is a charge concentrated at a single point. It is a very good model for the electric field of the nucleus of an atom, since the nucleus is so small compared to the atom. A point charge of strength q located at the origin has a charge density

$$\text{point charge at the origin:} \quad \rho(\vec{r}) = q\delta^3(\vec{r}) \tag{13.13}$$

Physical constants:

$\epsilon_0 = 8.854\,187\,817\ldots\ 10^{-12}\ \mathrm{C^2/Nm^2}$ $\qquad c = 299\,792\,458\ \mathrm{m/s} \approx 3\ 10^8\ \mathrm{m/s}$

Lorentz force law:

$$\vec{F} = q\left(\vec{\mathcal{E}} + \frac{\vec{v}}{c} \times c\vec{\mathcal{B}}\right) + \frac{q_m}{c}\left(c\vec{\mathcal{B}} - \frac{\vec{v}}{c} \times \vec{\mathcal{E}}\right)$$

Maxwell's equations:

$$\nabla \cdot \vec{\mathcal{E}} = \frac{1}{\epsilon_0}\rho \qquad \nabla \cdot c\vec{\mathcal{B}} = \frac{1}{\epsilon_0}\frac{\rho_m}{c}$$

$$\nabla \times \vec{\mathcal{E}} = -\frac{1}{c}\frac{\partial c\vec{\mathcal{B}}}{\partial t} - \frac{1}{\epsilon_0 c}\frac{\vec{\jmath}_m}{c} \qquad \nabla \times c\vec{\mathcal{B}} = \frac{1}{c}\frac{\partial \vec{\mathcal{E}}}{\partial t} + \frac{1}{\epsilon_0 c}\vec{\jmath}$$

$$\nabla \cdot \vec{\jmath} + \frac{\partial \vec{\rho}}{\partial t} = 0 \qquad \nabla \cdot \vec{\jmath}_m + \frac{\partial \vec{\rho}_m}{\partial t} = 0$$

Existence of a potential:

$\vec{\mathcal{E}} = -\nabla\varphi$ iff $\nabla \times \vec{\mathcal{E}} = 0$ $\qquad \vec{\mathcal{B}} = -\nabla\varphi_m$ iff $\nabla \times \vec{\mathcal{B}} = 0$

Point charge at the origin:

$$\varphi = \frac{q}{4\pi\epsilon_0}\frac{1}{r} \qquad \vec{\mathcal{E}} = \frac{q}{4\pi\epsilon_0}\frac{\vec{r}}{r^3} \qquad c\varphi_m = \frac{q_m}{4\pi\epsilon_0 c}\frac{1}{r} \qquad c\vec{\mathcal{B}} = \frac{q_m}{4\pi\epsilon_0 c}\frac{\vec{r}}{r^3}$$

Point charge at the origin in 2D:

$$\varphi = \frac{q'}{2\pi\epsilon_0}\ln\frac{1}{r} \qquad \vec{\mathcal{E}} = \frac{q'}{2\pi\epsilon_0}\frac{\vec{r}}{r^2} \qquad c\varphi_m = \frac{q'_m}{2\pi\epsilon_0 c}\ln\frac{1}{r} \qquad c\vec{\mathcal{B}} = \frac{q'_m}{2\pi\epsilon_0 c}\frac{\vec{r}}{r^2}$$

Charge dipoles:

$$\varphi = \frac{q}{4\pi\epsilon_0}\left[\frac{1}{|\vec{r}-\vec{r}_\oplus|} - \frac{1}{|\vec{r}-\vec{r}_\ominus|}\right] \qquad c\varphi_m = \frac{q_m}{4\pi\epsilon_0 c}\left[\frac{1}{|\vec{r}-\vec{r}_\oplus|} - \frac{1}{|\vec{r}-\vec{r}_\ominus|}\right]$$

$$\vec{\mathcal{E}} = \frac{q}{4\pi\epsilon_0}\left[\frac{\vec{r}-\vec{r}_\oplus}{|\vec{r}-\vec{r}_\oplus|^3} - \frac{\vec{r}-\vec{r}_\ominus}{|\vec{r}-\vec{r}_\ominus|^3}\right] \qquad c\vec{\mathcal{B}} = \frac{q_m}{4\pi\epsilon_0 c}\left[\frac{\vec{r}-\vec{r}_\oplus}{|\vec{r}-\vec{r}_\oplus|^3} - \frac{\vec{r}-\vec{r}_\ominus}{|\vec{r}-\vec{r}_\ominus|^3}\right]$$

$\vec{\wp} = q\,(\vec{r}_\oplus - \vec{r}_\ominus)$ $\quad E_{\mathrm{ext}} = -\vec{\wp}\cdot\vec{\mathcal{E}}_{\mathrm{ext}}$ $\qquad \vec{\mu} = q_m\,(\vec{r}_\oplus - \vec{r}_\ominus)$ $\quad E_{\mathrm{ext}} = -\vec{\mu}\cdot\vec{\mathcal{B}}_{\mathrm{ext}}$

Charge dipoles in 2D:

$$\varphi = \frac{q'}{2\pi\epsilon_0}\left[\ln\frac{1}{|\vec{r}-\vec{r}_\oplus|} - \ln\frac{1}{|\vec{r}-\vec{r}_\ominus|}\right] \qquad c\varphi_m = \frac{q'_m}{2\pi\epsilon_0 c}\left[\ln\frac{1}{|\vec{r}-\vec{r}_\oplus|} - \ln\frac{1}{|\vec{r}-\vec{r}_\ominus|}\right]$$

$$\vec{\mathcal{E}} = \frac{q'}{2\pi\epsilon_0}\left[\frac{\vec{r}-\vec{r}_\oplus}{|\vec{r}-\vec{r}_\oplus|^2} - \frac{\vec{r}-\vec{r}_\ominus}{|\vec{r}-\vec{r}_\ominus|^2}\right] \qquad c\vec{\mathcal{B}} = \frac{q'_m}{2\pi\epsilon_0 c}\left[\frac{\vec{r}-\vec{r}_\oplus}{|\vec{r}-\vec{r}_\oplus|^2} - \frac{\vec{r}-\vec{r}_\ominus}{|\vec{r}-\vec{r}_\ominus|^2}\right]$$

$\vec{\wp}' = q'\,(\vec{r}_\oplus - \vec{r}_\ominus)$ $\quad E'_{\mathrm{ext}} = -\vec{\wp}'\cdot\vec{\mathcal{E}}_{\mathrm{ext}}$ $\qquad \vec{\mu}' = q'_m\,(\vec{r}_\oplus - \vec{r}_\ominus)$ $\quad E'_{\mathrm{ext}} = -\vec{\mu}'\cdot\vec{\mathcal{B}}_{\mathrm{ext}}$

Table 13.1: Electromagnetics I: Fundamental equations and basic solutions.

Distributed charges:

$$\varphi = \frac{1}{4\pi\epsilon_0}\int_{\text{all }\underline{\vec{r}}} \frac{1}{|\vec{r}-\underline{\vec{r}}|}\rho(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \qquad c\varphi_m = \frac{1}{4\pi\epsilon_0}\int_{\text{all }\underline{\vec{r}}} \frac{1}{|\vec{r}-\underline{\vec{r}}|}\frac{\rho_m(\underline{\vec{r}})}{c}\,\mathrm{d}^3\underline{\vec{r}}$$

$$\vec{\mathcal{E}} = \frac{1}{4\pi\epsilon_0}\int_{\text{all }\underline{\vec{r}}} \frac{\vec{r}-\underline{\vec{r}}}{|\vec{r}-\underline{\vec{r}}|^3}\rho(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \qquad c\vec{\mathcal{B}} = \frac{1}{4\pi\epsilon_0}\int_{\text{all }\underline{\vec{r}}} \frac{\vec{r}-\underline{\vec{r}}}{|\vec{r}-\underline{\vec{r}}|^3}\frac{\rho_m(\underline{\vec{r}})}{c}\,\mathrm{d}^3\underline{\vec{r}}$$

$$\varphi \sim \frac{q}{4\pi\epsilon_0}\frac{1}{r} + \frac{1}{4\pi\epsilon_0}\frac{\vec{\wp}\cdot\vec{r}}{r^3} \qquad c\varphi_m \sim \frac{q_m}{4\pi\epsilon_0 c}\frac{1}{r} + \frac{1}{4\pi\epsilon_0}\frac{\vec{\mu}\cdot\vec{r}}{cr^3}$$

$$\vec{\mathcal{E}} \sim \frac{q}{4\pi\epsilon_0}\frac{\vec{r}}{r^3} + \frac{1}{4\pi\epsilon_0}\frac{3(\vec{\wp}\cdot\vec{r})\vec{r}-\vec{\wp}r^2}{r^5} \qquad c\vec{\mathcal{B}} \sim \frac{q_m}{4\pi\epsilon_0 c}\frac{\vec{r}}{r^3} + \frac{1}{4\pi\epsilon_0}\frac{3(\vec{\mu}\cdot\vec{r})\vec{r}-\vec{\mu}r^2}{cr^5}$$

$$q = \int \rho(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \quad \vec{\wp} = \int \underline{\vec{r}}\rho(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \qquad q_m = \int \rho_m(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \quad \vec{\mu} = \int \underline{\vec{r}}\rho_m(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}}$$

Ideal charge dipoles:

$$\varphi = \frac{1}{4\pi\epsilon_0}\frac{\vec{\wp}\cdot\vec{r}}{r^3} \qquad c\varphi_m = \frac{1}{4\pi\epsilon_0}\frac{\vec{\mu}\cdot\vec{r}}{cr^3}$$

$$\vec{\mathcal{E}} = \frac{1}{4\pi\epsilon_0}\frac{3(\vec{\wp}\cdot\vec{r})\vec{r}-\vec{\wp}r^2}{r^5} - \frac{\vec{\wp}}{3\epsilon_0}\delta^3(\vec{r}) \qquad c\vec{\mathcal{B}} = \frac{1}{4\pi\epsilon_0}\frac{3(\vec{\mu}\cdot\vec{r})\vec{r}-\vec{\mu}r^2}{cr^5} - \frac{\vec{\mu}}{3\epsilon_0 c}\delta^3(\vec{r})$$

Biot-Savart law for current densities and currents:

$$\vec{\mathcal{E}} = \frac{1}{4\pi\epsilon_0 c}\int_{\text{all }\underline{\vec{r}}} \frac{\vec{r}-\underline{\vec{r}}}{|\vec{r}-\underline{\vec{r}}|^3}\times\frac{\vec{\jmath}_m(\underline{\vec{r}})}{c}\,\mathrm{d}^3\underline{\vec{r}} \qquad c\vec{\mathcal{B}} = -\frac{1}{4\pi\epsilon_0 c}\int_{\text{all }\underline{\vec{r}}} \frac{\vec{r}-\underline{\vec{r}}}{|\vec{r}-\underline{\vec{r}}|^3}\times\vec{\jmath}(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}}$$

$$\vec{\mathcal{E}} = \frac{1}{4\pi\epsilon_0 c}\int_{\text{all }\underline{\vec{r}}} \frac{\vec{r}-\underline{\vec{r}}}{|\vec{r}-\underline{\vec{r}}|^3}\times\frac{I_m(\underline{\vec{r}})}{c}\,\mathrm{d}\underline{\vec{r}} \qquad c\vec{\mathcal{B}} = -\frac{1}{4\pi\epsilon_0 c}\int_{\text{all }\underline{\vec{r}}} \frac{\vec{r}-\underline{\vec{r}}}{|\vec{r}-\underline{\vec{r}}|^3}\times I(\underline{\vec{r}})\,\mathrm{d}\underline{\vec{r}}$$

2D field due to a straight current along the z-axis:

$$\varphi = \frac{I_m}{2\pi\epsilon_0 c^2}\theta \qquad \vec{\mathcal{E}} = -\frac{I_m}{2\pi\epsilon_0 c^2}\frac{1}{r}\hat{\imath}_\theta \qquad c\varphi_m = -\frac{I}{2\pi\epsilon_0 c}\theta \qquad c\vec{\mathcal{B}} = \frac{I}{2\pi\epsilon_0 c}\frac{1}{r}\hat{\imath}_\theta$$

Current dipole moment:

$$\vec{\wp} = -\frac{1}{2c}\int_{\text{all }\underline{\vec{r}}} \underline{\vec{r}}\times\frac{\vec{\jmath}_m(\underline{\vec{r}})}{c}\,\mathrm{d}^3\underline{\vec{r}} \qquad \vec{\mu} = \frac{1}{2}\int_{\text{all }\underline{\vec{r}}} \underline{\vec{r}}\times\vec{\jmath}(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} = \frac{q_\mathrm{c}}{2m_\mathrm{c}}\vec{L}$$

$$\vec{M} = \vec{\wp}\times\vec{\mathcal{E}}_\text{ext} \qquad E_\text{ext} = -\vec{\wp}\cdot\vec{\mathcal{E}}_\text{ext} \qquad \vec{M} = \vec{\mu}\times\vec{\mathcal{B}}_\text{ext} \qquad E_\text{ext} = -\vec{\mu}\cdot\vec{\mathcal{B}}_\text{ext}$$

Ideal current dipoles:

$$\varphi = \frac{1}{4\pi\epsilon_0}\frac{\vec{\wp}\cdot\vec{r}}{r^3} \qquad c\varphi_m = \frac{1}{4\pi\epsilon_0}\frac{\vec{\mu}\cdot\vec{r}}{cr^3}$$

$$\vec{\mathcal{E}} = \frac{1}{4\pi\epsilon_0}\frac{3(\vec{\wp}\cdot\vec{r})\vec{r}-\vec{\wp}r^2}{r^5} + \frac{2\vec{\wp}}{3\epsilon_0}\delta^3(\vec{r}) \qquad c\vec{\mathcal{B}} = \frac{1}{4\pi\epsilon_0}\frac{3(\vec{\mu}\cdot\vec{r})\vec{r}-\vec{\mu}r^2}{cr^5} + \frac{2\vec{\mu}}{3\epsilon_0 c}\delta^3(\vec{r})$$

Table 13.2: Electromagnetics II: Electromagnetostatic solutions.

where $\delta^3(\vec{r})$ is the three-dimensional delta function. A delta function is a spike at a single point that integrates to one, so the charge density above integrates to the total charge q.

The electric field lines of a point charge are radially outward from the charge; see for example figure 13.3 in the previous subsection. According to Coulomb's law, the electric field of a point charge is

$$\boxed{\text{electric field of a point charge:} \quad \vec{\mathcal{E}} = \frac{q}{4\pi\epsilon_0 r^2}\hat{\imath}_r} \tag{13.14}$$

where r is the distance from the charge, $\hat{\imath}_r$ is the unit vector pointing straight away from the charge, and $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is the permittivity of space. Now for static electric charges the electric field is minus the gradient of a potential φ,

$$\vec{\mathcal{E}} = -\nabla\varphi \qquad \nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

In everyday terms the potential φ is called the "voltage." It follows by integration of the electric field strength with respect to r that the potential of a point charge is

$$\boxed{\text{electric potential of a point charge:} \quad \varphi = \frac{q}{4\pi\epsilon_0 r}} \tag{13.15}$$

Multiply by $-e$ and you get the potential energy V of an electron in the field of the point charge. That was used in writing the Hamiltonians of the hydrogen and heavier atoms.

Delta functions are often not that easy to work with analytically, since they are infinite and infinity is a tricky mathematical thing. It is often easier to do the mathematics by assuming that the charge is spread out over a small sphere of radius ε, rather than concentrated at a single point. If it is assumed that the charge distribution is uniform within the radius ε, then it is

$$\text{spherical charge around the origin:} \quad \rho = \begin{cases} \dfrac{q}{\frac{4}{3}\pi\varepsilon^3} & \text{if } r \leqslant \varepsilon \\ 0 & \text{if } r > \varepsilon \end{cases} \tag{13.16}$$

Since the charge density is the charge per unit volume, the charge density times the volume $\frac{4}{3}\pi\varepsilon^3$ of the little sphere that holds it must be the total charge q. The expression above makes it so.

Figure 13.7 shows that outside the region with charge, the electric field and potential are exactly like those of a point charge with the same net charge q. But inside the region of charge distribution, the electric field varies linearly with radius, and becomes zero at the center. It is just like the gravity of earth: going above the surface of the earth out into space, gravity decreases like $1/r^2$ if r is the distance from the center of the earth. But if you go down below the

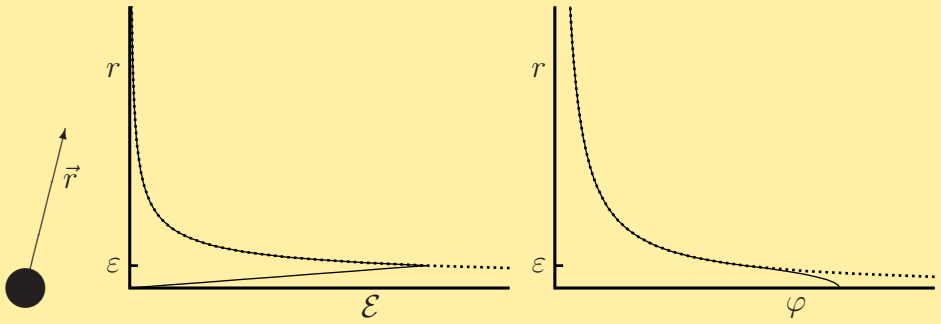


Figure 13.7: Electric field and potential of a charge that is distributed uniformly within a small sphere. The dotted lines indicate the values for a point charge.

surface of the earth, gravity decreases also and becomes zero at the center of the earth. If you want, you can derive the electric field of the spherical charge from Maxwell's first equation; it goes much in the same way that Coulomb's law was derived from it in the previous section.

If magnetic monopoles exist, they would create a magnetic field much like an electric charge creates an electric field. As table 13.1 shows, the only difference is the square of the speed of light c popping up in the expressions. (And that is really just a matter of definitions, anyway.) In real life, these expressions give an approximation for the magnetic field near the north or south pole of a very long thin magnet as long as you do not look inside the magnet.

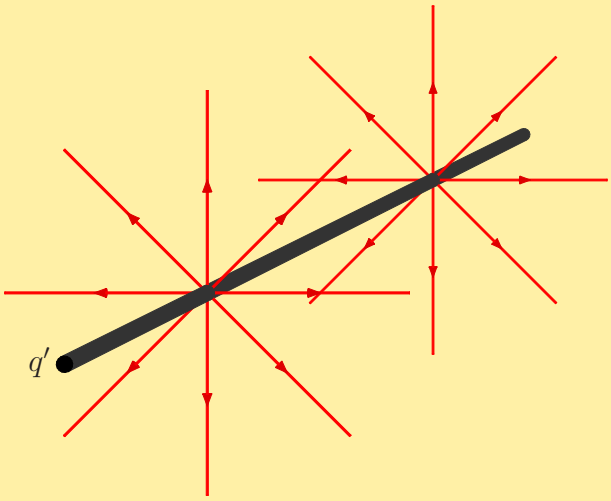


Figure 13.8: Electric field of a two-dimensional line charge.

A homogeneous distribution of charges along an infinite straight line is called a line charge. As shown in figure 13.8, it creates a two-dimensional field in the planes normal to the line. The line charge becomes a point charge within such a plane. The expression for the field of a line charge can be derived in much the

same way as Coulomb's law was derived for a three-dimensional point charge in the previous section. In particular, where that derivation surrounded the point charge by a spherical surface, surround the line charge by a cylinder. (Or by a circle, if you want to think of it in two dimensions.) The resulting expressions are given in table 13.1; they are in terms of the charge per unit length of the line q'. Note that in this section a prime is used to indicate that a quantity is per unit length.

13.3.2 Dipoles

A point charge can describe a single charged particle like an atom nucleus or electron. But much of the time in physics, you are dealing with neutral atoms or molecules. For those, the net charge is zero. The simplest model for a system with zero net charge is called the "dipole." It is simply a combination of a positive point charge q and a negative one $-q$, making the net charge zero.

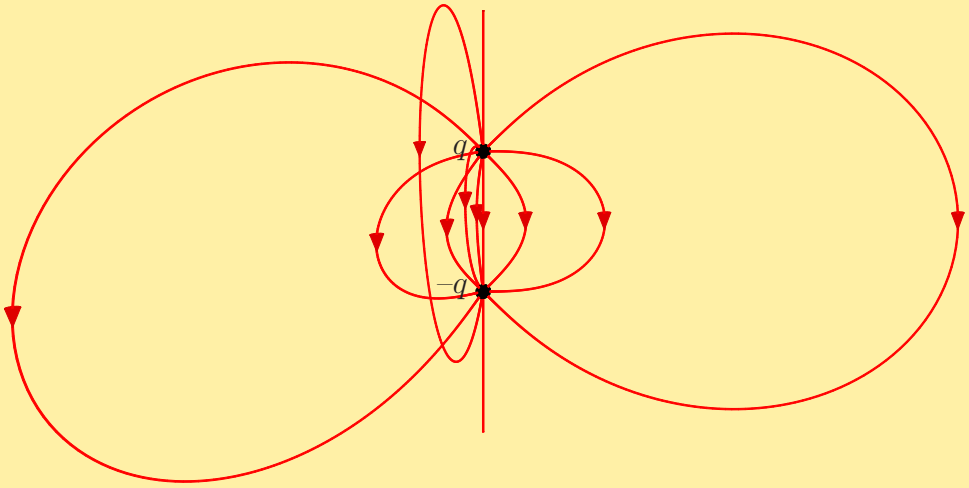


Figure 13.9: Field lines of a vertical electric dipole.

Figure 13.9 shows an example of a dipole in which the positive charge is straight above the negative one. Note the distinctive egg shape of the biggest electric field lines. The "electric dipole moment" $\vec{\wp}$ is defined as the product of the charge strength q times the connecting vector from negative to positive charge:

$$\text{electric dipole moment:} \quad \vec{\wp} = q(\vec{r}_\oplus - \vec{r}_\ominus) \tag{13.17}$$

where $\vec{r}_\oplus$ and $\vec{r}_\ominus$ are the positions of the positive and negative charges respectively.

The potential of a dipole is simply the sum of the potentials of the two charges:

$$\text{potential of an electric dipole:} \quad \varphi = \frac{q}{4\pi\epsilon_0}\frac{1}{|\vec{r} - \vec{r}_\oplus|} - \frac{q}{4\pi\epsilon_0}\frac{1}{|\vec{r} - \vec{r}_\ominus|} \tag{13.18}$$

Note that to convert the expressions for a charge at the origin to one not at the origin, you need to use the position vector measured from the location of the charge.

The electric field of the dipole can be found from either taking minus the gradient of the potential above, or from adding the fields of the individual point charges, and is

$$\text{field of an electric dipole:} \quad \vec{\mathcal{E}} = \frac{q}{4\pi\epsilon_0}\frac{\vec{r}-\vec{r}_\oplus}{|\vec{r}-\vec{r}_\oplus|^3} - \frac{q}{4\pi\epsilon_0}\frac{\vec{r}-\vec{r}_\ominus}{|\vec{r}-\vec{r}_\ominus|^3} \tag{13.19}$$

To obtain that result from taking the the gradient of the potential, remember the following important formula for the gradient of $|\vec{r}-\vec{r}_0|^n$ with n an arbitrary power:

$$\boxed{\frac{\partial|\vec{r}-\vec{r}_0|^n}{\partial r_i} = n|\vec{r}-\vec{r}_0|^{n-2}(r_i - r_{0,i}) \qquad \nabla_{\vec{r}}|\vec{r}-\vec{r}_0|^n = n|\vec{r}-\vec{r}_0|^{n-2}(\vec{r}-\vec{r}_0)} \tag{13.20}$$

The first expression gives the gradient in index notation and the second gives it in vector form. The subscript on ∇ merely indicates that the differentiation is with respect to $\vec{r}$, not $\vec{r}_0$. These formulae will be used routinely in this section. Using them, you can check that minus the gradient of the dipole potential does indeed give its electric field above.

Similar expressions apply for magnetic dipoles. The field outside a thin bar magnet can be approximated as a magnetic dipole, with the north and south poles of the magnet as the positive and negative magnetic point charges. The magnetic field lines are then just like the electric field lines in figure 13.9.

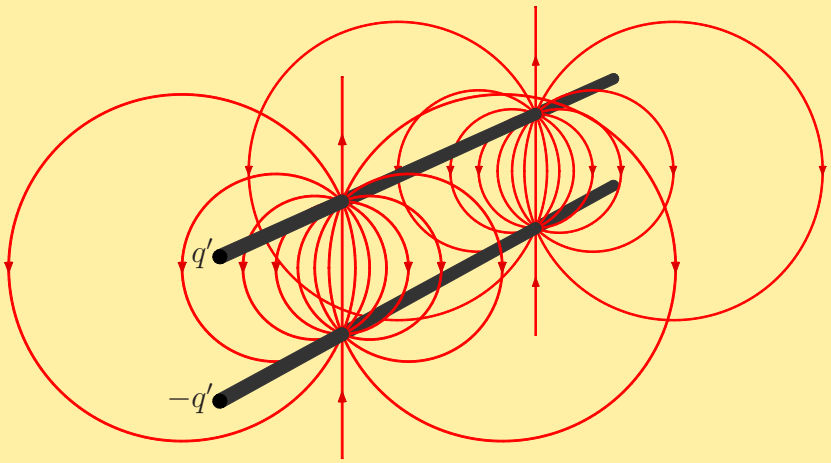


Figure 13.10: Electric field of a two-dimensional dipole.

Corresponding expressions can also be written down in two dimensions, for opposite charges distributed along parallel straight lines. Figure 13.10 gives

an example. In two dimensions, all field lines are circles passing through both charges.

A particle like an electron has an electric charge and no known size. It can therefore be described as an ideal point charge. But an electron also has a magnetic moment: it acts as a magnet of zero size. Such a magnet of zero size will be referred to as an "*ideal* magnetic dipole." More precisely, an ideal magnetic dipole is defined as the limit of a magnetic dipole when the two poles are brought vanishingly close together. Now if you just let the two poles approach each other without doing anything else, their opposite fields will begin to increasingly cancel each other, and there will be no field left when the poles are on top of each other. When you make the distance between the poles smaller, you also need to increase the strengths q_m of the poles to ensure that the

$$\text{magnetic dipole moment:} \quad \vec{\mu} = q_m(\vec{r}_\oplus - \vec{r}_\ominus) \tag{13.21}$$

remains finite. So you can think of an ideal magnetic dipole as infinitely strong magnetic poles infinitely close together.

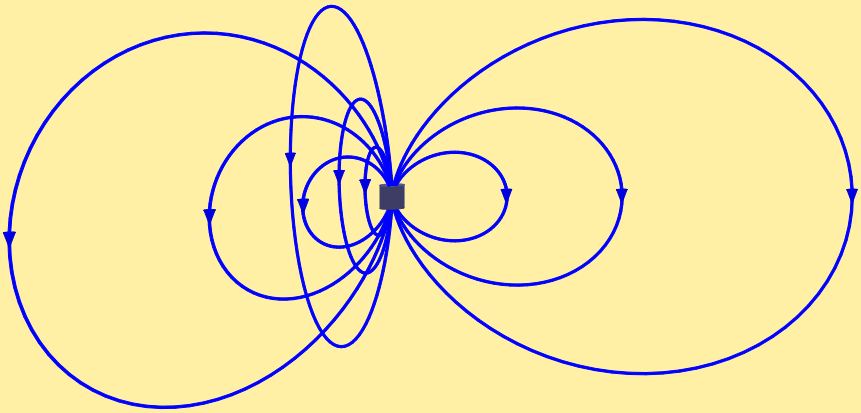


Figure 13.11: Field of an ideal magnetic dipole.

The field lines of a vertical ideal magnetic dipole are shown in figure 13.11. Their egg shape is in spherical coordinates described by, {D.72},

$$r = r_{\max} \sin^2\theta \qquad \phi = \text{constant} \tag{13.22}$$

To find the magnetic field itself, start with the magnetic potential of a nonideal dipole,

$$\varphi_m = \frac{q_m}{4\pi\epsilon_0 c^2}\left[\frac{1}{|\vec{r} - \vec{r}_\oplus|} - \frac{1}{|\vec{r} - \vec{r}_\ominus|}\right]$$

Now take the negative pole at the origin, and allow the positive pole to approach it vanishingly close. Then the potential above takes the generic form

$$\varphi_m = f(\vec{r} - \vec{r}_\oplus) - f(\vec{r}) \qquad f(\vec{r}) = \frac{q_m}{4\pi\epsilon_0 c^2}\frac{1}{|\vec{r}|}$$

Now according to the total differential of calculus, (or the multi-dimensional Taylor series theorem, or the definition of directional derivative), for small $\vec{r}_\oplus$ an expression of the form $f(\vec{r}-\vec{r}_\oplus)-f(\vec{r})$ can be approximated as

$$f(\vec{r}-\vec{r}_\oplus)-f(\vec{r}) \sim -\vec{r}_\oplus \cdot \nabla f \quad \text{for} \quad \vec{r}_\oplus \to 0$$

From this the magnetic potential of an ideal dipole at the origin can be found by using the expression (13.20) for the gradient of $1/|\vec{r}|$ and then substituting the magnetic dipole strength $\vec{\mu}$ for $q_m \vec{r}_\oplus$. The result is

$$\text{potential of an ideal magnetic dipole:} \quad \varphi_m = \frac{1}{4\pi\epsilon_0 c^2}\frac{\vec{\mu}\cdot\vec{r}}{r^3} \tag{13.23}$$

The corresponding magnetic field can be found as minus the gradient of the potential, using again (13.20) and the fact that the gradient of $\vec{\mu}\cdot\vec{r}$ is just $\vec{\mu}$:

$$\vec{\mathcal{B}} = \frac{1}{4\pi\epsilon_0 c^2}\frac{3(\vec{\mu}\cdot\vec{r})\vec{r}-\vec{\mu}r^2}{r^5} \tag{13.24}$$

Similar expressions can be written down for ideal electric dipoles and in two-dimensions. They are listed in tables 13.1 and 13.2. (The delta functions will be discussed in the next subsection.)

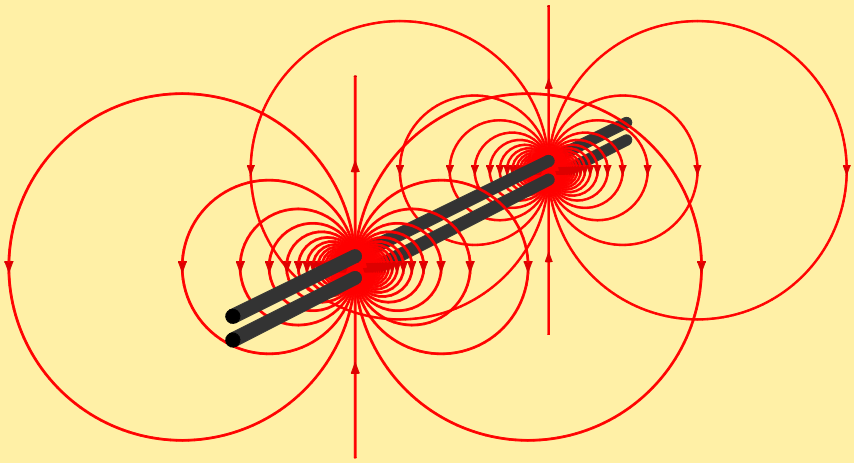


Figure 13.12: Electric field of an almost ideal two-dimensional dipole.

Figure 13.12 shows an *almost* ideal two-dimensional electric dipole. The spacing between the charges has been reduced significantly compared to that in figure 13.10, and the strength of the charges has been increased. For two-dimensional ideal dipoles, the field lines in a cross-plane are circles that all touch each other at the dipole.

13.3.3 Arbitrary charge distributions

Modeling electric systems like atoms and molecules and their ions as singular point charges or dipoles is not very accurate, except in a detailed quantum solution. In a classical description, it is more reasonable to assume that the charges are "smeared out" over space into a distribution. In that case, the charges are described by the charge per unit volume, called the charge density ρ. The integral of the charge density over volume then gives the net charge,

$$q_{\text{region}} = \int_{\text{region}} \rho(\underline{\vec{r}})\, \mathrm{d}^3\underline{\vec{r}} \tag{13.25}$$

As far as the potential is concerned, each little piece $\rho(\underline{\vec{r}})\, \mathrm{d}^3\underline{\vec{r}}$ of the charge distribution acts like a point charge at the point $\underline{\vec{r}}$. The expression for the potential of such a point charge is like that of a point charge at the origin, but with $\vec{r}$ replaced by $\vec{r} - \underline{\vec{r}}$. The total potential results from integrating over all the point charges. So, for a charge distribution,

$$\varphi(\vec{r}) = \frac{1}{4\pi\epsilon_0} \int_{\text{all } \underline{\vec{r}}} \frac{1}{|\vec{r} - \underline{\vec{r}}|} \rho(\underline{\vec{r}})\, \mathrm{d}^3\underline{\vec{r}} \tag{13.26}$$

The electric field and similar expression for magnetic charge distributions and in two dimensions may be found in table 13.2

Note that when the integral expression for the potential is differentiated to find the electric field, as in table 13.2, the integrand becomes much more singular at the point of integration where $\underline{\vec{r}} = \vec{r}$. This may be of importance in numerical work, where the more singular integrand can lead to larger errors. It may then be a better idea not to differentiate under the integral, but instead put the derivative of the charge density in the integral, like in

$$\mathcal{E}_x = -\frac{\partial \varphi}{\partial x} = -\frac{1}{4\pi\epsilon_0} \int_{\text{all } \underline{\vec{r}}} \frac{1}{|\vec{r} - \underline{\vec{r}}|} \frac{\partial \rho(\underline{\vec{r}})}{\partial \underline{x}}\, \mathrm{d}^3\underline{\vec{r}}$$

and similar for the y and z components. That you can do that may be verified by noting that differentiating $\vec{r} - \underline{\vec{r}}$ with respect to x is within a minus sign the same as differentiating with respect to $\underline{x}$, and then you can use integration by parts to move the derivative to ρ.

Now consider the case that the charge distribution is restricted to a very small region around the origin, or equivalently, that the charge distribution is viewed from a very large distance. For simplicity, assume the case that the charge distribution is restricted to a small region around the origin. In that case, $\underline{\vec{r}}$ is small wherever there is charge; the integrand can therefore be approximated by a Taylor series in terms of $\underline{\vec{r}}$ to give:

$$\varphi = \frac{1}{4\pi\epsilon_0} \int_{\text{all } \underline{\vec{r}}} \left[\frac{1}{|\vec{r}|} + \frac{\vec{r}}{|\vec{r}|^3} \cdot \underline{\vec{r}} + \ldots \right] \rho(\underline{\vec{r}})\, \mathrm{d}^3\underline{\vec{r}}$$

where (13.20) was used to evaluate the gradient of $1/|\underline{\vec{r}} - \vec{r}|$ with respect to $\underline{\vec{r}}$.

Since the fractions no longer involve $\underline{\vec{r}}$, they can be taken out of the integrals and so the potential simplifies to

$$\varphi = \frac{q}{4\pi\epsilon_0}\frac{1}{r} + \frac{1}{4\pi\epsilon_0}\frac{\vec{\wp}\cdot\vec{r}}{r^3} + \ldots \qquad q \equiv \int_{\text{all } \underline{\vec{r}}} \rho(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \quad \vec{\wp} \equiv \int_{\text{all } \underline{\vec{r}}} \underline{\vec{r}}\rho(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \tag{13.27}$$

The leading term shows that a distributed charge distribution will normally look like a point charge located at the origin when seen from a sufficient distance. However, if the net charge q is zero, like happens for a neutral atom or molecule, it will look like an ideal dipole, the second term, when seen from a sufficient distance.

The expansion (13.27) is called a "multipole expansion." It allows the effect of a complicated charge distribution to be described by a few simple terms, assuming that the distance from the charge distribution is sufficiently large that its small scale features can be ignored. If necessary, the accuracy of the expansion can be improved by using more terms in the Taylor series. Now recall from the previous section that one advantage of Maxwell's equations over Coulomb's law is that they allow you to describe the electric field at a point using purely local quantities, rather than having to consider the charges everywhere. But using a multipole expansion, you can simplify the effects of distant charge distributions. Then Coulomb's law *can* become competitive with Maxwell's equations, especially in cases where the charge distribution is restricted to a relatively limited fraction of the total space.

The previous subsection discussed how an ideal dipole could be created by decreasing the distance between two opposite charges with a compensating increase in their strength. The multipole expansion above shows that the same ideal dipole is obtained for a continuous charge distribution, provided that the net charge q is zero.

The electric field of this ideal dipole can be found as minus the gradient of the potential. But caution is needed; the so-obtained electric field may not be sufficient for your needs. Consider the following ballpark estimates. Assume that the charge distribution has been contracted to a typical small size ε. Then the net positive and negative charges will have been increased by a corresponding factor $1/\varepsilon$. The electric field within the contracted charge distribution will then have a typical magnitude $1/\varepsilon|\underline{\vec{r}} - \vec{r}|^2$, and that means $1/\varepsilon^3$, since the typical size of the region is ε. Now a quantity of order $1/\varepsilon^3$ can integrate to a finite amount even if the volume of integration is small of order ε^3. In other words, there seems to be a possibility that the electric field may have a delta function hidden within the charge distribution when it is contracted to a point. And so it does. The correct delta function is derived in derivation {D.72} and shown in table 13.2. It is important in applications in quantum mechanics where you need some integral of the electric field; if you forget about the delta function,

you will get the wrong result.

13.3.4 Solution of the Poisson equation

The previous subsections stumbled onto the solution of an important mathematical problem, the Poisson equation. The Poisson equation is

$$\nabla^2\varphi = f \tag{13.28}$$

where f is a given function and φ is the unknown one to be found. The Laplacian ∇^2 is also often found written as Δ.

The reason that the previous subsection stumbled on to the solution of this equation is that the electric potential φ satisfies it. In particular, minus the gradient of φ gives the electric field; also, the divergence of the electric field gives according to Maxwell's first equation the charge density ρ divided by ϵ_0. Put the two together and it says that $\nabla^2\varphi = -\rho/\epsilon_0$. So, identify the function f in the Poisson equation with $-\rho/\epsilon_0$, and there you have the solution of the Poisson equation.

Because it is such an important problem, it is a good idea to write out the abstract mathematical solution without the "physical entourage" of (13.26):

$$\nabla^2\varphi = f \quad \Longrightarrow \quad \varphi(\vec{r}) = \int_{\text{all } \underline{\vec{r}}} G(\vec{r}-\underline{\vec{r}})f(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} \qquad G(\vec{r}) = -\frac{1}{4\pi|\vec{r}|} \tag{13.29}$$

The function $G(\vec{r}-\underline{\vec{r}})$ is called the Green's function of the Laplacian. It is the solution for φ if the function f is a delta function at point $\underline{\vec{r}}$. The integral solution of the Poisson equation can therefore be understood as dividing function f up into spikes $f(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}}$; for each of these spikes the contribution to φ is given by corresponding Green's function.

It also follows that applying the Laplacian on the Green's function produces the three-dimensional delta function,

$$\nabla^2 G(\vec{r}) = \delta^3(\vec{r}) \qquad G(\vec{r}) = -\frac{1}{4\pi|\vec{r}|} \tag{13.30}$$

with $|\vec{r}| = r$ in spherical coordinates. That sometimes pops up in quantum mechanics, in particular in perturbation theory. You might object that the Green's function is infinite at $\vec{r} = 0$, so that its Laplacian is undefined there, rather than a delta function spike. And you would be perfectly right; just saying that the Laplacian of the Green's function is the delta function is not really justified. However, if you slightly round the Green's function near $\vec{r} = 0$, say like φ was rounded in figure 13.7, its Laplacian does exist everywhere. The Laplacian of this rounded Green's function is a spike confined to the region of rounding, and it integrates to one. (You can see the latter from applying the divergence theorem on a sphere enclosing the region of rounding.) If you then

contract the region of rounding to zero, this spike becomes a delta function in the limit of no rounding. Understood in this way, the Laplacian of the Green's function is indeed a delta function.

The multipole expansion for a charge distribution can also be converted to purely mathematical terms:

$$\varphi = -\frac{1}{4\pi r}\int_{\text{all } \underline{\vec{r}}} f(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} - \frac{\vec{r}}{4\pi r^3}\cdot\int_{\text{all } \underline{\vec{r}}} \underline{\vec{r}} f(\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}} + \ldots \tag{13.31}$$

(Of course, delta functions are infinite objects, and you might wonder at the mathematical rigor of the various arguments above. However, there are solid arguments based on "Green's second integral identity" that avoid the infinities and produce the same final results.)

13.3.5 Currents

Streams of moving electric charges are called currents. The current strength I through an electric wire is defined as the amount of charge flowing through a cross section per unit time. It equals the amount of charge q' per unit length times its velocity v;

$$I \equiv q'v \tag{13.32}$$

The current density $\vec{\jmath}$ is defined as the current per unit volume, and equals the charge density times the charge velocity. Integrating the current density over the cross section of a wire gives its current.

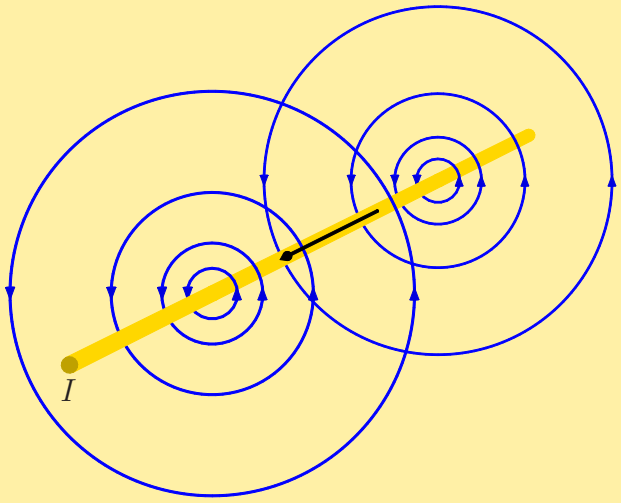


Figure 13.13: Magnetic field lines around an infinite straight electric wire.

As shown in figure 13.13, electric wires are encircled by magnetic field lines. The strength of this magnetic field may be computed from Maxwell's fourth equation. To do so, take an arbitrary field line circle. The field strength is

constant on the line by symmetry. So the integral of the field strength along the line is just $2\pi r\mathcal{B}$; the perimeter of the field line times its magnetic strength. Now the Stokes' theorem of calculus says that this integral is equal to the curl of the magnetic field integrated over the interior of the field line circle. And Maxwell's fourth equation says that that is $1/\epsilon_0 c^2$ times the current density integrated over the circle. And the current density integrated over the circle is just the current through the wire. Put it all together to get

$$\text{magnetic field of an infinite straight wire:} \quad \mathcal{B} = \frac{I}{2\pi\epsilon_0 c^2 r} \tag{13.33}$$

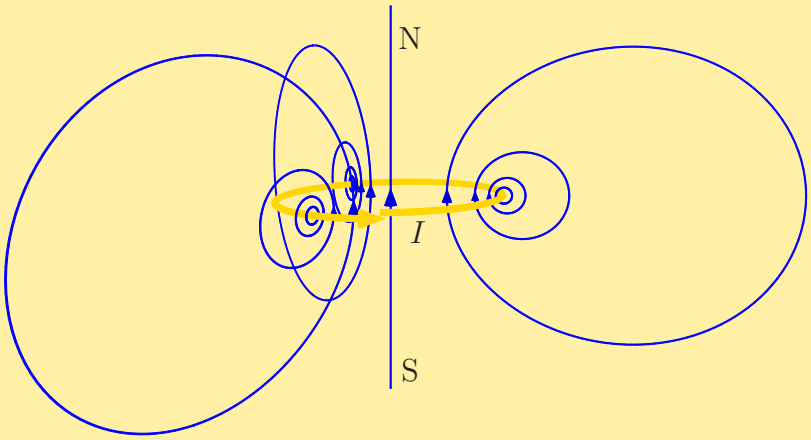


Figure 13.14: An electromagnet consisting of a single wire loop. The generated magnetic field lines are in blue.

An infinite straight wire is of course not a practical way to create a magnetic field. In a typical electromagnet, the wire is spooled around an iron bar. Figure 13.14 shows the field produced by a single wire loop, in vacuum. To find the fields produced by curved wires, use the so-called "Biot-Savart law" listed in table 13.2 and derived in {D.72}. You need it when you end up writing a book on quantum mechanics and have to plot the field.

Of course, while figure 13.14 does not show it, you will also need a lead from your battery to the electromagnet and a second lead back to the other pole of the battery. These two leads form a two-dimensional "current dipole," as shown in figure 13.15, and they produce a magnetic field too. However, the currents in the two leads are opposite; one coming from the battery and other returning to it, so the magnetic fields that they create are opposite. Therefore, if you strand the wires very closely together, their magnetic fields will cancel each other, and not mess up that of your electromagnet.

It may be noted that if you bring the wires close together, whatever is left of the field has circular field lines that touch at the dipole. In other words, a horizontal ideal current dipole produces the same field as a two-dimensional

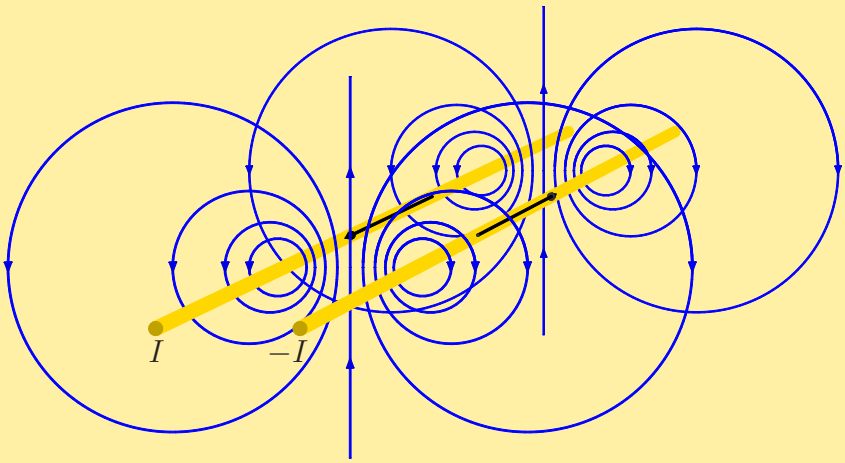


Figure 13.15: A current dipole.

vertical ideal charge dipole. Similarly, the horizontal wire loop, if small enough, produces the same field lines as a three-dimensional vertical ideal charge dipole. (However, the delta functions are different, {D.72}.)

13.3.6 Principle of the electric motor

The previous section discussed how Maxwell's third equation allows electric power generation using mechanical means. The converse is also possible; electric power allows mechanical power to be generated; that is the principle of the electric motor.

It is possible because of the Lorentz force law, which says that a charge q moving with velocity $\vec{v}$ in a magnetic field $\vec{\mathcal{B}}$ experiences a force pushing it sideways equal to

$$\vec{F} = q\vec{v} \times \vec{\mathcal{B}}$$

Consider the wire loop in an external magnetic field sketched in figure 13.16. The sideways forces on the current carriers in the wire produce a net moment $\vec{M}$ on the wire loop that allows it to perform useful work.

To be more precise, the forces caused by the component of the magnetic field normal to the wire loop are radial and produce no net force nor moment. However, the forces caused by the component of the magnetic field parallel to the loop produce forces normal to the plane of the loop that do generate a net moment. Using spherical coordinates aligned with the wire loop as in figure 13.17, the component of the magnetic field parallel to the loop equals $\mathcal{B}_{\text{ext}} \sin\theta$. It causes a sideways force on each element $r\mathrm{d}\phi$ of the wire equal to

$$\mathrm{d}F = \underbrace{q' r \mathrm{d}\phi}_{\mathrm{d}q} \underbrace{v \mathcal{B}_{\text{ext}} \sin\theta \sin\phi}_{\vec{v}\times\vec{\mathcal{B}}_{\text{parallel}}}$$

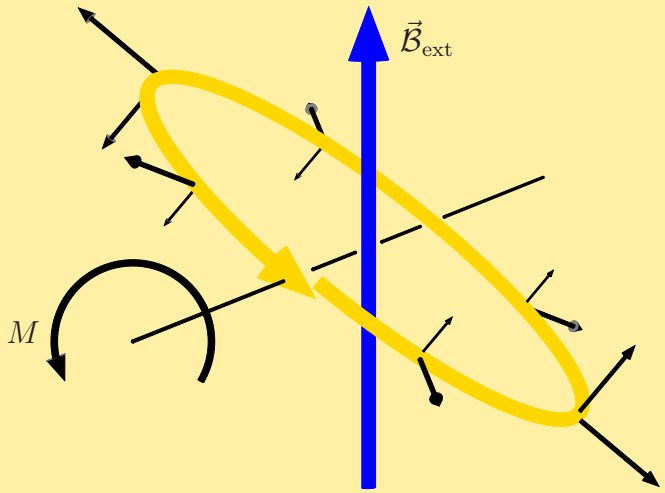


Figure 13.16: Electric motor using a single wire loop. The Lorentz forces (black vectors) exerted by the external magnetic field on the electric current carriers in the wire produce a net moment M on the loop. The self-induced magnetic field of the wire and the corresponding radial forces are not shown.

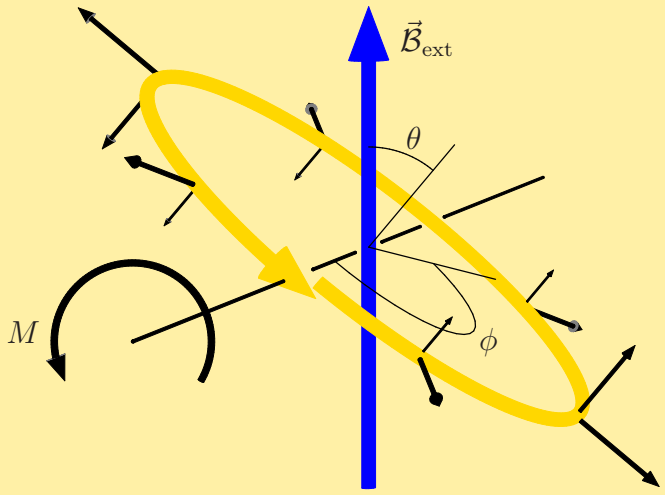


Figure 13.17: Variables for the computation of the moment on a wire loop in a magnetic field.

where q' is the net charge of current carriers per unit length and v their velocity. The corresponding net force integrates to zero. However the moment does not; integrating

$$\mathrm{d}M = \underbrace{r\sin\phi}_{\text{arm}}\underbrace{q'r\mathrm{d}\phi v\mathcal{B}_{\text{ext}}\sin\theta\sin\phi}_{\text{force}}$$

produces

$$M = \pi r^2 q' v\mathcal{B}_{\text{ext}}\sin\theta$$

If the work $M\mathrm{d}\theta$ done by this moment is formulated as a change in energy of the loop in the magnetic field, that energy is

$$E_{\text{ext}} = -\pi r^2 q' v\mathcal{B}_{\text{ext}}\cos\theta$$

The magnetic dipole moment $\vec{\mu}$ is defined as the factor that only depends on the wire loop, independent of the magnetic field. In particular $\mu = \pi r^2 q' v$ and it is taken to be in the axial direction. So the moment and energy can be written more concisely as

$$\vec{M} = \vec{\mu}\times\vec{\mathcal{B}}_{\text{ext}} \qquad E_{\text{ext}} = -\vec{\mu}\cdot\vec{\mathcal{B}}_{\text{ext}}$$

Yes, $\vec{\mu}$ also governs how the magnetic field looks at large distances; feel free to approximate the Biot-Savart integral for large distances to check.

A book on electromagnetics would typically identify $q'v$ with the current through the wire I and πr^2 with the area of the loop, so that the magnetic dipole moment is just IA. This is then valid for a flat wire loop of any shape, not just a circular one.

But this is a book on quantum mechanics, and for electrons in orbits about nuclei, currents and areas are not very useful. In quantum mechanics the more meaningful quantity is angular momentum. So identify $2\pi r q'$ as the total electric charge going around in the wire loop, and multiply that with the ratio $m_{\text{c}}/q_{\text{c}}$ of mass of the current carrier to its charge to get the total mass going around. Then multiply with rv to get the angular momentum L. In those terms, the magnetic dipole moment is

$$\vec{\mu} = \frac{q_{\text{c}}}{2m_{\text{c}}}\vec{L} \tag{13.34}$$

Usually the current carrier is an electron, so $q_{\text{c}} = -e$ and $m_{\text{c}} = m_{\text{e}}$.

These results apply to any arbitrary current distribution, not just a circular wire loop. Formulae are in table 13.2 and general derivations in {D.72}.

13.4 Particles in Magnetic Fields

Maxwell's equations are fun, but back to real quantum mechanics. The serious question in this section is how a magnetic field $\vec{\mathcal{B}}$ affects a quantum system, like say an electron in an hydrogen atom.

Well, if the Hamiltonian (13.2) for a charged particle is written out and cleaned up, {D.73}, it is seen that a constant magnetic field adds two terms. The most important of the two is

$$\boxed{H_{\mathcal{B}L} = -\frac{q}{2m}\vec{\mathcal{B}} \cdot \widehat{\vec{L}}} \tag{13.35}$$

where q is the charge of the particle, m its mass, $\vec{\mathcal{B}}$ the external magnetic field, assumed to be constant on the scale of the atom, and $\widehat{\vec{L}}$ is the orbital angular momentum of the particle.

In terms of classical physics, this can be understood as follows: a particle with angular momentum $\vec{L}$ can be pictured to be circling around the axis through $\vec{L}$. Now according to Maxwell's equations, a charged particle going around in a circle acts as a little electromagnet. Think of a version of figure 13.6 using a circular path. And a little magnet wants to align itself with an ambient magnetic field, just like a magnetic compass needle aligns itself with the magnetic field of earth.

In electromagnetics, the effective magnetic strength of a circling charged particle is described by the so called orbital "magnetic dipole moment" $\vec{\mu}_L$, defined as

$$\vec{\mu}_L \equiv \frac{q}{2m}\vec{L}. \tag{13.36}$$

In terms of this magnetic dipole moment, the energy is

$$H_{\mathcal{B}L} = -\vec{\mu}_L \cdot \vec{\mathcal{B}}. \tag{13.37}$$

which is the lowest when the magnetic dipole moment is in the same direction as the magnetic field.

The scalar part of the magnetic dipole moment, to wit,

$$\gamma_L = \frac{q}{2m} \tag{13.38}$$

is called the "gyromagnetic ratio." But since in quantum mechanics the orbital angular momentum comes in chunks of size $\hbar$, and the particle is usually an electron with charge $q = -e$, much of the time you will find instead the "Bohr magneton"

$$\boxed{\mu_{\rm B} = \frac{e\hbar}{2m_{\rm e}} \approx 9.274\ 10^{-24}\ {\rm J/T}} \tag{13.39}$$

used. Here T stands for Tesla, the kg/C-s unit of magnetic field strength.

Please, all of this is serious; this is not a story made up by this book to put physicists in a bad light. Note that the original formula had four variables in it: q, m, $\vec{\mathcal{B}}$, and $\widehat{\vec{L}}$, and the three new names they want you to remember are less than that.

The big question now is: since electrons have spin, build-in angular momentum, do they still act like little magnets even if not going around in a circle? The answer is yes; there is an additional term in the Hamiltonian due to spin. Astonishingly, the energy involved pops out of Dirac's relativistic description of the electron, {D.74}. The energy that an electron picks up in a magnetic field due to its inherent spin is:

$$\boxed{H_{\mathcal{B}S} = -g_e \frac{q}{2m_{\rm e}} \vec{\mathcal{B}} \cdot \widehat{\vec{S}} \qquad g_e \approx 2 \quad q = -e} \tag{13.40}$$

(This section uses again S to indicate spin angular momentum.) The constant g is called the "g-factor". Since its value is 2, electron spin produces twice the magnetic dipole strength as the same amount of orbital angular momentum would. That is called the "magnetic spin anomaly," [52, p. 222].

It should be noted that really the g-factor of an electron is about 0.1% larger than 2 because of the quantization of the electromagnetic field ignored in the Dirac equation. The quantized electromagnetic field, whose particle is the photon, has quantum uncertainty. You can think of it qualitatively as virtual photons popping up and disappearing continuously according to the energy-time uncertainty $\Delta E \Delta t \approx \hbar$, allowing particles with energy ΔE to appear as long as they don't stay around longer than a very brief time Δt. "Quantum electrodynamics" says that to a better approximation $g \approx 2 + \alpha/\pi$ where $\alpha = e^2/4\pi\epsilon_0\hbar c \approx 1/137$ is called the fine structure constant. This correction to g, due to the possible interaction of the electron with a virtual photon, [19, p. 116], is called the "anomalous magnetic moment," [25, p. 273]. (The fact that physicists have not yet defined potential deviations from the quantum electrodynamics value to be "magnetic spin anomaly anomalous magnetic moment anomalies" is an anomaly.) The prediction of the g-factor of the electron is a test for the accuracy of quantum electrodynamics, and so this g-factor has been measured to exquisite precision. At the time of writing, (2008), the experimental value is 2.002 319 304 362, to that many correct digits. Quantum electrodynamics has managed to get things right to more than ten digits by including more and more, increasingly complex interactions with virtual photons and virtual electron/positron pairs, [19], one of the greatest achievements of twentieth century physics.

You might think that the above formula for the energy of an electron in a magnetic field should also apply to protons and neutrons, since they too are spin ½ particles. However, this turns out to be untrue. Protons and neutrons are not elementary particles, but consist of three "quarks." Still, for both electron and proton spin the gyromagnetic ratio can be written as

$$\gamma_S = g \frac{q}{2m} \tag{13.41}$$

but while the g-factor of the electron is 2, the measured one for the proton is 5.59.

Do note that due to the much larger mass of the proton, its actual magnetic dipole moment is much less than that of an electron despite its larger g-factor. Still, under the right circumstances, like in nuclear magnetic resonance, the magnetic dipole moment of the proton is crucial despite its relative small size.

For the neutron, the charge is zero, but the magnetic moment is not, which would make its g-factor infinite! The problem is that the quarks that make up the neutron *do* have charge, and so the neutron can interact with a magnetic field even though its *net* charge is zero. When the *proton* mass and charge are arbitrarily used in the formula, the neutron's g-factor is -3.83. More generally, nuclear magnetic moments are expressed in terms of the "nuclear magneton"

$$\boxed{\mu_{\rm N} = \frac{e\hbar}{2m_{\rm p}} \approx 5.050\,78\ 10^{-27}\ {\rm J/T}} \tag{13.42}$$

that is based on proton charge and mass. Therefore nuclear g-factors are simply twice the nuclear magnetic moment in magnetons. (Needless to say, some authors leave out the factor 2 for that additional touch of confusion.)

At the start of this subsection, it was noted that the Hamiltonian for a charged particle has another term. So, how about it? It is called the "diamagnetic contribution," and it is given by

$$\boxed{H_{\mathcal{B}D} = \frac{q^2}{8m}\left(\vec{\mathcal{B}} \times \widehat{\vec{r}}\right)^2} \tag{13.43}$$

Note that a system, like an atom, minimizes this contribution by staying away from magnetic fields: it is positive and proportional to $\mathcal{B}^2$.

The diamagnetic contribution can usually be ignored if there is net orbital or spin angular momentum. To see why, consider the following numerical values:

$$\mu_{\rm B} = \frac{e\hbar}{2m_{\rm e}} \approx 5.788\ 10^{-5}\ {\rm eV/T} \qquad \frac{e^2a_0^2}{8m_{\rm e}} = 6.156\,5\ 10^{-11}\ {\rm eV/T^2}$$

The first number gives the magnetic dipole energy, for a quantum of angular momentum, per Tesla, while the second number gives the diamagnetic energy, for a Bohr-radius spread around the magnetic axis, per square Tesla.

It follows that it takes about a million Tesla for the diamagnetic energy to become comparable to the dipole one. Now at the time of this writing, (2008), the world record magnet that can operate continuously is right here at the Florida State University. It produces a field of 45 Tesla, taking in 33 MW of electricity and 4 000 gallons of cooling water per minute. The world record magnet that can produce even stronger brief magnetic pulses is also here, and it produces 90 Tesla, going on 100. (Still stronger magnetic fields are possible if you allow the magnet to blow itself to smithereens during the fraction of a second that it operates, but that is so messy.) Obviously, these numbers are way below a million Tesla. Also note that since atom energies are in electron volts or more, none of these fields are going to blow an atom apart.

13.5 Stern-Gerlach Apparatus

A constant magnetic field will exert a torque, but no net force on a magnetic dipole like an electron; if you think of the dipole as a magnetic north pole and south pole close together, the magnetic forces on north pole and south pole will be opposite and produce no net force on the dipole. However, if the magnetic field strength varies with location, the two forces will be different and a net force will result.

The Stern-Gerlach apparatus exploits this process by sending a beam of atoms through a magnetic field with spatial variation, causing the atoms to deflect upwards or downwards depending on their magnetic dipole strength. The magnetic dipole strengths of the atoms will be proportional to the relevant electron angular momenta, (the nucleus can be ignored because of the large mass in its gyromagnetic ratio), and that will be quantized. So the incoming beam will split into *distinct* beams corresponding to the quantized values of the electron angular momentum.

The experiment was a great step forward in the development of quantum mechanics, because there is really no way that classical mechanics can explain the splitting into separate beams; classical mechanics just has to predict a smeared-out beam. Angular momentum in classical mechanics can have any value, not just the values $m\hbar$ of quantum mechanics. Moreover, by capturing one of the split beams, you have a source of particles all in the *same* state without uncertainty, to use for other experiments or practical applications such as masers.

Stern and Gerlach used a beam of silver atoms in their experiment, and the separated beams deposited this silver on a plate. Initially, Gerlach had difficulty seeing any deposited silver on those plates because the layer was extremely thin. But fortunately for quantum mechanics, Stern was puffing his usual cheap cigars when he had a look, and the large amount of sulphur in the smoke was enough to turn some of the silver into jet-black silver sulfide, making it show clearly.

An irony is that that Stern and Gerlach assumed that that they had verified Bohr's orbital momentum. But actually, they had discovered spin. The net magnetic moment of silver's inner electrons is zero, and the lone valence electron is in a 5s orbit with zero orbital angular momentum. It was the spin of the valence electron that caused the splitting. While spin has half the strength of orbital angular momentum, its magnetic moment is about the same due to its g-factor being two rather than one.

To use the Stern Gerlach procedure with charged particles such as lone electrons, a transverse electric field must be provided to counteract the large Lorentz force that the magnet imparts on the moving electrons.

13.6 Nuclear Magnetic Resonance

Nuclear magnetic resonance, or NMR, is a valuable tool for examining nuclei, for probing the structure of molecules, in particular organic ones, and for medical diagnosis, as MRI. This section will give a basic quantum description of the idea. Linear algebra will be used.

13.6.1 Description of the method

First demonstrated independently by Bloch and Purcell in 1946, NMR probes nuclei with net spin, in particular hydrogen nuclei or other nuclei with spin $\frac{1}{2}$. Various common nuclei, like carbon and oxygen do not have net spin; this can be a blessing since they cannot mess up the signals from the hydrogen nuclei, or a limitation, depending on how you want to look at it. In any case, if necessary isotopes such as carbon 13 can be used which do have net spin.

It is not actually the spin, but the associated magnetic dipole moment of the nucleus that is relevant, for that allows the nuclei to be manipulated by magnetic fields. First the sample is placed in an extremely strong steady magnetic field. Typical fields are in terms of Tesla. (A Tesla is about 20 000 times the strength of the magnetic field of the earth.) In the field, the nucleus has two possible energy states; a ground state in which the spin component in the direction of the magnetic field is aligned with it, and an elevated energy state in which the spin is opposite {N.33}. (Despite the large field strength, the energy difference between the two states is extremely small compared to the thermal kinetic energy at room temperature. The number of nuclei in the ground state may only exceed those in the elevated energy state by say one in 100 000, but that is still a large absolute number of nuclei in a sample.)

Now perturb the nuclei with a second, much smaller and radio frequency, magnetic field. If the radio frequency is just right, the excess ground state nuclei can be lifted out of the lowest energy state, absorbing energy that can be observed. The "resonance" frequency at which this happens then gives information about the nuclei. In order to observe the resonance frequency very accurately, the perturbing rf field must be very weak compared to the primary steady magnetic field.

In Continuous Wave NMR, the perturbing frequency is varied and the absorption examined to find the resonance. (Alternatively, the strength of the primary magnetic field can be varied, that works out to the same thing using the appropriate formula.)

In Fourier Transform NMR, the perturbation is applied in a brief pulse just long enough to fully lift the excess nuclei out of the ground state. Then the decay back towards the original state is observed. An experienced operator can then learn a great deal about the environment of the nuclei. For example, a nucleus in a molecule will be shielded a bit from the primary magnetic field by

the rest of the molecule, and that leads to an observable frequency shift. The amount of the shift gives a clue about the molecular structure at the nucleus, so information about the molecule. Additionally, neighboring nuclei can cause resonance frequencies to split into several through their magnetic fields. For example, a single neighboring perturbing nucleus will cause a resonance frequency to split into two, one for spin up of the neighboring nucleus and one for spin down. It is another clue about the molecular structure. The time for the decay back to the original state to occur is another important clue about the local conditions the nuclei are in, especially in MRI. The details are beyond this author's knowledge; the purpose here is only to look at the basic quantum mechanics behind NMR.

13.6.2 The Hamiltonian

The magnetic fields will be assumed to be of the form

$$\vec{\mathcal{B}} = \mathcal{B}_0\hat{k} + \mathcal{B}_1\left(\hat{\imath}\cos\omega t - \hat{\jmath}\sin\omega t\right) \tag{13.44}$$

where $\mathcal{B}_0$ is the Tesla-strength primary magnetic field, $\mathcal{B}_1$ the very weak perturbing field strength, and ω is the frequency of the perturbation.

The component of the magnetic field in the xy-plane, $\mathcal{B}_1$, rotates around the z-axis at angular velocity ω. Such a rotating magnetic field can be achieved using a pair of properly phased coils placed along the x and y axes. (In Fourier Transform NMR, a single perturbation pulse actually contains a range of different frequencies ω, and Fourier transforms are used to take them apart.) Since the apparatus and the wave length of a radio frequency field is very large on the scale of a nucleus, spatial variations in the magnetic field can be ignored.

Now suppose you place a spin $\frac{1}{2}$ nucleus in the center of this magnetic field. As discussed in section 13.4, a particle with spin will act as a little compass needle, and its energy will be lowest if it is aligned with the direction of the ambient magnetic field. In particular, the energy is given by

$$H = -\vec{\mu}\cdot\vec{\mathcal{B}}$$

where $\vec{\mu}$ is called the magnetic dipole strength of the nucleus. This dipole strength is proportional to its spin angular momentum $\widehat{\vec{S}}$:

$$\vec{\mu} = \gamma\widehat{\vec{S}}$$

where the constant of proportionality γ is called the gyromagnetic ratio. The numerical value of the gyromagnetic ratio can be found as

$$\gamma = \frac{gq}{2m}$$

In case of a hydrogen nucleus, a proton, the mass m_p and charge $q_p = e$ can be found in the notations section, and the proton's experimentally found g-factor is $g_p = 5.59$.

The bottom line is that you can write the Hamiltonian of the interaction of the nucleus with the magnetic field in terms of a numerical gyromagnetic ratio value, spin, and the magnetic field:

$$H = -\gamma \widehat{\vec{S}} \cdot \vec{\mathcal{B}} \tag{13.45}$$

Now turning to the wave function of the nucleus, it can be written as a combination of the spin-up and spin-down states,

$$\Psi = a\uparrow + b\downarrow,$$

where $\uparrow$ has spin $\frac{1}{2}\hbar$ in the z-direction, along the primary magnetic field, and $\downarrow$ has $-\frac{1}{2}\hbar$. Normally, a and b would describe the spatial variations, but spatial variations are not relevant to the analysis, and a and b can be considered to be simple numbers.

You can use the concise notations of linear algebra by combining a and b in a two-component column vector (more precisely, a spinor),

$$\Psi = \begin{pmatrix} a \\ b \end{pmatrix}$$

In those terms, the spin operators become matrices, the so-called Pauli spin matrices of section 12.10,

$$\widehat{S}_x = \frac{\hbar}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \widehat{S}_y = \frac{\hbar}{2}\begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \quad \widehat{S}_z = \frac{\hbar}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{13.46}$$

Substitution of these expressions for the spin, and (13.44) for the magnetic field into (13.45) gives after cleaning up the final Hamiltonian:

$$H = -\frac{\hbar}{2}\begin{pmatrix} \omega_0 & \omega_1 e^{\mathrm{i}\omega t} \\ \omega_1 e^{-\mathrm{i}\omega t} & -\omega_0 \end{pmatrix} \qquad \omega_0 = \gamma \mathcal{B}_0 \quad \omega_1 = \gamma \mathcal{B}_1 \tag{13.47}$$

The constants ω_0 and ω_1 have the dimensions of a frequency; ω_0 is called the "Larmor frequency." As far as ω_1 is concerned, the important thing to remember is that it is much smaller than the Larmor frequency ω_0 because the perturbation magnetic field is small compared to the primary one.

13.6.3 The unperturbed system

Before looking at the perturbed case, it helps to first look at the unperturbed solution. If there is just the primary magnetic field affecting the nucleus, with

no radio-frequency perturbation ω_1, the Hamiltonian derived in the previous subsection simplifies to

$$H = -\frac{\hbar}{2}\begin{pmatrix} \omega_0 & 0 \\ 0 & -\omega_0 \end{pmatrix}$$

The energy eigenstates are the spin-up state, with energy $-\frac{1}{2}\hbar\omega_0$, and the spin-down state, with energy $\frac{1}{2}\hbar\omega_0$.

The difference in energy is in relativistic terms exactly equal to a photon with the Larmor frequency ω_0. While the treatment of the electromagnetic field in this discussion will be classical, rather than relativistic, it seems clear that the Larmor frequency must play more than a superficial role.

The unsteady Schrödinger equation tells you that the wave function evolves in time like $\mathrm{i}\hbar\dot{\Psi} = H\Psi$, so if $\Psi = a\uparrow + b\downarrow$,

$$\mathrm{i}\hbar\begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} = -\frac{\hbar}{2}\begin{pmatrix} \omega_0 & 0 \\ 0 & -\omega_0 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}$$

The solution for the coefficients a and b of the spin-up and -down states is:

$$a = a_0 e^{\mathrm{i}\omega_0 t/2} \qquad b = b_0 e^{-\mathrm{i}\omega_0 t/2}$$

if a_0 and b_0 are the values of these coefficients at time zero.

Since $|a|^2 = |a_0|^2$ and $|b|^2 = |b_0|^2$ at all times, the probabilities of measuring spin-up or spin-down do not change with time. This was to be expected, since spin-up and spin-down are energy states for the steady system. To get more interesting physics, you really need the unsteady perturbation.

But first, to understand the quantum processes better in terms of the ideas of nonquantum physics, it will be helpful to write the unsteady quantum evolution in terms of the *expectation values* of the angular momentum components. The expectation value of the z-component of angular momentum is

$$\langle S_z \rangle = |a|^2\frac{\hbar}{2} - |b|^2\frac{\hbar}{2}$$

To more clearly indicate that the value must be in between $-\hbar/2$ and $\hbar/2$, you can write the magnitude of the coefficients in terms of an angle α, the "precession angle",

$$|a| = |a_0| \equiv \cos(\alpha/2) \qquad |b| = |b_0| \equiv \sin(\alpha/2)$$

In terms of the so-defined α, you simply have, using the half-angle trig formulae,

$$\langle S_z \rangle = \frac{\hbar}{2}\cos\alpha$$

The expectation values of the angular momenta in the x and y directions can by found as the inner products $\langle\Psi|\widehat{S}_x\Psi\rangle$ and $\langle\Psi|\widehat{S}_y\Psi\rangle$, chapter 4.4.3. Substituting the representation in terms of spinors and Pauli spin matrices, and

cleaning up using the Euler formula (2.5), you get

$$\langle S_x \rangle = \frac{\hbar}{2} \sin\alpha \cos(\omega_0 t + \alpha) \quad \langle S_y \rangle = -\frac{\hbar}{2} \sin\alpha \sin(\omega_0 t + \alpha)$$

where α is some constant phase angle that is further unimportant.

The first thing that can be seen from these results is that the length of the expectation angular momentum vector is $\hbar/2$. Next, the component with the z-axis, the direction of the primary magnetic field, is at all times $\frac{1}{2}\hbar\cos\alpha$. That implies that the expectation angular momentum vector is under a constant angle α with the primary magnetic field.

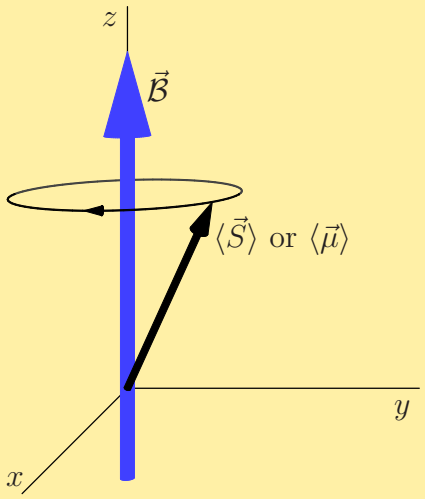


Figure 13.18: Larmor precession of the expectation spin (or magnetic moment) vector around the magnetic field.

The component in the xy-plane is $\frac{1}{2}\hbar\sin\alpha$, and this component rotates around the z-axis, as shown in figure 13.18, causing the end point of the expectation angular momentum vector to sweep out a circular path around the magnetic field $\vec{\mathcal{B}}$. This rotation around the z-axis is called "Larmor precession." Since the magnetic dipole moment is proportional to the spin, it traces out the same conical path.

Caution should be used against attaching too much importance to this classical picture of a precessing magnet. The expectation angular momentum vector is not a physically measurable quantity. One glaring inconsistency in the expectation angular momentum vector versus the true angular momentum is that the square magnitude of the expectation angular momentum vector is $\hbar^2/4$, three times smaller than the true square magnitude of angular momentum.

13.6.4 Effect of the perturbation

In the presence of the perturbing magnetic field, the unsteady Schrödinger equation $\mathrm{i}\hbar\dot{\Psi} = H\Psi$ becomes

$$\mathrm{i}\hbar \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} = -\frac{\hbar}{2} \begin{pmatrix} \omega_0 & \omega_1 e^{\mathrm{i}\omega t} \\ \omega_1 e^{-\mathrm{i}\omega t} & -\omega_0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \tag{13.48}$$

where ω_0 is the Larmor frequency, ω is the frequency of the perturbation, and ω_1 is a measure of the strength of the perturbation and small compared to ω_0.

The above equations can be solved exactly using standard linear algebra procedures, though the the algebra is fairly stifling {D.75}. The analysis brings in an additional quantity that will be called the "resonance factor"

$$f = \sqrt{\frac{\omega_1^2}{(\omega - \omega_0)^2 + \omega_1^2}} \tag{13.49}$$

Note that f has its maximum value, one, at "resonance," i.e. when the perturbation frequency ω equals the Larmor frequency ω_0.

The analysis finds the coefficients of the spin-up and spin-down states to be:

$$a = \left[a_0 \left(\cos\left(\frac{\omega_1 t}{2f}\right) - \mathrm{i}f\frac{\omega - \omega_0}{\omega_1} \sin\left(\frac{\omega_1 t}{2f}\right) \right) + b_0 \mathrm{i}f \sin\left(\frac{\omega_1 t}{2f}\right) \right] e^{\mathrm{i}\omega t/2} \tag{13.50}$$

$$b = \left[b_0 \left(\cos\left(\frac{\omega_1 t}{2f}\right) + \mathrm{i}f\frac{\omega - \omega_0}{\omega_1} \sin\left(\frac{\omega_1 t}{2f}\right) \right) + a_0 \mathrm{i}f \sin\left(\frac{\omega_1 t}{2f}\right) \right] e^{-\mathrm{i}\omega t/2} \tag{13.51}$$

where a_0 and b_0 are the initial coefficients of the spin-up and spin-down states.

This solution looks pretty forbidding, but it is not that bad in application. The primary interest is in nuclei that start out in the spin-up ground state, so you can set $|a_0| = 1$ and $b_0 = 0$. Also, the primary interest is in the probability that the nuclei may be found at the elevated energy level, which is

$$|b|^2 = f^2 \sin^2\left(\frac{\omega_1 t}{2f}\right) \tag{13.52}$$

That is a pretty simple result. When you start out, the nuclei you look at are in the ground state, so $|b|^2$ is zero, but with time the rf perturbation field increases the probability of finding the nuclei in the elevated energy state eventually to a maximum of f^2 when the sine becomes one.

Continuing the perturbation beyond that time is bad news; it decreases the probability of elevated states again. As figure 13.19 shows, over extended times, there is a flip-flop between the nuclei being with certainty in the ground state, and having a probability of being in the elevated state. The frequency at which the probability oscillates is called the "Rabi flopping frequency". The author's

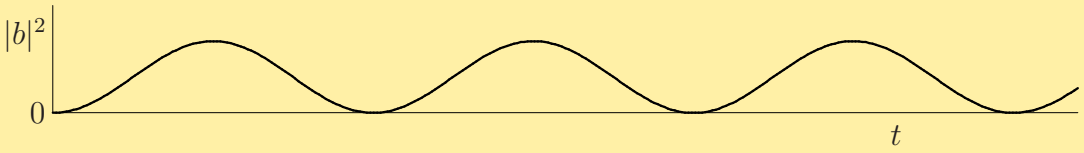


Figure 13.19: Probability of being able to find the nuclei at elevated energy versus time for a given perturbation frequency ω.

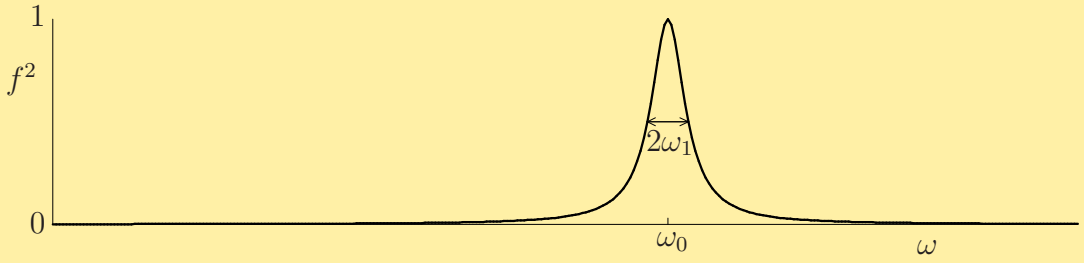


Figure 13.20: Maximum probability of finding the nuclei at elevated energy.

sources differ about the precise definition of this frequency, but the one that seems to be most logical is ω_1/f.

Anyway, by keeping up the perturbation for the right time you can raise the probability of elevated energy to a maximum of f^2. A plot of f^2 against the perturbing frequency ω is called the "resonance curve," shown in figure 13.20. For the perturbation to have maximum effect, its frequency ω must equal the nuclei's Larmor frequency ω_0. Also, for this frequency to be very accurately observable, the "spike" in figure 13.20 must be narrow, and since its width is proportional to $\omega_1 = \gamma\mathcal{B}_1$, that means the perturbing magnetic field must be very weak compared to the primary magnetic field.

There are two qualitative ways to understand the need for the frequency of the perturbation to equal the Larmor frequency. One is geometrical and classical: as noted in the previous subsection, the expectation magnetic moment precesses around the primary magnetic field with the Larmor frequency. In order for the small perturbation field to exert a long-term downward "torque" on this precessing magnetic moment as in figure 13.21, it must rotate along with it. If it rotates at any other frequency, the torque will quickly reverse direction compared to the magnetic moment, and the vector will start going up again. The other way to look at it is from a relativistic quantum perspective: if the magnetic field frequency equals the Larmor frequency, its photons have exactly the energy required to lift the nuclei from the ground state to the excited state.

At the Larmor frequency, it would naively seem that the optimum time to maintain the perturbation is until the expectation spin vector is vertically down; then the nucleus is in the exited energy state with certainty. If you then allow nature the time to probe its state, every nucleus will be found to be in the excited state, and will emit a photon. (If not messed up by some

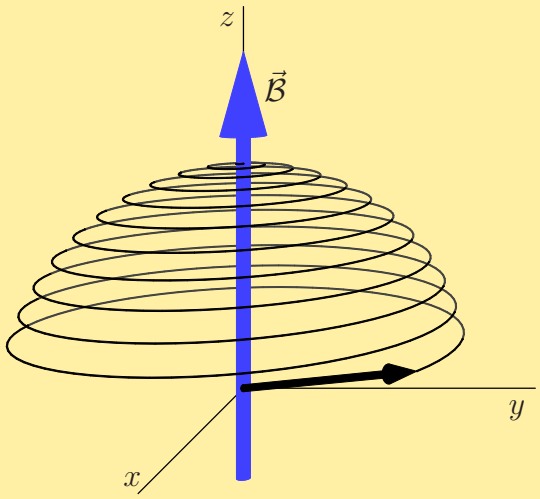


Figure 13.21: A perturbing magnetic field, rotating at precisely the Larmor frequency, causes the expectation spin vector to come cascading down out of the ground state.

collision or whatever, little in life is ideal, is it?) However, according to actual descriptions of NMR devices, it is better to stop the perturbation earlier, when the expectation spin vector has become horizontal, rather than fully down. In that case, nature will only find half the nuclei in the excited energy state after the perturbation, presumably decreasing the radiation yield by a factor 2. The classical explanation that is given is that when the (expectation) spin vector is precessing at the Larmor frequency in the horizontal plane, the radiation is most easily detected by the coils located in that same plane. And that closes this discussion.

Chapter 14

Nuclei [Unfinished Draft]

This chapter has not been finished. Since I think some parts of it are already of interest, like the figures, I am posting it as is. The reader beware, much of it has been poorly proofread, if at all. The figures should be fine.

So far, the focus in this book has been mostly on electrons. That is normal because electrons are important like nothing else for the physical properties of matter. Atomic nuclei appear in the story only as massive anchors for the electrons, holding onto the electrons with their positive electric charge. But then there is nuclear energy. Here the nuclei call the shots. Nuclei are discussed in this chapter.

The theory of nuclear structure is much less advanced than that of the electronic structure of atoms. Unlike the electromagnetic forces, the nuclear forces are very poorly understood. Examining them with well-understood electromagnetic probes is limited since nuclear forces are extremely strong, resisting manipulation. Accurate direct measurement of quantities of interest is usually not possible.

Nuclear physicists responded to that with a tidal wave of ingenious experiments, usually leveraging one accepted fact to deduce the next one (and at the same time check the old one). Much of this data is presented in this chapter in the form of overview figures. This is intended to allow you to understand the big picture.

Some important approximate quantum models have been developed by nuclear physicists to explain all that data. This chapter also tries to explain these models in relatively simple terms.

The first few sections of the chapter give an overview of key concepts important for understanding nuclei. It is highly recommended that you read these before reading any later sections in this chapter.

But first one word of caution about the figures. Most of their data has been carefully machine-read from standard nuclear data bases. However, the used data bases date from around the year 2003. So check any data you get from the figures for any more recent updates that may be available. Also note that various

figures that depend on relatively delicate mathematical analysis were machine produced too. Typically this was done using reasonable simplifications and/or a priori assumptions. Use such figures to understand the big picture, but do not pick individual data from them without checking it. A simple automated procedure processing about 3 000 different nuclei from some key data using an approximate model cannot compete with a nuclear specialist analyzing a single nucleus based on all the extensive knowledge that is available for that one nucleus.

14.1 Fundamental Concepts

This section describes the most basic facts about nuclei. These facts will be taken for granted in the rest of this chapter.

Nuclei consist of protons and neutrons. Protons and neutrons are therefore called "nucleons." Neutrons are electrically neutral, but protons are positively charged. In particular, the electric charge of a proton has the same magnitude as the charge of an electron, but has opposite sign. Since opposite charges attract, the protons in a nucleus attract electrons. Despite that, the electrons do not end up inside the nucleus. They have much larger quantum mechanical uncertainty in position than the much heavier nucleons. So the electrons form a "cloud" around the tiny nucleus, producing an atom.

Since charges of the same sign repel, protons mutually repel each other. That is due to the same electric "Coulomb" force that allows them to attract electrons. By itself, the Coulomb force between the protons in a nucleus would cause the nucleus to fly apart immediately. But nucleons, both protons and neutrons, also attract each other through another force, the "nuclear force." It is this force that keeps a nucleus together.

The nuclear force is very strong, which allows it to dominate electromagnetic forces like the repulsive Coulomb force in stable nuclei. But the nuclear force is also very short range, extending over no more than a few femtometers. (A femtometer, or fm, equals 10^{-15} m. It is sometimes called a fermi after famous nuclear physicist Enrico Fermi. While not approved by SI, Fermi was one of the good guys, so we should make allowances.) In big nuclei, nucleons are only held together to other nucleons in their immediate neighborhood by the nuclear force. But the protons are repulsed by other protons *everywhere* in the nucleus. If the nucleus gets too big, this repulsion becomes so big that the nucleus can no longer be stable. Lead, with 82 protons, is the heaviest element that can be stable, and then only if it contains a suitable number of neutrons to keep the protons somewhat apart.

The strength of the nuclear force is about the same regardless of the type of nucleons involved, protons or neutrons. That is called "charge independence."

More restrictively, but even more accurately, the nuclear force is the same if you swap the nucleon types. In other words, the nuclear force is the same if you replace all protons by neutrons and vice-versa. That is called "charge symmetry." For example, if you swap the nucleon type of a pair of protons, you get a pair of neutrons. Therefore the nuclear force between a pair of protons is very accurately the same as the one between a pair of neutrons, all else being equal. (The already mentioned Coulomb repulsion between the protons is additional and not the same.) But if you swap the nucleon type of a pair of protons, or of a pair of neutrons, you do not get a proton and a neutron. So the nuclear force between a proton and a neutron is less accurately the same as that between two protons or two neutrons.

The nuclear force is not a fundamental one. It is just an effect of the "color force" or "strong force" between the "quarks" of which protons and neutrons consist. That is why the nuclear force is also often called the "residual strong force." It is much like how the Van der Waals force between molecules is not a fundamental one; that force is a residual of the electromagnetic force between the electrons and nuclei of which molecules exist, {A.33}.

However, the theory of the color force,"quantum chromedynamics," is well beyond the scope of this book. It is also not really important for nanotechnology. In fact, it is not all that important for nuclear engineering either because the details of the theory are uncertain, and numerical solution is intractable, [19].

Despite the fact that the nuclear force is poorly understood, physicists can say some things with confidence. First of all,

Nuclei are normally in the ground state.

The "ground state" is the quantum state of lowest energy E. Nuclei can also be in "excited" states of higher energy. However, a bit of thermal energy is not going to excite a nucleus. Differences between nuclear energy levels are extremely large on a microscopic scale. That is why nuclear bombs and nuclear reactors can create so much energy. Still, nuclear reactions will typically leave nuclei in excited states. Usually such states decay back to the ground state very quickly. (In special cases, it may take forever.)

It should be noted that if a nuclear state is not stable, it implies that it has a very slight uncertainty in energy, compare chapter 7.4.1. This uncertainty in energy is commonly called the "width" Γ of the state. The discussion here will almost always ignore the uncertainty in energy.

A second general property of nuclei is:

Nuclear states have definite nuclear mass $m_{\rm N}$.

You may be surprised by this statement. It seems trivial. You would expect that the nuclear mass is simply the sum of the masses of the protons and neutrons that make up the nucleus. But Einstein's famous relation $E = mc^2$ relates

energy and mass. The nuclear mass is slightly *less* than the combined mass of the protons and neutrons from which it is made. The difference is the binding energy that keeps the nucleus together, expressed in mass units. (In other words, divided by the square speed of light c^2.) Sure, even for nuclear energies the changes in nuclear mass due to binding energy are tiny. But physicists can measure nuclear masses to very great accuracy. Different nuclear states have different binding energies. So they have slightly different nuclear masses.

(Similarly, a hydrogen atom has less mass than a free proton and a free electron. But here the difference, a few eV, is far too small to note. Since nuclear binding energies are millions of times bigger, in nuclei the effect is much more important.)

It may be noted that binding energies are almost never expressed in mass units in nuclear physics. Instead masses are expressed in energy units! And not in Joule either. The energy units used are almost invariably "electron volts" (eV). Never use an SI unit when talking to nuclear physicists. They will immediately know that you are one of those despised nonexperts. Just call it a "blah." In the unlikely case that they ask, tell them "That is what Fermi called it."

Next,

Nuclear states have definite nuclear spin $j_{\rm N}$.

Here the "nuclear spin" $j_{\rm N}$ is the quantum number of the net nuclear angular momentum. The magnitude of the net nuclear angular momentum itself is

$$J = \sqrt{j_{\rm N}(j_{\rm N}+1)}\hbar$$

Nuclei in excited energy states usually have different angular momentum than in the ground state.

The name nuclear "spin" may seem inappropriate since net nuclear angular momentum includes not just the spin of the nucleons but also their orbital angular momentum. But since nuclear energies are so large, in many cases nuclei act much like elementary particles do. Externally applied electromagnetic fields are not by far strong enough to break up the internal nuclear structure. And the angular momentum of an elementary particle is appropriately called spin. However, the fact that "nuclear spin" is two words and "azimuthal quantum number of the net nuclear angular momentum" is nine might conceivably also have something to do with the terminology.

According to quantum mechanics, $j_{\rm N}$ must be integer or half-integer. In particular, $j_{\rm N}$ must be an integer if the number of nucleons is even ($j_{\rm N} = 0$ or 1 or 2 or ...). If the number of nucleons is odd, $j_{\rm N}$ must be half an odd integer ($j_{\rm N} = \frac{1}{2}$ or $\frac{3}{2}$ or $\frac{5}{2}$ or ...).

The fact that nuclei have definite angular momentum does not depend on the details of the nuclear force. It is a consequence of the very fundamental

observation that empty space has no "build-in" preferred direction. That issue was explored in more detail in chapter 7.3.

(Many references use the symbol J also for j_N for that spicy extra bit of confusion. So one reference tells you that the eigenvalue [singular] of J^2 is $J(J+1)$, leaving the $\hbar^2$ away from conciseness. No kidding. One popular book uses I instead of j_N and reserves J for electronic angular momentum. At least this reference uses a bold face I to indicate the angular momentum itself, as a vector.)

Consider also the component J_z of the nuclear angular momentum in a selected z direction. According to quantum mechanics, this component can have the measurable values $-j_N$, $-j_N+1$, $-j_N+2$, ..., j_N-1, or j_N. Note also that if J_z has a definite value for nonzero j_N, then the components in orthogonal directions have uncertain values and are therefore not that interesting for analysis.

Finally,

Nuclear states have definite parity.

Here "parity" is what happens to the wave function when the nucleus is mirrored and then rotated 180° around the axis normal to the mirror, chapter 7.3. (Mathematically, this corresponds to inverting every $\vec{r}$ position vector measured from the center of gravity into $-\vec{r}$. Since the rotation is already covered by angular momentum, the important step is the mirroring.) The wave function can either stay the same, (called parity 1 or even parity), or it can change sign, (called parity -1 or odd parity). The fact that nuclei have definite parity too does not depend on the details of the nuclear force. It is a consequence of the fact that the forces of nature behave the same way when seen in the mirror.

Or actually, there is one force of nature, the still unmentioned so-called "weak force" that does *not* behave the same way when seen in the mirror. But the weak force is, like it says, weak. On nuclear scales, it is many orders of magnitude smaller than the nuclear and electromagnetic forces. So, while the weak force introduces some quantum-mechanical uncertainty in the parity of nuclei, this uncertainty is usually negligibly small. The chances of finding a nucleus in a given energy state with the "wrong" parity can be ballparked at 10^{-14}, [31, pp. 313ff]. That is almost always negligible. Only if, say, a nuclear process is strictly impossible solely because of parity, then the uncertainty in parity might give it a very slight possibility of occurring anyway.

Parity is commonly indicated by π because π is the Greek letter "p" and is not used for anything else in science. And physicists usually list the spin and parity of a nucleus together in the form J^π. If you have two quantities like spin and parity that have nothing to do with one another, what is better than show one as a superscript of the other? But do not start raising J to the power π! You should be translating this into common sense as follows:

$$J^\pi \quad \Rightarrow \quad j_N^\pm \quad \Rightarrow \quad j_N \text{ and } \pm$$

As a numerical example, 3^- means a nucleus with spin 3 and odd parity. It does not mean a nucleus with spin 1/3, (which is not even possible; spins can only be integer or half-integer.)

Key Points

- Nuclei form the centers of atoms.
- Nuclei consist of protons and neutrons. Therefore protons and neutrons are called nucleons.
- Protons and neutrons themselves consist of quarks. But for practical purposes, you may as well forget about that.
- Neutrons are electrically neutral. Protons are positively charged.
- Nucleons are held together by the so-called nuclear force.
- The nuclear force is approximately independent of whether the nucleons are protons or neutrons. That is called charge independence. Charge symmetry is a more accurate, but also more limited version of charge independence.
- Nuclear states, including the ground state, have definite nuclear energy E. The differences in energy between nuclear states are so large that they produce small but measurable differences in the nuclear mass $m_{\rm N}$.
- Nuclear states also have definite nuclear spin $j_{\rm N}$. Nuclear spin is the azimuthal quantum number of the net angular momentum of the nucleus. Many references indicate it by J or I.
- Nuclear states have definite parity π. At least they do if the so-called weak force is ignored.
- Never use an SI unit when talking to a nuclear physicist.

14.2 Draft: The Simplest Nuclei

This subsection introduces the simplest nuclei and their properties.

14.2.1 Draft: The proton

The simplest nucleus is the hydrogen one, just a single proton. It is trivial. Or at least it is if you ignore the fact that that proton really consists of a conglomerate of three quarks held together by gluons. A proton has an electric charge e that is the same as that of an electron but opposite in sign (positive). It has the same spin s as an electron, $\frac{1}{2}$. Spin is the quantum number of inherent angular

momentum, chapter 5.4. Also like an electron, a proton has a magnetic dipole moment μ. In other words, it acts as a little electromagnet.

However, the proton is roughly 2 000 times heavier than the electron. On the other hand the magnetic dipole moment of a proton is roughly 700 times smaller than that of an electron. The differences in mass and magnetic dipole moment are related, chapter 13.4. In terms of classical physics, a lighter particle circles around a lot faster for given angular momentum.

Actually, the proton has quite a large magnetic moment for its mass. The proton has the same spin and charge as the electron but is roughly 2 000 times heavier. So logically speaking the proton magnetic moment should be roughly 2 000 times smaller than the one of the electron, not 700 times. The explanation is that the electron is an elementary particle, but the proton is not. The proton consists of two up quarks, each with charge $\frac{2}{3}e$, and one down quark, with charge $-\frac{1}{3}e$. All three quarks have spin $\frac{1}{2}$. Since the quarks have significantly lower effective mass than the proton, they have correspondingly higher magnetic moments. Even though the spins of the quarks are not all aligned in the same direction, the resulting net magnetic moment is still unusually large for the proton net charge, mass, and spin.

Key Points

- The proton is the nucleus of a normal hydrogen atom.
- It really consists of three quarks, but ignore that.
- It has the opposite charge of an electron, positive.
- It has spin $\frac{1}{2}$.
- It is roughly 2 000 times heavier than an electron.
- It has a magnetic dipole moment. But this moment is roughly 700 times smaller than that of an electron.

14.2.2 Draft: The neutron

It is hard to call a lone neutron a nucleus, as it has no net charge to hold onto any electrons. In any case, it is somewhat academic, since a lone neutron disintegrates in on average about 10 minutes. The neutron emits an electron and an antineutrino and turns into a proton. That is an example of what is called "beta decay." Neutrons in nuclei can be stable.

A neutron is slightly heavier than a proton. It too has spin $\frac{1}{2}$. And despite the zero net charge, it has a magnetic dipole moment. The magnetic dipole moment of a neutron is about two thirds of that of a proton. It is in the direction opposite to the spin rather than parallel to it like for the proton.

The reason that the neutron has a dipole moment is that the three quarks that make up a neutron do have charge. A neutron contains one up quark with

a charge of $\frac{2}{3}e$ and two down quarks with a charge of $-\frac{1}{3}e$ each. That makes the net charge zero, but the magnetic dipole moment can be and is nonzero.

Key Points

- The neutron is slightly heavier than the proton.
- It too has spin ½.
- It has no charge.
- Despite that, it does have a comparable magnetic dipole moment.
- Lone neutrons are unstable. They suffer beta decay.

14.2.3 Draft: The deuteron

The smallest nontrivial nucleus consists of one proton and one neutron. This nucleus is called the deuteron. (An atom with such a nucleus is called deuterium). Just like the proton-electron hydrogen atom has been critical for deducing the structure of atoms, so the proton-neutron deuteron has been very important in deducing knowledge about the internal structure of nuclei.

However, the deuteron is not by far as simple a two-particle system as the hydrogen atom. It is also much harder to analyze. For the hydrogen atom, spectroscopic analysis of its excited quantum states provided a gold mine of information. Unfortunately, it turns out that the deuteron is so weakly bound that it has no excited quantum states. If you try to excite it by adding energy, it falls apart.

The experimental binding energy of the deuteron is only about 2.22 MeV. Here a MeV is the energy that an electron would pick up in a one-million voltage difference. For an electron, that would be a gigantic energy. But for a nucleus it is ho-hum indeed. A typical stable nucleus has a binding energy on the order of 8 MeV per nucleon.

In any case, it is lucky that that 2.22 MeV of binding energy is there at all. If the deuteron would not bind, life as we know it would not exist. The formation of nuclei heavier than hydrogen, including the carbon of life, begins with the deuteron.

The lack of excited states makes it hard to understand the deuteron. In addition, spin has a major effect on the force between the proton and neutron. In the hydrogen atom, that effect exists but it is extremely small. In particular, in the true hydrogen atom ground state the electron and proton align their spins in opposite directions. That produces the so-called singlet state of zero net spin, chapter 5.5.6. However, the electron and proton can also align their spins in the same direction, at least as far as angular momentum uncertainty allows. That produces the so-called triplet state of unit net spin. For the hydrogen atom, it

turns out that the triplet state has very slightly higher energy than the singlet state, {A.39}.

In case of the deuteron, however, the triplet state has the lowest energy. And the singlet state has so much more energy that it is not even bound. Almost bound maybe, but definitely not bound. For the proton and neutron to bind together at all, they *must* align their spins into the triplet state.

As a result, a nucleus consisting of two protons (the diproton) or of two neutrons (the dineutron) does not exist. That is despite the fact that two protons or two neutrons attract each other almost the same as the proton and neutron in the deuteron. The problem is the antisymmetrization requirement that two identical nucleons must satisfy, chapter 5.6. A spatial ground state should be symmetric. (See addendum {A.40} for more on that.) To satisfy the antisymmetrization requirement, the spin state of a diproton or dineutron must then be the antisymmetric singlet state. But only the triplet state is bound.

(You might guess that the diproton would also not exist because of the Coulomb repulsion between the two protons. But if you ballpark the Coulomb repulsion using the models of {A.41}, it is less than a third of the already small 2.22 MeV binding energy. In general, the Coulomb force is quite small for light nuclei.)

There is another qualitative difference between the hydrogen atom and the deuteron. The hydrogen atom has zero orbital angular momentum in its ground state. In particular, the quantum number of orbital angular momentum l equals zero. That makes the spatial structure of the atom spherically symmetric.

But orbital angular momentum is not conserved in the deuteron. In terms of classical physics, the forces between the proton and neutron are not exactly along the line connecting them. They deviate from the line based on the directions of the nucleon spins.

In terms of quantum mechanics, this gets phrased a bit differently. The potential does not commute with the orbital angular momentum operators. Therefore the ground state is not a state of definite orbital angular momentum. The angular momentum is still limited by the experimental observations that the deuteron has spin 1 and even parity. That restricts the orbital angular momentum quantum number l to the possible values 0 or 2, {A.40}. Various evidence shows that there is a quantum probability of about 95% that $l = 0$ and 5% that $l = 2$.

One consequence of the nonzero orbital angular momentum is that the magnetic dipole strength of the deuteron is not exactly what would be expected based on the dipole strengths of proton and neutron. Since the charged proton has orbital angular momentum, its acts like a little electromagnet not just because of its spin, but also because of its orbital motion.

Another consequence of the nonzero orbital angular momentum is that the charge distribution of the deuteron is not exactly spherically symmetric. This asymmetric charge distribution allows the deuteron to interact with gradients

in an external electric field. It is said that the deuteron has a nonzero "electric quadrupole moment."

Roughly speaking, you may think of the charge distribution of the deuteron as elongated in the direction of its spin. That is not quite right, quantum-mechanically speaking, since angular momentum has uncertainty in direction. Therefore, instead consider the quantum state in which the deuteron spin has its maximum component, $\hbar$, in the chosen z-direction. In that state, the charge distribution is elongated in the z-direction.

The nonzero orbital momentum also shows up in experiments where various particles are scattered off deuterons.

To be sure, the precise probability of the $l = 2$ state has never been established. However, assume that the deuteron is modeled as composed of a proton and a neutron. (Although in reality it is a system of 6 quarks.) And assume that the proton and neutron have the same properties as they have in free space. (That is almost certainly not a good assumption; compare the next section.) For such a model the $l = 2$ state needs to have about 4% probability to get the magnetic moment right. Similar values can be deduced from the quadrupole moment and scattering experiments, [31].

Key Points

- The deuteron consists of a proton and a neutron.
- The deuteron is the simplest nontrivial nucleus. The diproton and the dineutron do not exist.
- The deuteron has spin 1 and even parity. The binding energy is 2.225 MeV.
- There are no excited states. The ground state of lowest energy is all there is.
- The deuteron has a nonzero magnetic dipole moment.
- It also has a nonzero electric quadrupole moment.

14.2.4 Draft: Property summary

Table 14.1 gives a summary of the properties of the three simplest nuclei. The electron is also included for comparison.

The first data column gives the mass. Note that nuclei are thousands of times heavier than electrons. As far as the units are concerned, what is really listed is the energy equivalent of the masses. That means that the mass is multiplied by the square speed of light following the Einstein mass-energy relation. The resulting energies in Joules are then converted to MeV. An MeV is the energy that an electron picks up in a 1 million voltage difference. Yes it is crazy, but

	m MeV	q	$r_{\rm ch}$ fm	$j_{\rm N}$	π	μ $\mu_{\rm N}$	Q $e\,\mathrm{fm}^2$
electron	0.511	$-e$	0	$\frac{1}{2}$	$+1$	$-1\,838.282$	0
Proton	938.272	e	0.86	$\frac{1}{2}$	$+1$	2.793	0
Neutron	939.565	0	—	$\frac{1}{2}$	$+1$	-1.913	0
Deuteron	1 875.612	e	2.14	1	$+1$	0.857	0.286

$$\mathrm{fm} = 10^{-15}\ \mathrm{m} \qquad \mathrm{MeV} \approx 1.602\ 10^{-13}\ \mathrm{J}$$
$$e \approx 1.602\ 10^{-19}\ \mathrm{C} \qquad \mu_{\rm N} = \frac{e\hbar}{2m_{\rm p}} \approx 5.051\ 10^{-27}\ \mathrm{J/T}$$

Table 14.1: Properties of the electron and of the simplest nuclei.

that is how you will almost always find masses listed in nuclear references. So you may as well get used to it.

It can be verified from the given numbers that the deuteron mass is indeed smaller than the sum of proton and neutron masses by the 2.225 MeV of binding energy. It is a tenth of a percent, but it is very accurately measurable.

The second column gives the charge. Note that all these charges are whole multiples of the proton charge e. However, that is not a fundamental requirement of physics. In particular, "up" quarks have charge $\frac{2}{3}e$ while "down" quarks have charge $-\frac{1}{3}e$. The proton contains two up quarks and a down one, producing net charge e. The neutron contains one up quark and two down ones, producing zero net charge.

The third column gives the charge radius. That is a measure of the spatial extent of the charge distribution. The electron is, as far as is known, a point particle with no internal structure. For the neutron, with no net charge, it is not really clear what to define as charge radius.

The fourth column shows the quantum number of net angular momentum. For the first three particles, that is simply their spin. For the deuteron, it is the nuclear spin. That includes both the spins and the orbital angular momenta of the proton and neutron that make up the deuteron.

The fifth column is parity. It is even in all cases. More complicated nuclei can have negative parity.

The sixth column is the magnetic dipole moment. It is expressed in terms of the so-called nuclear magneton $\mu_{\rm N}$. A proton circling around with one quantum unit of orbital angular momentum has one nuclear magneton of magnetic moment due to its motion. (Which would be in addition to the intrinsic magnetic moment listed in the table. Note that magnetic moments are vectors like

angular momenta; they may cancel each other when summed.)

As the table shows, nuclei have much smaller magnetic moments than electrons. That is due to their much larger masses. However, using a magnetic field of just the right frequency, nuclear magnetic moments can be observed. That is how nuclear magnetic resonance works, chapter 13.6. Therefore nuclear magnetic moments are important for many applications, including medical ones like MRI.

The last column lists the electric quadrupole strength. That is a measure for the deviation of the nuclear charge distribution from a spherically symmetric shape. It is a complicating factor in nuclear magnetic resonance. Or an additional source of information, depending on your view point. Nuclei with spin less than 1 do not have electric quadrupole moments. (That is an implied consequence of the relation between angular momentum and symmetry in quantum mechanics.)

Note that the SI length unit of femtometer works very nicely for nuclei. So, since physicists hate perfection, they define a new non-SI unit called the barn b. A barn is 100 fm^2. So you will likely find the quadrupole moment of the deuteron listed as 0.00286 eb. Note the additional leading zeros. Some physicists do not like them, for good reason, and then use millibarn, giving 2.86 $e$mb. However, the quadrupole moments for many heavy nuclei are quite large in terms of millibarn. For example, einsteinium-253 has around 6 700 $e$mb. Anytime now, physicists are bound to figure out that centibarn works even better than millibarn. When that happens, let's all agree that we will not tell them that a centibarn is the same as that hated fm^2. The charge e is commonly left away from the definition of the quadrupole moment, giving it units of area.

Key Points

- The properties of the simplest nuclei are summarized in table 14.1.

14.3 Draft: Overview of Nuclei

This section introduces basic terminology and concepts of nuclei. It also gives an overview of the ways that they can decay.

The number of protons in a nucleus is called its "atomic number" Z. Since each proton has an electric charge e, equal to 1.602 18 10^{-19} C, the total nuclear charge is Ze. While protons attract nearby protons and neutrons in the nucleus with the short-range nuclear force, they also repel other protons by the long-range Coulomb force. This force too is very strong at nuclear distances. It makes nuclei with more than 82 protons unstable, because for such large nuclei the longer range of the Coulomb forces becomes a major factor.

The number of neutrons in a nucleus is its neutron number N. Neutrons have no charge, so they do not produce Coulomb repulsions. Therefore, the right amount of neutrons has a stabilizing effect on nuclei. However, too many neutrons is not stable either, because neutrons by themselves are unstable particles that fall apart in about 10 minutes. Combined with protons in a nucleus, neutrons can be stable.

Since neutrons have no charge, they also do not attract the electrons in the atom or molecule that the nucleus is in. Therefore only the atomic number Z is of much relevance for the chemical properties of an atom. It determines the position in the periodic table of chemistry, chapter 5.9. Nuclei with the same atomic number Z, so with the same place in the periodic table, are called "isotopes." (In Greek, "iso" means equal and "topos" place.)

However, the number of neutrons does have a secondary effect on the chemical properties, because it changes the mass of the nucleus. And the number of neutrons is of critical importance for the nuclear properties. Nuclei with the same number of neutrons are called "isotones." How clever, to replace the p in isotopes with an n.

The name of a nucleus indicates its number of protons Z; for example, "hydrogen" means $Z = 1$, "helium" $Z = 2$. To also indicate the number of neutrons, the convention is to follow the name by the "mass number, or "nucleon number" $A = N + Z$. It gives the total number of nucleons in the nucleus.

For example, the normal hydrogen nucleus, which consists of a lone proton, is hydrogen-1. The deuterium nucleus, which contains both a proton and a neutron, is hydrogen-2, indicating that it contains two nucleons total. Because it has the same charge as the normal hydrogen nucleus, a deuterium atom behaves chemically almost the same as a normal hydrogen atom. For example, you can create water with deuterium and oxygen just like you can with normal hydrogen and oxygen. Such water is called "heavy water." Don't drink it, however; the difference in chemical properties is still sufficient to upset biological systems. Trace amounts are harmless, as can be appreciated from the fact that deuterium occurs naturally. About 1 in 6 500 hydrogen nuclei in water on earth are deuterium ones.

The normal helium nucleus contains two protons plus two neutrons, so it is called helium-4. There is a stable isotope, helium-3, that has only one neutron. In the atmosphere, one in a million helium atoms has a helium-3 nucleus. While normally, there is no big difference between the two isotopes, at very low cryogenic temperatures they do behave very differently. The reason is that both protons and neutrons have spin ½, as do electrons, so a difference of one neutron switches the net atomic spin between half-integer (helium-3) and integer (helium-4). That makes the helium-3 atom a fermion but the helium-4 one a boson. At extremely low temperatures it makes a big difference in behavior, chapter 11.

In terms of symbols, it is conventional to precede the element symbol by

the mass number as a superscript and the atomic number as a subscript. So normal hydrogen-1 is indicated by ${}^{1}_{1}\mathrm{H}$, hydrogen-2 by ${}^{2}_{1}\mathrm{H}$, helium-3 by ${}^{3}_{2}\mathrm{He}$, and helium-4 by ${}^{4}_{2}\mathrm{He}$.

Nuclei with the same mass number A are called "isobars." Yes, this conflicts with the established usage of the word isobar for lines of constant pressure in meteorology, but in this case physicists have blown it. There is not likely to be any resulting confusion unless there is a nuclear winter.

Sometimes the element symbol is also followed by the number of neutrons as a subscript. However, that then raises the question whether H_2 stands for hydrogen-3 or a hydrogen molecule. The neutron number can readily by found by subtracting the atomic number from the mass number,

$$\boxed{N = A - Z} \tag{14.1}$$

so this book will leave it out.

It may also be noted that the atomic number is technically redundant, since the chemical symbol already implies the number of protons. It is often left away, because that confuses people who do not remember the atomic number of every chemical symbol in the periodic table. To create further confusion, deuterium is often indicated by chemical symbol D instead of H. It is hilarious to see people who have forgotten this search through a periodic table for element "D." For additional fun, the unstable hydrogen-3 nucleus, with one proton and two neutrons, is also called the "tritium" nucleus, or "triton," and indicated by T instead of ${}^{3}_{1}\mathrm{H}$. The helium-3 nucleus is also called the "helion." Fortunately for us all, helion starts with an h.

The nuclei mentioned above are just a tiny sample of the total of 256 nuclei that are stable and a much greater number still that are observed but unstable. It is conventional to represent both the stable and unstable nuclei in a "Chart Of the Nuclides," (CON), like the one shown in figure 14.1. In the CON, the tiny green squares are the stable nuclei. Squares of colors other than green represent unstable nuclei. The horizontal position of each square gives the number of neutrons N (selected "magic" values are listed along the horizontal axis). The vertical position gives the number of protons Z, (the crucial number that determines what chemical properties an atom with that nucleus has). Note that $Z = 82$ is lead, the last element with at least one stable nucleus. In fact, lead has four stable isotopes and naturally occuring lead atoms have a fair chance of having any one of the four as nucleus. The fact that lead is so unusually stable has a lot to do with the fact that lead is right on top of the $Z = 82$ magic line, and close to the $N = 126$ magic line. But more on that later.

Before continuing the discussion of the CON, a graphical problem must be addressed. While you cannot argue about taste, 99.9% of readers would surely agree that 14.1 is *(a)* ugly as hell, and *(b)* requires a magnifying glass to read. The "Chart Of the Nuclides" is well suited for printing out on two yards of paper and hanging on the wall of your office in its full glory. But in a book, it really

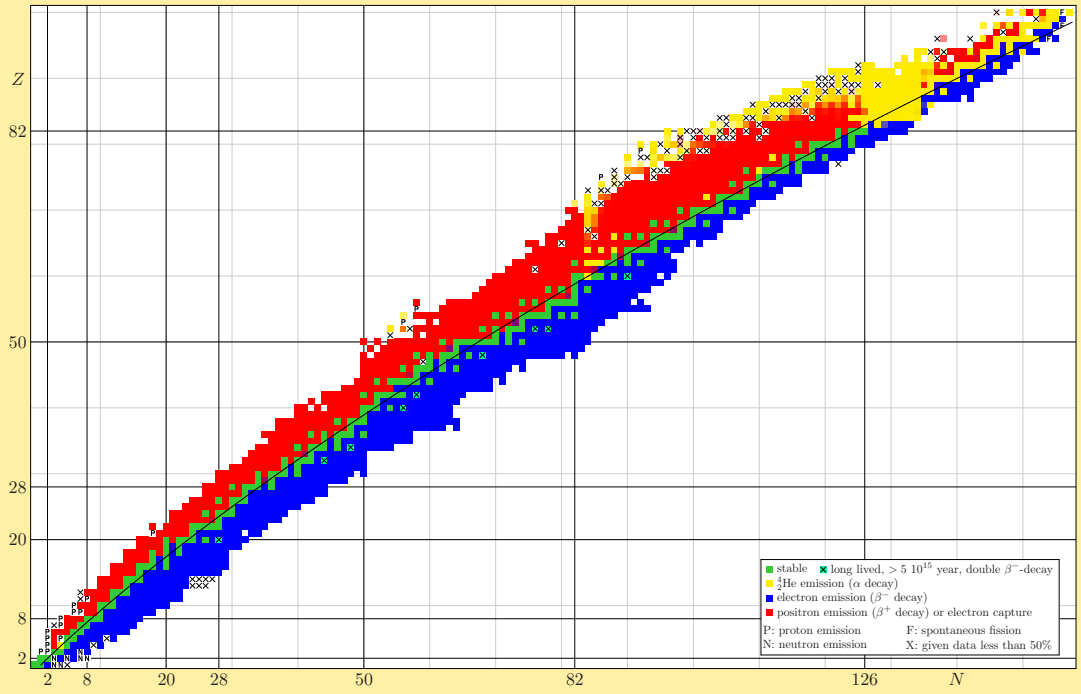


Figure 14.1: Chart of the nuclides.

does not work. It could be made slightly bigger if printed out sideways, but rotating a monitor with coffee cups on it and cables attached is a bit awkward. And a crick in your neck is not that great either.

Based on these considerations, from now on, this book will no longer plot the neutron number N along the horizontal axis, but the "neutron excess" $N - Z$. The neutron excess is how many more neutrons there are than protons. That is an important number, maybe even more important than the absolute number of neutrons. The corresponding modified chart of the nuclides is in figure 14.2. It will be called the RECON (Revised Chart of the Nuclei). It gives you something you can view in comfort.

But admittedly there are some disadvantages. In the RECON the isotones (the lines connecting nuclei with the same number of neutrons) are no longer vertical; now they slope down by 45°. That cannot be helped. Similarly the isobars, the lines connecting nuclei with the same total number of nucleons, no longer slope down by 45° like in the CON. In the RECON they slope down with the smaller slope 1/2: going down one square to the previous chemical element now requires that you go two squares to the right to stay on the same isobar.

So be it. The good news is that the neutron excess is a lot more relevant to nuclear stability than the absolute number of nucleons. For example, at low values of Z, the green band in RECON figure 14.2 is vertical, demonstrating very clearly that indeed for light nuclei, the number of neutrons must be about

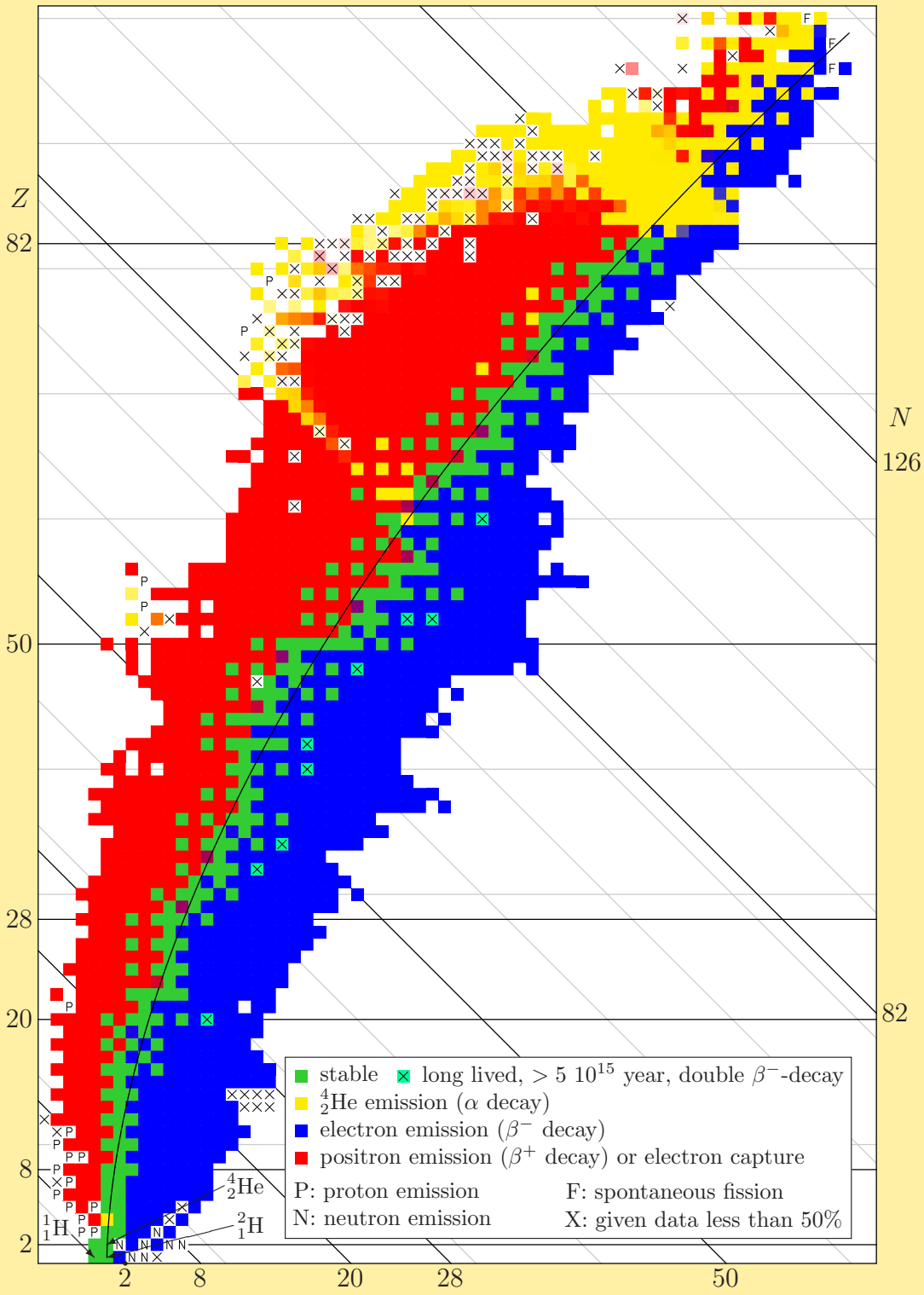


Figure 14.2: Nuclear decay modes. [pdf][con]

equal to the number of protons. Also for heavier nuclei, the RECON shows much more clearly exactly how much the relative number of neutrons goes up to mitigate the effect of the Coulomb repulsions between protons.

There is an other advantage to the RECON. It has to do with the fact that those nuclei in which both the number of protons and the number of neutrons is even, the "even-even" nuclei, are found to have enhanced stability. On the other hand those nuclei in which both the number of protons and the number of neutrons is odd, the "odd-odd" nuclei, are found to have reduced stability. Simply put, protons like to pair up, and so do neutrons.

It works out that in the RECON, the even-even and odd-odd nuclei end up on the same vertical lines. (These vertical lines alternate with vertical lines of even-odd and odd-even nuclei.) So wherever in figure 14.2 you see a vertical line with alternating green and non-green squares, well, the stable green squares are the even-even nuclei and the non-green ones in between the odd-odd ones. The pattern very convincingly demonstrates that indeed even-even nuclei are a lot more stable than odd-odd ones. (In the CON, the equivalent lines slant by 45° and are not by far as striking.)

In the intermediate vertical lines in the RECON, where you do not see such a periodic variation of stability, you find the even-odd and odd-even nuclei. Note that on these lines, the vertical extent of green squares is much less than on the adjacent lines with even-even nuclei. This demonstrates graphically that even-even nuclei are not just a lot more stable than odd-odd ones; they are also a lot more stable than even-odd and odd-even ones.

All this also makes it easy to figure out whether a given nucleus is is even-even or odd-odd in the RECON. Look whether the vertical line it is on has a series of alternating green and non-green squares; if so, then that is a line of even-even and odd-odd nuclei. The adjacent two lines then contain even-odd and odd-even nuclei. (In the region of heaviest nuclei, you can typically look at the yellow "alpha-decay" nuclei as a substitute for the green nuclei.) Alternatively, if you see two green squares immediately above each other in the RECON above $Z = 8$, then that vertical line consists of even-odd and odd-even nuclei; there are no stable odd-odd nuclei above $Z = 7$. (Conversely, the RECON illustrates that quite clearly too.)

Note also that the mass number A is odd on the even-odd, odd-even vertical lines. And A is even on the even-even, odd-odd vertical lines. While the mass number by iself does not have that much physical meaning, nuclear physicists often use "odd mass number nuclei" as a shorthand for "even-odd and odd-even nuclei."

If you are really a CON man or woman, there is nothing wrong with that. You can always click on the [con] link provided in the legend of the figure to load the figure in CON format as a separate pdf file. Conversely, if you really like the RECON format and you want to print it and hang it on your wall, click on the [pdf] link instead for a printable version. And either type of pdf can be

readily magnified to see details more clearly. [On linux I have to set preferences in buggy Adobe Acrobat reader to always open links before it works.]

Let's look at some of the details of RECON figure 14.2. The leftmost green square in the bottom row ($Z = 1$) is the hydrogen-1 nucleus, and the green square immediately to the right of it is hydrogen-2, deuterium. The green squares on the second-lowest row ($Z = 2$) are helium-3 and helium-4 respectively.

Like in the CON, isotopes are found on the same horizontal line in the RECON. As mentioned, the horizontal position of each square in RECON figure 14.2 indicates the neutron excess. For example, hydrogen-2 and helium-4 both have equal numbers of protons and neutrons. So they are at the same horizontal position, zero, in the figure. Similarly, hydrogen-1 and helium-3 both have a neutron excess of minus one. The figure shows that stable light nuclei have about the same number of neutrons as protons. However, for the heaviest nuclei, there are about 50% more neutrons than protons. For heavy nuclei, too many closely packed protons would mean too much Coulomb repulsion.

Many isotopes are unstable and decay spontaneously, liberating energy. For example, consider the blue square to the right of ${}^{2}_{1}\mathrm{H}$ in figure 14.2. That is ${}^{3}_{1}\mathrm{H}$, hydrogen-3 or tritium. It is unstable. After on average about twelve years, it will turn into helium-3. In particular, one of the two neutrons turns into a positively charged proton. So there are still three nucleons, the mass number has stayed the same, but the atomic number has increased one unit. In terms of RECON figure 14.2, the nucleus has changed into one that is one place up and two places to the left.

Since charge is conserved, the creation of the positive charge can only happen if the neutron emits a compensating negative charge; the neutron does so by emitting an electron. For historical reasons, a decay process of this type is called "beta decay" (β-decay) instead of "electron emission;" initially it was not recognized that the observed radiation was merely high energy electrons. And the name could not be changed later, because that would add clarity. (An antineutrino is also emitted, but it is almost impossible to detect: solar neutrinos will readily travel all the way through the earth with only a miniscule chance of being captured.)

Nuclei with too many neutrons tend to use beta decay to turn the excess into protons in order to become stable. Figure 14.2 shows nuclei that suffer beta decay in blue. Since in the decay process they move towards the left, they move towards the stable green area. Although not shown in the figure, a lone neutron also suffers beta decay after about 10 minutes and so turns into a proton.

If nuclei have too many protons rather than too many neutrons, they can turn their excess protons into neutrons by emitting a positron. The positron, the anti-particle of the electron, carries away one unit of positive charge, turning a positively charged proton into a neutral neutron.

However, a nucleus has a much easier way to get rid of one unit of net

positive charge: it can swipe a negatively charged electron from the atom it is in. This is called "electron capture" (EC). An electron neutrino is emitted in this process.

Electron capture is also called K-capture of L-capture, depending on the electron shell from which the electron is swiped. It is also referred to as "inverse beta decay," especially within the context of "neutron stars." These stars are so massive that their atoms collapse under gravity and the electrons and protons combine into neutrons. These stars then emit enormous amounts of high-energy neutrinos, taking along a large amount of the available energy of the star.

Of course, "inverse beta decay" is not really inverse beta decay, because in beta decay the emitted electron does not go into an empty atomic orbit, and in beta decay no neutrino is absorbed; instead an antineutrino is emitted.

Positron emission is also often called "beta-plus decay" (β^+-decay). After all, if you do have obsolete terminology, it is fun to use it to the fullest. Note that NUBASE 2003 uses the term beta-plus decay to indicate either positron emission or electron capture. In analogy with the beta-plus terminology, electron emission is also commonly called beta-minus decay or negatron emission. Some physicists leave the "r" away to save trees and talk about positons and negatons.

The nuclei that suffer beta-plus decay or electron capture are shown as red squares in figure 14.2. In the decay, a proton turns into a neutron, so the nucleus moves one place down and two places towards the right. That means that these nuclei too move towards the stable green area.

There are a variety of other ways in which nuclei may decay. If the number of protons or neutrons is really excessive, the nucleus may just kick out one of the bums instead of convert it. Nuclei that do that are marked with "P," respectively "N" in figure 14.2,

Similarly, heavy nuclei that are weakened by Coulomb repulsions tend to just throw some nucleons out. Commonly, a ${}^4_2$He helium-4 nucleus is emitted, as this is a very stable nucleus that does not require much energy to create. Such an emission is called "alpha decay" (α-decay) because helium-4 emission would be easily understandable. Alpha decay reduces the mass number A by 4 and the atomic number Z by 2. The nucleus moves two places straight down in RECON figure 14.2.

If nuclei are really oversized, they may just fall apart completely; that is called spontaneous fission.

Another process, "gamma decay," is not shown in figure 14.2. In gamma decay, an *excited* nucleus transitions to a lower energy state and emits the released energy as very energetic electromagnetic radiation. This is much like the spontaneous decay of excited electron levels in atoms, which too releases electromagnetic radiation. However, the electromagnetic radiation emitted in gamma decay is much more powerful than that emitted by atomic electrons, as nuclear energies are so much higher than those of atomic electrons. Unlike

the decays shown in figure 14.2, in gamma decay the type of nucleus does not change; there is no change in the number of protons nor neutrons.

Unlike gamma decay, the nuclear decays shown in figure 14.2 are from their their non-excited "ground state." But the shown decays are commonly associated with *additional* gamma radiation, since the decay tends to leave the changed nucleus in an excited state.

Gamma decay as a separate process, not directly caused by another process, is often referred to as an "isomeric transition" (IT) or "internal transition." In nuclear physics, an isomer is a long-lived excited state of a nucleus.

Besides gamma decay, a second way that an excited nucleus can get rid of excess energy is by throwing an electron from the atomic electron cloud surrounding the nucleus out of the atom. You or I would probably call that something like electron ejection. But what better name for throwing an electron, that is already outside the nucleus to start with, completely out of the *atom* than "*internal* conversion" (IC)? It can produce some of that hilarious confusion with the similar sounding term "internal transition." Internal conversion is usually included in the term "isomeric transition."

Figure 14.2 mixes colors if more than one decay mode occurs for a nucleus. The dominant decay is often immediately followed by another decay process. The subsequent decay is not shown. Data are from NUBASE 2003, without any later updates. The blank square right at the stable region is silver-106, and has a half-life of 24 minutes. Other sources list it as decaying through the expected electron capture or positron emission. But NUBASE 2003 lists that contribution as unknown and only mentions that beta-minus decay is negligible.

RE	RaA	RaB	RaC	RaC1	RaC2	RaD	RaE	RaF
$^{222}_{86}$Rn	$^{218}_{84}$Po	$^{214}_{82}$Pb	$^{214}_{83}$Bi	$^{214}_{84}$Po	$^{210}_{81}$Tl	$^{210}_{82}$Pb	$^{210}_{83}$Bi	$^{210}_{84}$Po

Table 14.2: Alternate names for nuclei.

Since so many outsiders know what nuclear symbols mean, physicists prefer to use obsolete names to confuse them. Table 14.2 has a list of names used. The abbreviations refer to historical names for decay products of radium (radium emanation, radium A, etc.)

Key Points

- Nuclei consist of protons and neutrons held together by the nuclear force.
- Protons and neutrons are collectively referred to as nucleons.
- Protons also repel each other by the Coulomb force.

- The number of protons in a nucleus is the atomic number Z. The number of neutrons is the neutron number N. The total number of nucleons $Z + N$ is the mass number or nucleon number A.
- Nuclei with the same number of protons Z correspond to atoms with the same place in the periodic table of chemistry. Therefore nuclei with the same atomic number are called isotopes.
- To promote confusion, nuclei with the same number of neutrons N are called isotones, and nuclei with the same total number of nucleons A are called isobars.
- For an example nuclear symbol, consider ${}^{4}_{2}\mathrm{He}$. It indicates a helium atom nucleus consisting of $A = 4$ nucleons, the left superscript, of which $Z = 2$ are protons, the left subscript. Since it would not be helium if it did not have 2 protons, that subscript is often left away. If you do not remember Z for, say, ${}^{146}\mathrm{Pm}$, you can look it up in a periodic table, like 5.8. But avoid doing so with "elements" D and T.
- Since these rules are too simple, physicists often drag up obsolete symbols like "RE" and "RaF" from the dark history of nuclear physics. You can look these up in a table above.
- The name for the nucleus with symbol ${}^{4}_{2}\mathrm{He}$ is helium-4, where the 4 is again the number of nucleons A.
- An odd mass number A corresponds to either an even-odd nucleus, a nucleus in which the number of protons is even and the number of nucleons odd, or to an odd-even nucleus, in which it is the other way around. An even mass number A corresponds to either an even-even nucleus, which tends to have relatively high stability, or to an odd-odd nucleus, which tends to have relatively low stability. The vertical columns in a RECON plot correspond alternatingly to odd and even mass numbers A. The two types of columns look very different.
- Nuclei can decay by various mechanisms. To promote confusion, emission of a helium-4 nucleus is called alpha decay or *alpha* decay. Emission of an electron is called beta decay, or β decay, or beta-minus decay, or β^- decay, or negatron emission, or negaton emission, but *never* electron emission. To do the latter would be severely frowned upon by physicists. Emission of a positron (positon) may be called beta-plus decay, or β^+ decay, but either term might be used to also indicate electron capture (EC), depending on who uses the term. Electron capture may also be called K-capture or L-capture or even inverse beta decay, though it is not. More extreme decay mechanisms are proton or neutron emission, and spontaneous fission. Kicking an electron in the electron cloud outside the nucleus completely free of the atom is called *internal* conversion. Mere emission of electromagnetic radiation is called gamma decay or γ decay.

🗝 No, this is not a story made up by this book to put physicists in a bad light.

14.4 Draft: Magic numbers

In nuclear physics, there are certain special values for the number of protons or the number of neutrons that keep popping up. Those are the values shown by horizontal and diagonal lines in the decay plot figure 14.2:

$$\text{magic numbers:} \quad 2, 8, 20, 28, 50, 82, 126, \ldots \tag{14.2}$$

These "magic" numbers were historically found to be associated with unusual stability properties. For example, the magic number of 82 neutrons occurs in 7 stable nuclei, more stable nuclei than for any other number of neutrons. The runners-up are 20 and 50 neutrons, also both magic numbers, that each occur in 5 stable nuclei.

Nuclei that have a magic number of protons also tend to have unusual stability. For example, the element with the most stable isotopes is tin, with 10 of them. Tin has $Z = 50$ protons, a magic number. To be sure, the runner up, Xenon with nine stable isotopes, has $Z = 54$, not a magic number, but the heaviest of these nine stable isotopes has a magic number of neutrons.

The last element to have any stable isotopes at all is lead, and its number of protons $Z = 82$ is magic. The lead isotope ${}^{208}_{82}\text{Pb}$, with 82 protons and 126 neutrons, is doubly magic, and it shows. It holds the triple records of being the heaviest nucleus that is stable, the heaviest element that is stable, and the highest number of neutrons that is stable.

The doubly magic ${}^{4}_{2}\text{He}_2$ nucleus, the alpha particle, is stable enough to be emitted in alpha decays of other nuclei.

Nuclei with magic numbers also have unusually great isotopic presence on earth as well as cosmic abundance. The reason for the magic numbers will eventually be explained through a simple quantum model for nuclei called the "shell model." Their importance will further be apparent throughout the figures in this chapter.

14.5 Draft: Radioactivity

Nuclear decay is governed by chance. It is impossible to tell exactly when any specific nucleus will decay. Therefore, the decay is phrased in terms of statistical quantities like specific decay rate, lifetime and half-life. This section explains how these are defined.

14.5.1 Draft: Half-life and decay rate

As a generic example of an unstable nucleus, consider tritium, an isotope of hydrogen. The ${}^{3}_{1}\mathrm{H}$ tritium nucleus, the triton, consists of one proton and two neutrons. The triton suffers beta decay. Eventually it will eject an electron and an antineutrino. This turns one of the two neutrons into a proton. So the decay turns the triton into the ${}^{3}_{2}\mathrm{He}$ helium nucleus isotope called the "helion." The original triton is lost.

Unlike the "normal" ${}^{4}_{2}\mathrm{He}$ helium nucleus, the helion contains only one neutron. However, it is stable; there will not be any further spontaneous decays.

(Note that the triton decays even though it has two neutrons, a magic number. But the helion it decays to has two protons, also magic. And just like a lone neutron decays into a less heavy proton, two neutrons in the tiny triton is just too much of a neutron excess compared to the negligible additional Coulomb repulsion in the helion.)

The big question in this section is, *when* will an unstable nucleus like the triton decay? Unfortunately, there is no complete anwer to that question. A given triton might last for 10 years, or it might last for 20 years or whatever. It could last less than a year, though that is not very likely. It could last for 100 years, even though that is much less likely still. But there is no way to tell for sure.

However, suppose you take a very large number of tritons and record for each how long the triton lives. Then you can average all these times and you get a number that is called the "lifetime" τ of the triton. The correct term would be *expected* or *average* lifetime, but this is physics, remember. Correct terms are not allowed. (To be fair, some physicists do use the proper term "mean lifetime" instead of just lifetime.)

If you average over enough tritons, you will find this mean lifetime of tritons to be almost 18 years. But not a single triton will decay after exactly 18 years. It is much like the expected lifetime of a newborn baby in the USA is, say, 80 years. Despite that, almost no one dies on their 80th birthday. Some die at birth or as kids.

Still, there is one big difference between people and nuclei. If you have a person who is 80 years old, surely you do not expect them to live until they become 160 years old. But if you have a bunch of tritons, on average this bunch of tritons will last for another 18 years. That is regardless of how long these tritons have survived already when you start observing them. Nuclear decay is a completely random process that occurs "out of the blue;" it does not depend on any previous history of the nucleus.

There is another issue. Unless you are an accountant by calling, why would you want to sit down, measure lifetimes of nuclei, and average them? What is the use?

A physically much more relevant scenario is that you have managed to create

a large number of tritons, and you would like to know how long they will last for doing experiments. In particular, you might want to know how long it takes before half of the tritons you created with blood, tears, and tax-payer money, are gone. This physically more meaningful time period is called the "half-life" $\tau_{1/2}$. It is related to the mean lifetime by a simple factor $\ln 2$:

$$\boxed{\tau_{1/2} = \tau \ln 2} \tag{14.3}$$

Note that $\ln 2$ is less than one. Half-life is somewhat shorter than mean lifetime.

The half-life of the triton is 12.32 years. So if you initially have a large collection of tritium nuclei, after 12.32 years only half will be left. After 24.64 years, only a quarter will remain, and after a century only 0.4%. After a millennium only 4 10^{-23}% will remain. (Since a gram of tritium contains about 2 10^{23} atoms, after a millenium, not a *single atom* would be left of a gram of tritium, if you managed to create that many of them!)

Now the earth is over 4 million millenia old. So you will appreciate that almost *none* of the tritium ever present on earth still exists. Some new tritium is continuously being created in the atmosphere by high-energy cosmic rays, but because of the geologically short half-life, there is no measurable accumulation. The total amount of tritium present on earth is virtually zero.

As in an earlier section, figure 14.3 shows again the decay processes of the nuclei. But unlike in the earlier figure 14.2, this time the square size of each nucleus has been adjusted to illustrate its half-life.

For the full size squares in the figure, the half-life is 10^{18} seconds or longer. Now 10^{18} seconds is about twice the estimated age of the universe since the Big Bang. So for nuclei that really have full size squares in figure 14.3, most of these nuclei that the universe ever created is likely to be still around.

On the other hand, if the square size of a nucleus is even *slightly* smaller than full size, then most of these nuclei that the universe ever created will have decayed. You may note that the square size of the ${}^{3}_{1}\mathrm{H}$ triton is noticably smaller than full size.

On the other hand, all stable green nuclei have full size squares. You might also note bismuth-209, ${}^{209}_{83}\mathrm{Bi}$. While not actually stable, for all practical puposes it is. Its half-life of 6 10^{26} seconds exceeds the age of the unverse by a factor of over a billion. In fact, it took physicists until 2003 to observe that bismuth-209 did actually suffer alpha decay; until then it was believed stable. (Note that bismuth 209 has 126 neutrons, a magic number.) The various double-beta-decay light green nuclei live even longer, on the order of 10^{30} seconds.

Based on various arguments, it was decided to take the minimum half-life shown in figure 14.3 to be one nanosecond. Clearly the figure needs *some* lower limit. And a nanosecond is really fast for alpha decay, and much faster than any beta decay. (Note that you do not see any really small red or blue squares in figure 14.3, and only a few yellow ones.) And for gamma decay, a nanosecond is often used as a cut-off between "prompt" and "isomeric" decay.

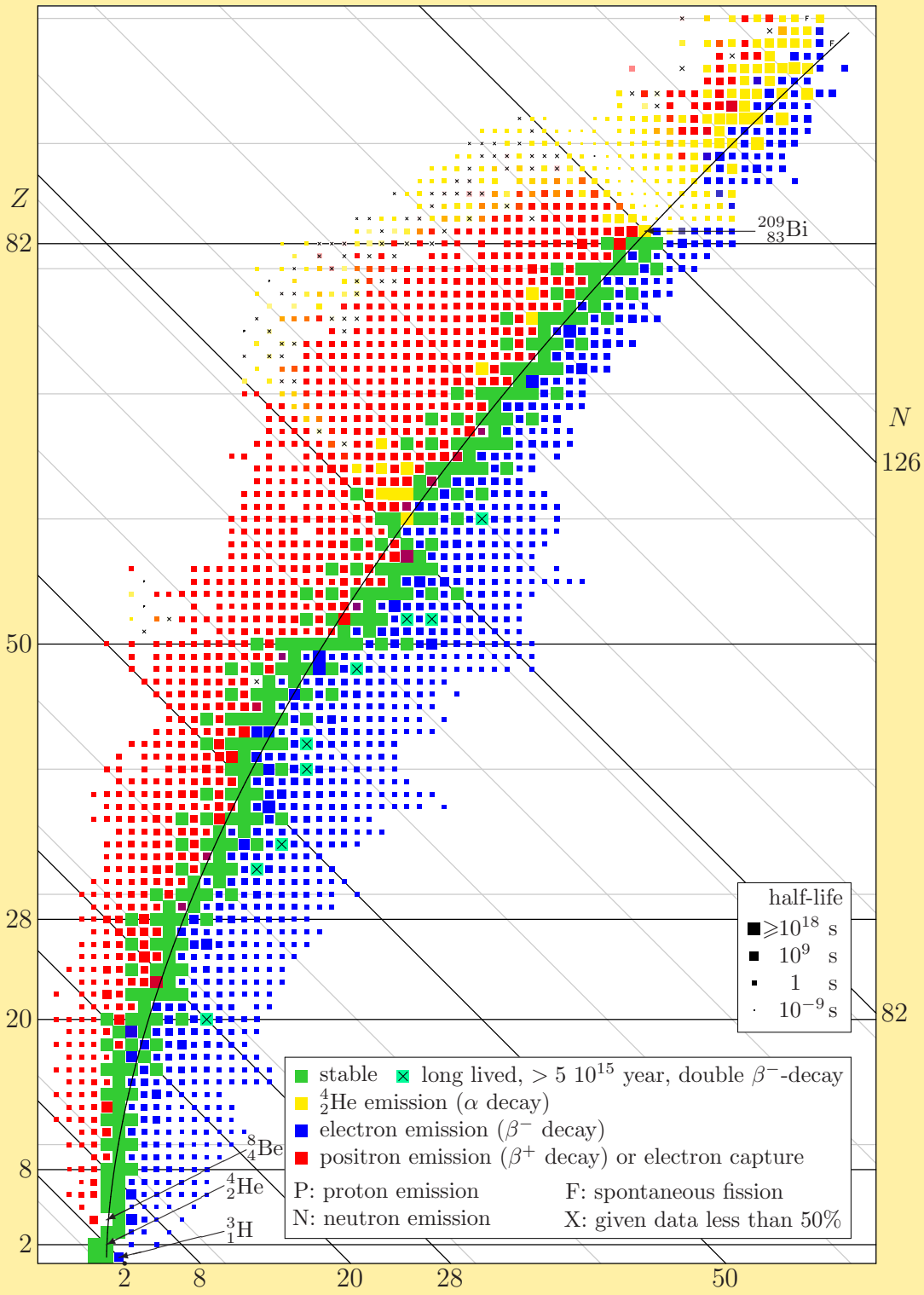


Figure 14.3: Nuclear half-lifes. [pdf][con]

Still, some nuclear decay processes are much quicker than a nanosecond. For example, you might note that the light nuclei that decay though proton or neutron emission in figure 14.2 have disappeared in figure 14.3. Such a decay process may have a half life on the order of 100 ys, (i.e. 100 10^{-24} seconds). That is much faster than the normal decays. In fact, in terms of quasi-classical physics, these decay times are comparable to the time for a nucleon with a speed of say a tenth of that of light to "move" just once through a nucleus with a size of the order of femtometers.

Another notable nucleus that has disappeared is ${}^{8}_{4}\mathrm{Be}$, beryllium-8. Berylium-8 falls apart in two ${}^{4}_{2}\mathrm{He}$, helium-4, alpha particles. The half-life of that process is just 67 as (i.e. 67 10^{-18} seconds). While it is technically called alpha-decay because an alpha particle is emitted, it is physically very different from the normal alpha decay of heavy nuclei. In the alpha decay of heavy nuclei, a heavy nucleus is left, not a second alpha particle. In particular, the berylium-8 decay is many orders of magnitude faster than the normal alpha decays discussed in section 14.11.

Besides (mean) lifetime and half-life, there is one more related term that is commonly used in describing nuclear decays. It is called the "specific decay rate" λ. The specific decay rate is the relative amount of nuclei in a large sample that is lost per unit time. Mathematically, the specific decay rate is just the inverse of the mean lifetime τ:

$$\boxed{\lambda \equiv \frac{1}{\tau}} \tag{14.4}$$

To better understand the various variables mathematically, it may be worthwhile to see how the mentioned relationships between them arise. First, according to the very definition of the decay rate λ above, if the current number of nuclei is I, then the number of nuclei that are lost, call it $-\mathrm{d}I$, in an infinitesimally small time interval $\mathrm{d}t$ is given by

$$-\mathrm{d}I = \lambda I \,\mathrm{d}t \tag{14.5}$$

This can be integrated after moving the I to the left-hand side. The result shows that if the amount of nuclei at time zero is I_0, then at an arbitrary later time t the amount of remaining nuclei I is

$$I = I_0 e^{-\lambda t} \tag{14.6}$$

(To check this expression, just differentiate it.) To find the half-life, you can set $I = \frac{1}{2} I_0$and $t = \tau_{1/2}$; that shows that $\tau_{1/2}$ must be $\ln 2/\lambda$. Also, from the expression above, you can compute the (average) lifetime as

$$\tau = \int_0^\infty t \frac{-\mathrm{d}I}{\mathrm{d}t}\,\mathrm{d}t \Big/ \int_0^\infty \frac{-\mathrm{d}I}{\mathrm{d}t}\,\mathrm{d}t$$

giving $\tau = 1/\lambda$. That then gives $\tau_{1/2} = \ln 2 \tau$.

14.5.2 Draft: More than one decay process

One very important point must be emphasized. Many nuclei undergo more than one decay process. In that case, each decay process has its own decay rate, independent of the other decay processes. In such cases,

Always add specific decay rates, never lifetimes or half-lifes.

The sum of the specific decay rates gives the total specific decay rate of the nucleus. The reciprocal of that total is the actual lifetime. Multiply by $\ln 2$ to get the actual half-life.

14.5.3 Draft: Other definitions

You probably think that having three different names, the specific decay rate λ, the lifetime τ, and the half-life $\tau_{1/2}$, for essentially the same physical quantity, is no good. You want more! Physicists are only too happy to oblige. How about using the term "decay constant" instead of specific decay rate? Its redeeming feature is that "constant" is a much more vague term, maximizing confusion. Even better, how does "disintegration constant" sound? Especially since the nucleus clearly does not disintegrate in decays other than spontaneous fission? Why not call it "specific activity,"come to think of it? Activity is another of these vague terms that the hated nonspecialists cannot make heads or tails of.

How about calling the product λI the "decay rate" or"disintegration rate" or simply the "activity?"

You probably want some units to go with that! What is more logical than to take the decay rate or activity to be in units of "curie," with symbol Ci and equal 3.7 10^{10} decays per second. (Of course you guessed that straight away. If you add 3 and 7 you get 10, not?) There is also the "becquerel," Bq, equal to 1 decay per second, defined but almost never used. Why not "dpm," disintegrations per minute, come to think of it? Why not indeed. The minute is just the additional unit the SI system needs, and using an acronym is great for creating confusion.

Of course the "activity" only tells you the amount of decays, not how bad the generated radiation is for your health. The "exposure" is the ionization produced by the radiation in a given mass of air, in SI units of Coulomb per kg. Exposure is very important for all people made of air. Of course, a better unit than a blasted SI one is needed, so the "roentgen" or "röntgen" R is defined to 2.58 10^{-4} C/kg. Why not?

But health-wise you may be more interested in the "absorbed dose" or "total ionizing dose" or "TID." That is the radiation energy absorbed per unit mass. That would be in J/kg or "gray," Gy, in SI units, but people really use the "rad" which is one hundredth of a gray.

If an organ or tissue absorbs a given dose of radiation, it is likely to be a lot worse if all that radiation is concentrated near the surface than if it is spread

out. The "quality factor" Q or the somewhat differently defined "radiation weighting factor" $w_{\rm R}$ is designed to correct for that fact. X-rays, beta rays, and gamma rays have radiation weighting factors (quality factors) of 1, but energetic neutrons, alpha rays and heavier nuclei go up to 20. Higher quality means worse for your health. Of course you already guessed that.

The bad effects of the radiation on your health are taken to be approximately given by the "equivalent dose," equal to the average absorbed dose of the organ or tissue times the radiation weighting factor. It is in SI units of J/kg, called the "sievert" Sv, but people really use the "rem," equal to one hundredth of a sievert. Note that the units of dose and equivalent dose are equal; the name is just a way to indicate what quantity you are talking about. It works if you can remember all these names.

To get the "effective dose" for your complete body, the equivalent doses for the organs and tissues must still be multiplied by "tissue weighting factors and summed. The weighting factors add up to one when summed over all the parts of your body. The ICRP defines "dose equivalent" different from equivalent dose. Dose equivalent is used on an operational basis. The personal dose equivalent is defined as the product of the dose at a point at an appropriate depth in tissue, (usually below the point where the dosimeter is worn), times the quality factor (not the radiation weighting factor).

14.6 Draft: Mass and energy

Nuclear masses are not what you would naively expect. For example, since the deuterium nucleus consists of one proton and one neutron, you might assume its mass is the sum of that of a proton and a neutron. It is not. It is less.

This weird effect is a consequence of Einstein's famous relation $E = mc^2$, in which E is energy, m mass, and c the speed of light, chapter 1.1.2. When the proton and neutron combine in the deuterium nucleus, they lower their total energy by the binding energy that keeps the two together. According to Einstein's relation, that means that the mass goes down by the binding energy divided by c^2. In general, for a nucleus with Z protons and N neutrons,

$$\boxed{m_{\rm nucleus} = Zm_{\rm p} + Nm_{\rm n} - \frac{E_{\rm B}}{c^2}} \tag{14.7}$$

where

$$m_{\rm p} = 1.672\,621\ 10^{-27}\ {\rm kg} \qquad m_{\rm n} = 1.674\,927\ 10^{-27}\ {\rm kg}$$

are the mass of a lone proton respectively a lone neutron at rest, and $E_{\rm B}$ is the binding energy. This result is very important for nuclear physics, because mass is something that can readily be measured. Measure the mass accurately and you know the binding energy.

In fact, even a normal hydrogen atom has a mass lower than that of a proton and electron by the 12.6 eV (electron volt) binding energy between proton and electron. But scaled down by c^2, the associated change in mass is negligible.

In contrast, nuclear binding energies are on the scale of MeV instead of eV, a million times higher. It is the devastating difference between a nuclear bomb and a stick of dynamite. Or between the almost limitless power than can be obtained from peaceful nuclear reactors and the limited supply of fossil fuels.

At nuclear energy levels the changes in mass become noticeable. For example, deuterium has a binding energy of 2.224 5 MeV. The proton has a rest mass that is equivalent to 938.272 013 MeV in energy, and the neutron 939.565 561 MeV. (You see how accurately physicists can measure masses.) Therefore the mass of the deuteron nucleus is lower than the combined mass of a proton and a neutron by about 0.1%. It is not big, but observable. Physicists are able to measure masses of reasonably stable nuclei extremely accurately by ionizing the atoms and then sending them through a magnetic field in a mass spectrograph or mass spectrometer. And the masses of unstable isotopes can be inferred from the end products of nuclear reactions involving them.

As the above discussion illustrates, in nuclear physics masses are often expressed in terms of their equivalent energy in MeV instead of in kg. To add further confusion and need for conversion factors, still another unit is commonly used in nuclear physics and chemistry. That is the "unified atomic mass unit" (u), also called "Dalton," (Da) or "universal mass unit" to maximize confusion. The "atomic mass unit" (amu) is an older virtually identical unit, or rather two virtually identical units, since physicists and chemists used different versions of it in order to achieve that supreme perfection in confusion.

These units are chosen so that atomic or nuclear masses expressed in terms them are approximately equal to the number of nucleons, (within a percent or so.) The current official definition is that a carbon-12, ${}^{12}_{6}\mathrm{C}$, atom has a mass of exactly 12 u. That makes 1 u equivalent 931.494 028 MeV. That is somewhat less than the mass of a free proton or a neutron.

One final warning about nuclear masses is in order. Almost always, it is atomic mass that is reported instead of nuclear mass. To get the nuclear mass, the rest mass of the electrons must be subtracted, and a couple of additional correction terms applied to compensate for their binding energy, [37]:

$$\boxed{m_{\text{nucleus}} = m_{\text{atom}} - Z m_{\text{e}} + A_e Z^{2.39} + B_e Z^{5.35}} \tag{14.8}$$

$$m_{\text{e}} = 0.510\,998\,91 \text{ MeV} \quad A_e = 1.443\,81\ 10^{-5} \text{ MeV} \quad B_e = 1.554\,68\ 10^{-12} \text{ MeV}$$

The nuclear mass is taken to be in MeV. So it is really the rest mass energy, not the mass, but who is complaining? Just divide by c^2 to get the actual mass. The final two correction terms are really small, especially for light nuclei, and are often left away (but not in the data presented here).

14.7 Draft: Binding energy

The binding energy of a nucleus is the energy that would be needed to take it apart into its individual protons and neutrons. Binding energy explains the overall trends in nuclear reactions.

As explained in the previous section, the binding energy $E_{\rm B}$ can be found from the mass of the nucleus. The specific binding energy is defined as the binding energy per nucleon, $E_{\rm B}/A$. Figure 14.4 shows the specific binding energy of the nuclei with known masses. The highest specific binding energy is 8.8 MeV, and occurs for ${}^{62}_{28}$Ni nickel. Nickel has 28 protons, a magic number. However, nonmagic ${}^{58}_{26}$Fe and ${}^{56}_{26}$Fe are right on its heels.

Nuclei can therefore lower their total energy by evolving towards the nickel-iron region. Light nuclei can "fusion" together into heavier ones to do so. Heavy nuclei can emit alpha particles or fission, fall apart in smaller pieces.

Figure 14.4 also shows that the binding energy of most nuclei is roughly 8 MeV per nucleon. However, the very light nuclei are an exception; they tend to have a quite small binding energy per nucleon. In a light nucleus, each nucleon only experiences attraction from a small number of other nucleons. For example, deuterium only has a binding energy of 1.1 MeV per nucleon.

The big exception to the exception is the doubly magic ${}^{4}_{2}$He nucleus, the alpha particle. It has a stunning 7.07 MeV binding energy per nucleon, exceeding its immediate neighbors by far.

The ${}^{8}_{4}$Be beryllium nucleus is not bad either, also with 7.07 MeV per nucleon, almost exactly as high as ${}^{4}_{2}$He, though admittedly that is achieved using eight nucleons instead of only four. But clearly, ${}^{8}_{4}$Be is a lot more tightly bound than its immediate neighbors.

It is therefore ironic that while various of those neighbors are stable, the much more tightly bound ${}^{8}_{4}$Be is not. It falls apart in about 67 as (i.e. 67 10^{-18} s), a tragic consequence of being able to come neatly apart into two alpha particles that are just a tiny bit more tightly bound. It is the only alpha decay among the light nuclei. It is an exception to the rule that light nuclei prefer to fusion into heavier ones.

But despite its immeasurably short half-life, do not think that ${}^{8}_{4}$Be is not important. Without it there would be no life on earth. Because of the absence of stable intermediaries, the Big Bang produced no elements heavier than beryllium, (and only trace amounts of that) including no carbon. As Hoyle pointed out, the carbon of life is formed in the interior of aging stars when ${}^{8}_{4}$Be captures a third alpha particle, to produce ${}^{12}_{6}$C, which is stable. This is called the "triple alpha process." Under the extreme conditions in the interior of collapsing stars, given time this process produces significant amounts of carbon despite the extremely short half-life of ${}^{8}_{4}$Be. The process is far too slow to have occurred in the Big Bang, however.

For ${}^{12}_{6}\mathrm{C}_{6}$ carbon, the superior number of nucleons has become big enough to

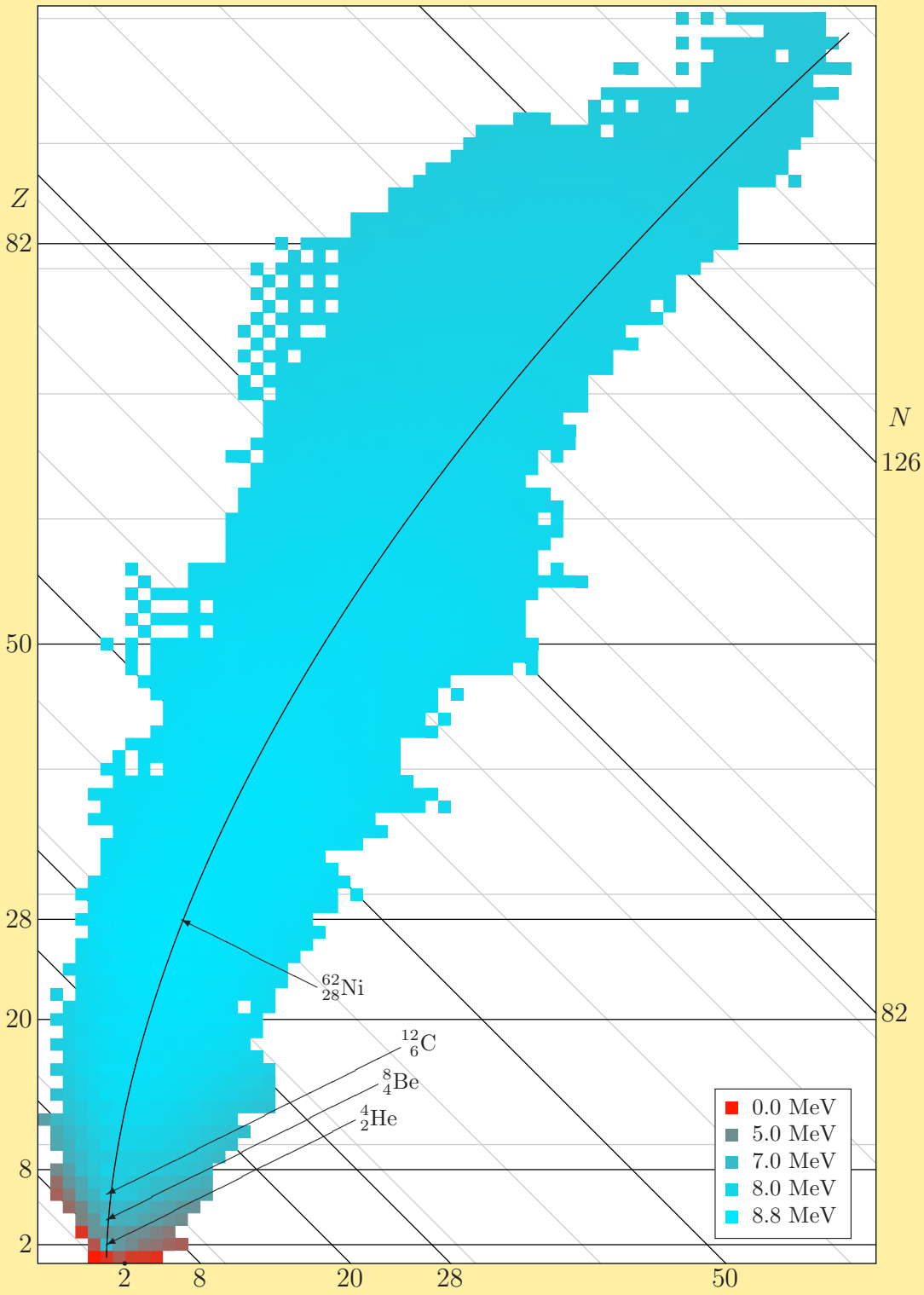


Figure 14.4: Binding energy per nucleon. [pdf][con]

overcome the doubly magic advantage of the three corresponding alpha particles. Carbon-12's binding energy is 7.68 MeV per nucleon, greater than that of alpha particles.

14.8 Draft: Nucleon separation energies

Nucleon separation energies are the equivalent of atomic ionization energies, but for nuclei. The proton separation energy is the minimum energy required to remove a proton from a nucleus. It is how much the rest mass energy of the nucleus is less than that of the nucleus with one less proton and a free proton.

Similarly, the neutron separation energy is the energy needed to remove a neutron. Figures 14.5 and 14.6 show proton and neutron separation energies as grey tones. Note that these energies are quite different from the average binding energy per nucleon given in the previous subsection. In particular, it takes a lot of energy to take another proton out of an already proton-deficient nucleus. And the same for taking a neutron out of an already neutron deficient nucleus.

In addition, the vertical striping in 14.5 shows that the proton separation energy is noticeably higher if the initial number of protons is even than if it is odd. Nucleons of the same kind like to pair up. If a proton is removed from a nucleus with an even number of protons, a pair must be broken up, and that requires additional energy. The neutron separation energy 14.6 shows diagonal striping for similar reasons; neutrons too pair up.

There is also a visible step down in overall grey level at the higher magic numbers. It is not dramatic, but real. It illustrates that the nucleon energy levels come in "shells" terminated by magic numbers. In fact, this step down in energy *defines* the magic numbers. That is discussed further in section 14.12.

Figures 14.7 and 14.8 show the energy to remove two protons, respectively two neutrons from even-even nuclei. This show up the higher magic numbers more clearly as the pairing energy effect is removed as a factor.

14.9 Draft: Modeling the Deuteron

This book largely limits itself to relatively simple, but effective, models for nuclei. However, the deuteron, the deuterium nucleus, is the most simple nontrivial nucleus, as it consists of only a single proton. So, to give a rough idea of what sort of more advanced nuclear theories there are out there, this one section will look at the deuteron in some more detail.

Addendum {A.41} explores some of the nuclear potentials that you can write down to model the deuteron. Simple potentials such as those described there give a lot of insight in the deuteron. They give ballpark values for the potential and kinetic energies. They also give an explanation for the observations that the deuteron is only weakly bound and that there is no second bound state.

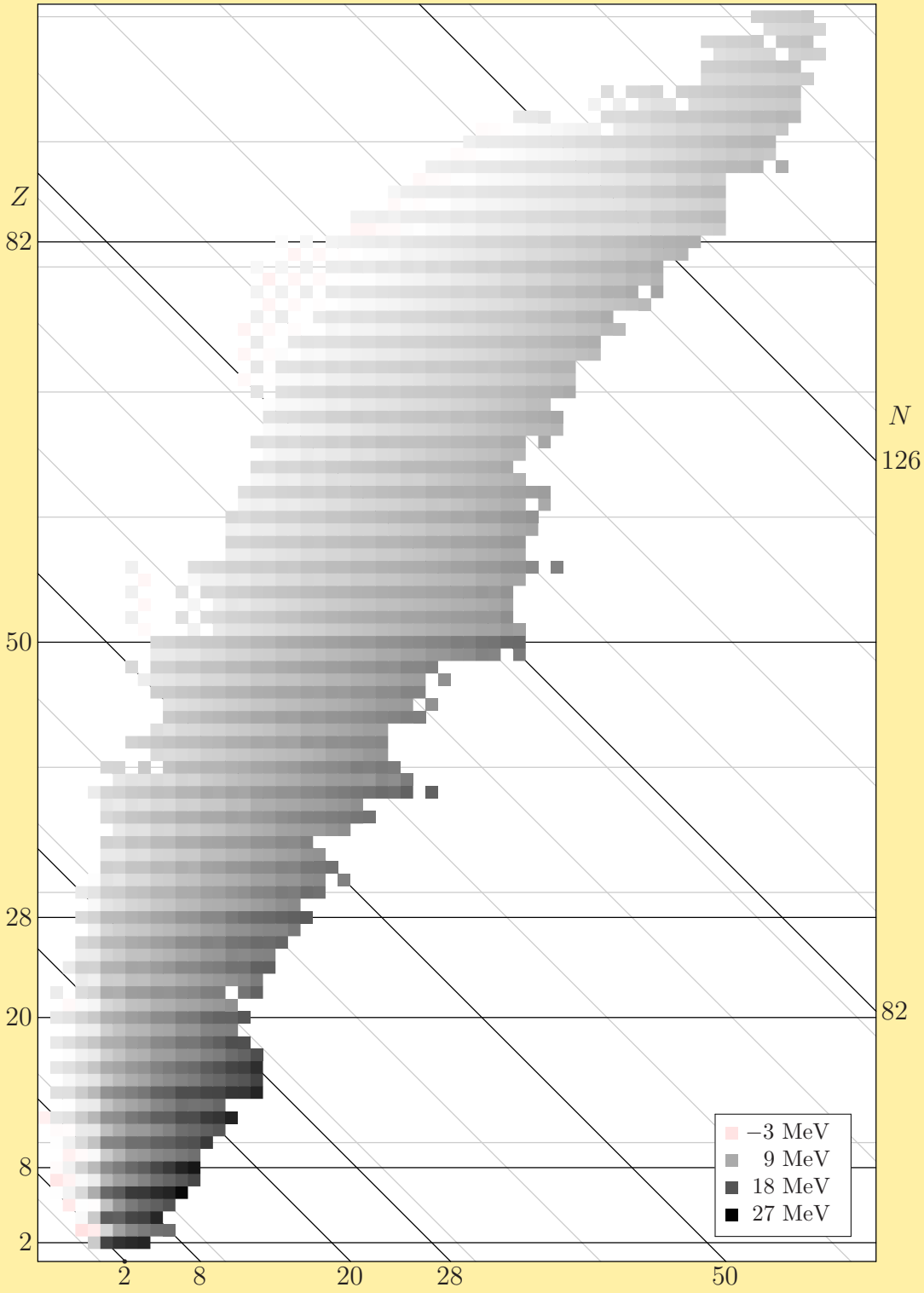


Figure 14.5: Proton separation energy. [pdf][con]

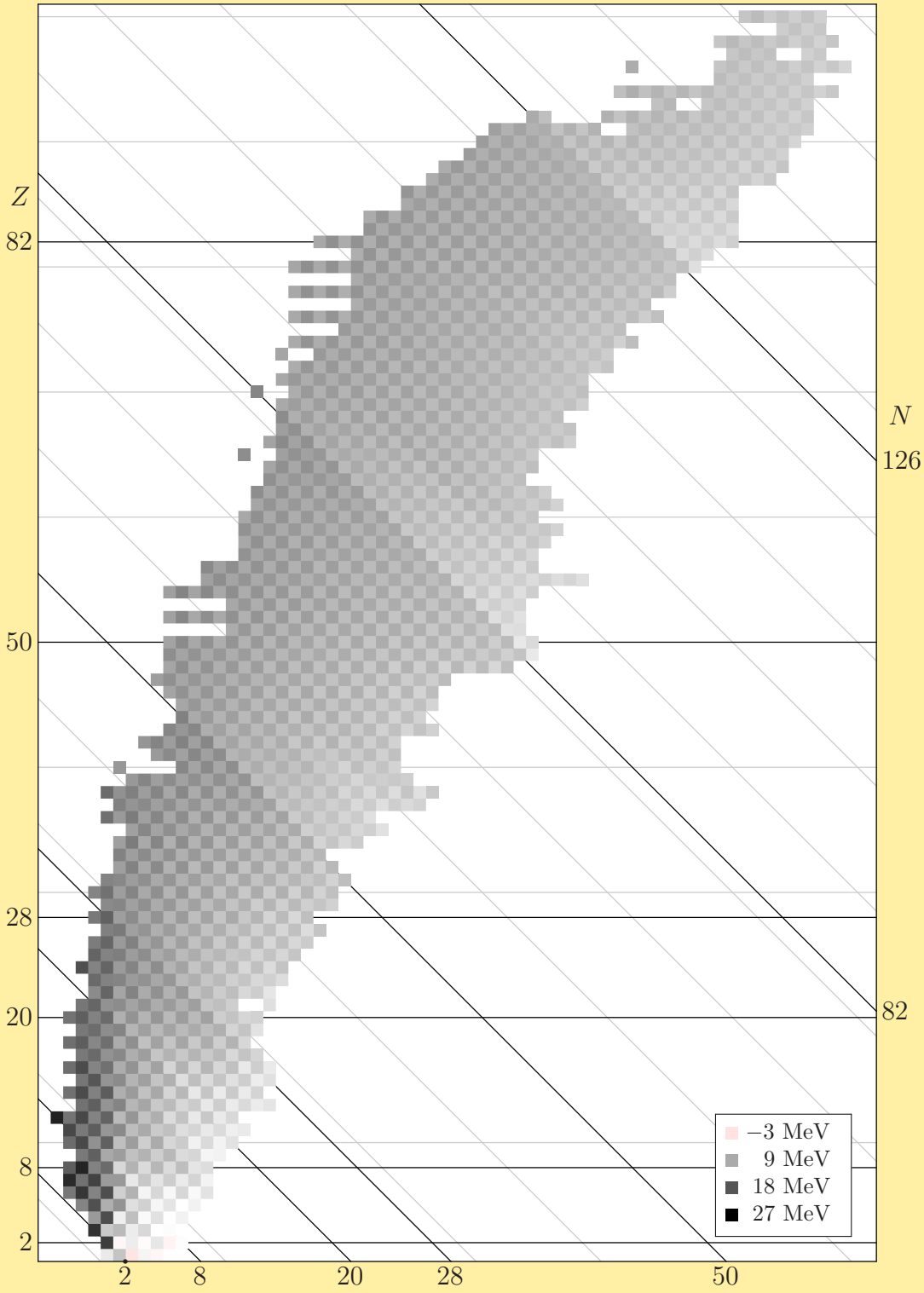


Figure 14.6: Neutron separation energy. [pdf][con]

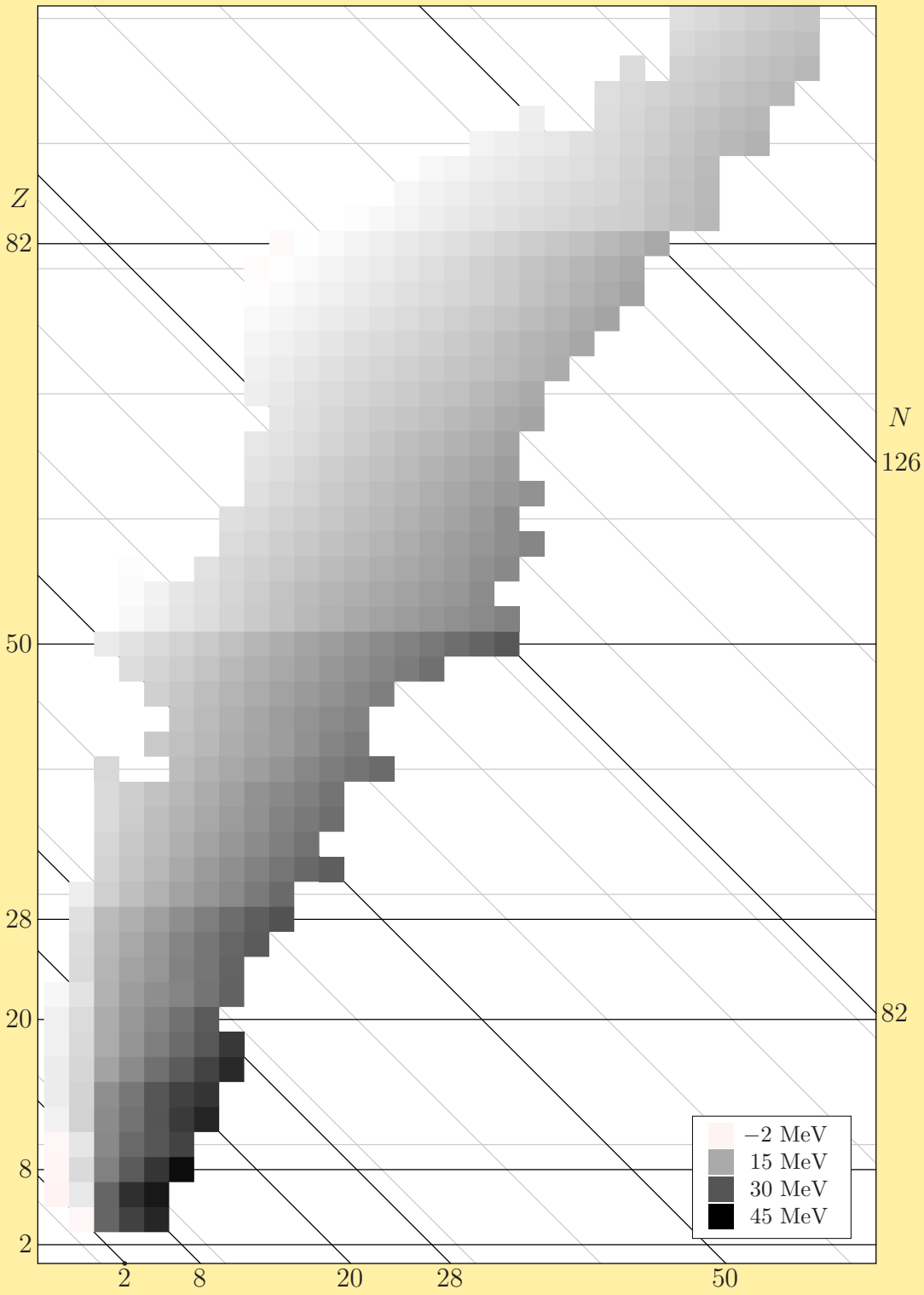


Figure 14.7: Proton pair separation energy. [pdf][con]

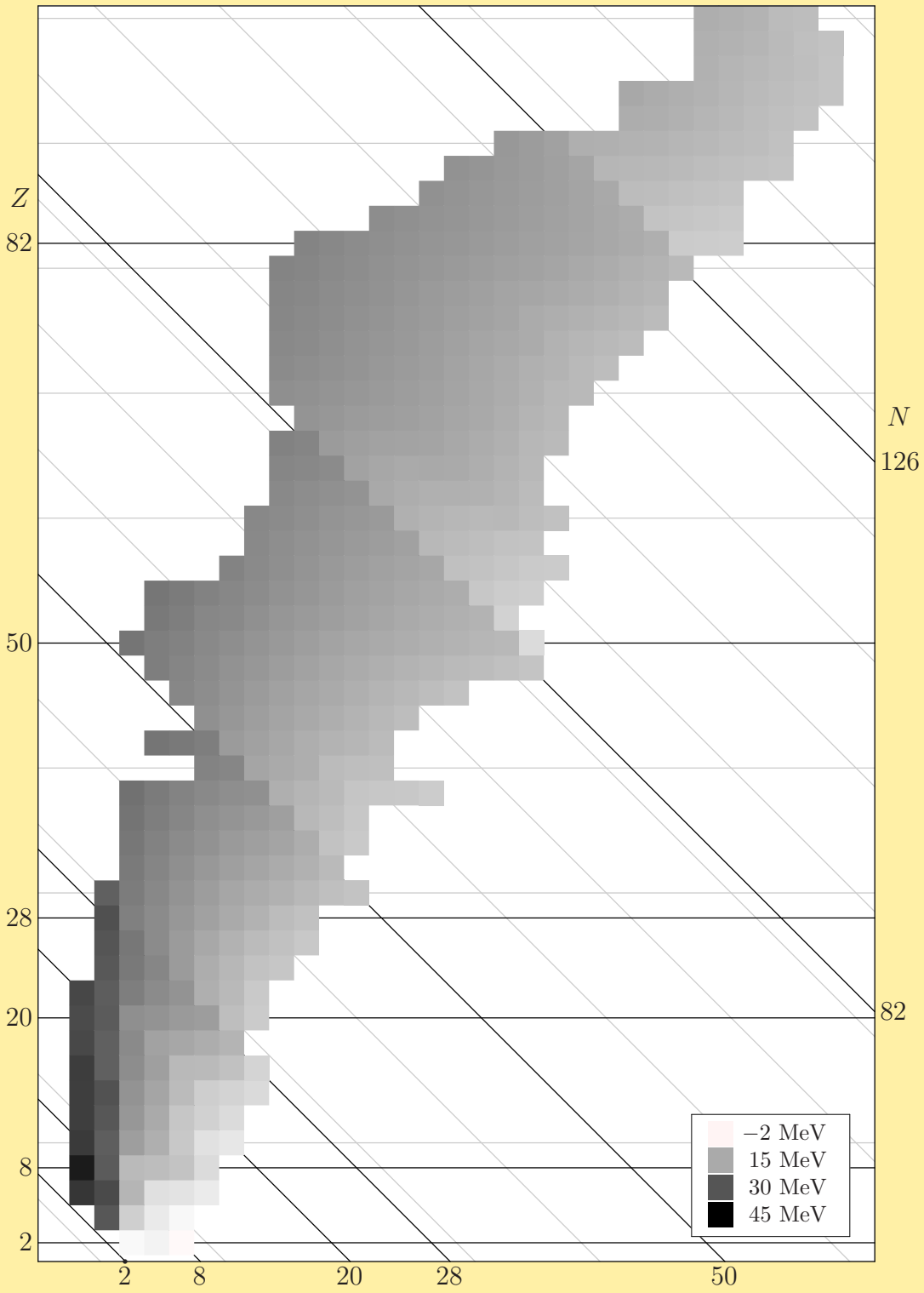


Figure 14.8: Neutron pair separation energy. [pdf][con]

They show the relatively large size of the deuteron: there is a good chance that the proton and neutron can be found way apart. Simple additions to the potential can describe the spin dependence and the violation of orbital angular momentum conservation.

However, there is a problem. These potentials describe a deuteron consisting of a proton and a neutron. But the proton and neutron are not elementary particles: each consists of three quarks.

This internal structure should not be a much of a concern when the proton and neutron are relatively far apart. But if the two get really close? Surely in that case the physics should be described in terms of six quarks that interact through gluons and the Pauli exclusion principle, [5, p. 95]? Eventually the proton and neutron must lose their identity. Then there is no longer any reasonable justification for a picture of a free-space proton interacting with a free-space neutron.

A rough idea of the scales involved may be obtained by looking at charge radii. The charge radius of a particle is a measure of the size of its charge distribution. Now the proton has a charge radius of about 0.88 fm. The deuteron has a charge radius of 2.14 fm. So at least the charge radius of a proton is not that much smaller than the size of the deuteron. And a quantum description of the deuteron needs to consider all possible nucleon spacings. Clearly for the smallest nucleon spacings, the two must intrude nontrivially into each other's space.

Consider another example of compound structures, noble gas atoms. When such atoms are relatively far apart, they can be modeled well as point particles forming an ideal gas. You might add some simple Van der Waals forces to that picture. However, when the atoms get pushed close together, the electromagnetic interactions between them become much more complex. And if you try to push the atoms even closer together, they resist that very strongly. The reason is the Pauli exclusion principle, chapter 5.10. More than two electrons cannot be pushed in the same spatial state.

The big difference is of course that the electromagnetic interactions of the electrons and nuclei that make up atoms are well understood. There is as yet no good way to describe the color force interactions between the quarks that make up nucleons. (Certainly not at the relatively low energies of importance for nuclear structure.)

Physicists have developed a model that is somewhere intermediate between that of interacting free-space nucleons and interacting quarks. In the model, the forces between nucleons are produced by the exchange of particles called "pions." That is much like how in relativistic quantum mechanics, electromagnetic forces are produced by the exchange of photons. Or how the forces between quarks are believed to be produced by the exchange of gluons. These exchanged particles are "virtual" ones.

Roughly speaking, relativistic mass-energy equivalence combined with quan-

tum uncertainty in energy allows some possibility for these particles to be found near the real particles. (See addendum {A.42} for a more quantitative description of these ideas.)

The picture is now that the proton and neutron are elementary particles but "dressed" in a coat of virtual pions. Pions consist of two quarks. More precisely, they consist of a quark and an antiquark. There are three pions; the positively charged π^+, the neutral π^0, and the negatively charged π^-. Pions have no spin. They have negative intrinsic parity. The charged pions have a mass of 140 MeV, while the uncharged one has a slightly smaller mass of 135 MeV.

When the neutron and proton exchange a pion, they also exchange its momentum. That produces a force between them.

There are a number of redeeming features to this model:

1. It explains the short range of the nuclear force. That is worked out in addendum {A.42}.

   However, the same result is often derived much quicker and easier in literature. That derivation does not require any Hamiltonians to be written down, or even any mathematics above the elementary school level. It uses the popular so-called "energy-time uncertainty equality," chapter 7.2.2,

   any energy difference you want $\times$ any time difference you want $= \frac{1}{2}\hbar$

   To apply it to pion exchange, replace "any energy difference you want" with the rest mass energy of the virtual pion that supposedly pops up, about 138 MeV on average. Replace "any time difference you want" with the time that the pion exists. (Model the pion here as a classical particle with definite values of position versus time.) From the time that the pion exists, you can compute how far it travels. That is because clearly it must travel with about half the speed of light: it cannot travel with a speed less than zero nor more than the speed of light. Put in the numbers, ignore the stupid factors $\frac{1}{2}$ because you do not need to be that accurate, and it shows that the pion would travel about 1.4 fm if it had a position to begin with. That range of the pion is consistent with the experimental data on the range of the nuclear force.

   (OK, someone might object the pions do most decidedly not pop up and disappear again. The ground state of a nucleus is an energy eigenstate and those are stationary, chapter 7.1.4. But why worry about such minor details?)

2. It gives a reasonable explanation of the anomalous magnetic moments of the proton and neutron. The magnetic moment of the proton can be written as

$$\mu_{\rm p} = g_{\rm p}\tfrac{1}{2}\mu_{\rm N} \quad g_{\rm p} = 5.586 \qquad \mu_{\rm N} = \frac{e\hbar}{2m_{\rm p}} \approx 5.051\ 10^{-27}\ {\rm J/T}$$

Here μ_N is called the nuclear magneton. It is formed with the charge and mass of the proton. Now, if the proton was an elementary particle with spin one-half, the factor g_p should be two, chapter 13.4. Or at least close to it. The electron is an elementary particle, and its similarly defined g-factor is 2.002. (Of course that uses the electron mass instead of the proton one.)

The magnetic moment of the neutron can similarly be written

$$\mu_n = g_n \tfrac{1}{2}\mu_N \quad g_n = -3.826 \qquad \mu_N = \frac{e\hbar}{2m_p} \approx 5.051\ 10^{-27}\ \text{J/T}$$

Note that the proton charge and energy are used here. In fact, if you consider the neutron as an elementary particle with no charge, it should not have a magnetic moment in the first place.

Suppose however that the neutron occasionally briefly flips out a negatively charged virtual π^-. Because of charge conservation, that will leave the neutron as a positively charged proton. But the pion is much lighter than the proton. Lighter particles produce much greater magnetic moments, all else being the same, chapter 13.4. In terms of classical physics, lighter particles circle around a lot faster for the same angular momentum. To be sure, as a spinless particle the pion has no intrinsic magnetic moment. However, because of parity conservation, the pion should have at least one unit of orbital angular momentum to compensate for its intrinsic negative parity. So the neutral neutron acquires a big chunk of unexpected negative magnetic moment.

Similar ideas apply for the proton. The proton may temporarily flip out a positively charged π^+, leaving itself a neutron. Because of the mass difference, this can be expected to significantly increase the proton magnetic moment.

Apparently then, the virtual π^+ pions increase the g-factor of the proton from 2 to 5.586, an increase of 3.586. So you would expect that the virtual π^- pions decrease the g-factor of the neutron from 0 to -3.586. That is roughly right, the actual g-factor of the neutron is -3.826.

The slightly larger value seems logical enough too. The fact that the proton turns occasionally into a neutron should decrease its total magnetic moment. Conversely, the neutron occasionally turns into a proton. Assume that the neutron has half a unit of spin in the positive chosen z-direction. That is consistent with one unit of orbital angular momentum for the π^- and minus half a unit of spin for the proton that is left behind. So the proton spins in the opposite direction of the neutron, which means that it increases the magnitude of the negative neutron magnetic moment. (The other possibility, that

the π^{-1} has no z-momentum and the proton has half a positive unit, would decrease the magnitude. But this possibility has only half the probability, according to the Clebsch-Gordan coefficients figure 12.6.)

3. The same ideas also provide an explanation for a problem with the magnetic moments of heavier nuclei. These do not fit theoretical data that well, figure 14.42. A closer examination of these data suggests that the intrinsic magnetic moments of nucleons are smaller inside nuclei than they are in free space. That can reasonably be explained by assuming that the proximity of other nucleons affects the coats of virtual pions.
4. It explains a puzzling observation when high-energy neutrons are scattered off high-energy protons going the opposite way. Because of the high energies, you would expect that only a few protons and neutrons would be significantly deflected from their path. Those would be caused by the few collisions that happen to be almost head-on. And indeed relatively few are scattered to say about 90° angles. But an unexpectedly large number seems to get scattered almost 180°, straight back to where they came from. That does not make much sense.

   The more reasonable explanation is that the proton catches a virtual π^- from the neutron. Or the neutron catches a virtual π^+ from the proton. Either process turns the proton into a neutron and vice-versa. So an apparent reflected neutron is really a proton that kept going straight but changed into a neutron. And the same way for an apparent reflected proton.
5. It can explain why charge independence is less perfect than charge symmetry. A pair of neutrons can only exchange the neutral π^0. Exchange of a charged pion would create a proton and a nucleon with charge -1. The latter does not exist. (At least not for measurable times, nor at the energies of most interest here.) Similarly, a pair of protons can normally only exchange a π^0. But a neutron and a proton can also exchange charged pions. So pion interactions are not charge independent.
6. The nuclear potential that can be written down analytically based on pion exchange works very well at large nucleon spacings. This potential is called the "OPEP," for One Pion Exchange Potential, {A.42}.

There are also drawbacks to the pion exchange approach.

For one, the magnetic moments of the neutron and proton can be reasonably explained by simply adding those of the constituent quarks, [31, pp. 74, 745]. To be sure, that does not directly affect the question whether the pion exchange model is useful. But it does make the dressed nucleon picture look quite con-

trived.

A bigger problem is nucleon spacings that are not so large. One-pion exchange is generally believed to be dominant for nucleon spacings above 3 fm, and reasonable for spacings above 2 fm, [36, p. 135, 159], [5, p. 86, 91]. However, things get much more messy for spacings shorter than that. That includes the vital range of spacings of the primary nucleon attractions and repulsions. For these, two-pion exchange must be considered. In addition, excited pion states and an excited nucleon state need to be included. That is much more complicated. See addendum {A.42} for a brief intro to some of the issues involved.

And for still shorter nucleon spacings, things get very messy indeed, including multi-pion exchanges and a zoo of other particles. Eventually the question must be at what spacing nucleons lose their distinctive character and a model of quarks exchanging gluons becomes unavoidable. Fortunately, very close spacings correspond to very high energies since the nucleons strongly repel each other at close range. So very close spacings may not be that important for most nuclear physics.

Because of the above and other issues, many physicists use a less theoretical approach. The OPEP is still used at large nucleon spacings. But at shorter spacings, relatively simple chosen potentials are used. The parameters of those "phenomenological" potentials are adjusted to match the experimental data.

It makes things a lot simpler. And it is not clear whether the theoretical models used at smaller nucleon spacings are really that much more justified. However, phenomenological potentials do require that large numbers of parameters are fit to experimental data. And they have a nasty habit of not working that well for experimental data different from that used to define their parameters, [32].

Regardless of potential used, it is difficult to come up with an unambiguous probability for the $l = 2$ orbital angular momentum. Estimates hover around the 5% value, but a clear value has never been established. This is not encouraging, since this probability is an integral quantity. If it varies nontrivially from one model to the next, then there is no real convergence on a single deuteron model. Of course, if the proton and neutron are modeled as interacting clouds of particles, it may not even be obvious how to define their orbital angular momentum in the first place, [Phys. Rev. C 19,20 (1979) 1473,325]. And that in turn raises questions in what sense these models are really well-defined two-particle ones.

Then there is the very big problem of generalizing all this to systems of three or more nucleons. One current hope is that closer examination of the underlying quark model may produce a more theoretically justified model in terms of nucleons and mesons, [32].

14.10 Draft: Liquid drop model

Nucleons attract each other with nuclear forces that are not completely understood, but that are known to be short range. That is much like molecules in a classical liquid drop attract each other with short-range Van der Waals forces. Indeed, it turns out that a liquid drop model can explain many properties of nuclei surprisingly well. This section gives an introduction.

14.10.1 Draft: Nuclear radius

The volume of a liquid drop, hence its number of molecules, is proportional to the cube of its radius R. Conversely, the radius is proportional to the cube root of the number of molecules. Similarly, the radius of a nucleus is approximately equal to the cube root of the number of nucleons:

$$R \approx R_A \sqrt[3]{A} \qquad R_A = 1.23 \text{ fm} \tag{14.9}$$

Here A is the mass number, equal to the number of protons Z plus the number of neutrons N. Also fm stands for "femtometer," equal to 10^{-15} meter; it may be referred to as a "fermi" in some older references. Enrico Fermi was a great curse for early nuclear physicists, quickly doing all sorts of things before they could.

It should be noted that the above nuclear radius is an average one. A nucleus does not stop at a very sharply defined radius. (And neither would a liquid drop if it only contained 100 molecules or so.) Also, the constant R_A varies a bit with the nucleus and with the method used to estimate the radius. Values from 1.2 to 1.25 are typical. This book will use the value 1.23 stated above.

It may be noted that these results for the nuclear radii are quite solidly established experimentally. Physicists have used a wide variety of ingenious methods to verify them. For example, they have bounced electrons at various energy levels off nuclei to probe their Coulomb fields, and alpha particles to also probe the nuclear forces. They have examined the effect of the nuclear size on the electron spectra of the atoms; these effects are very small, but if you substitute a muon for an electron, the effect becomes much larger since the muon is much heavier. They have dropped pi mesons on nuclei and watched their decay. They have also compared the energies of nuclei with Z protons and N neutrons against the corresponding "mirror nuclei" that have with N protons and Z neutrons. There is good evidence that the nuclear force is the same when you swap neutrons with protons and vice versa, so comparing such nuclei shows up the Coulomb energy, which depends on how tightly the protons are packed together. All these different methods give essentially the same results for the nuclear radii. They also indicate that the neutrons and protons are well-mixed throughout the nucleus, [31, pp. 44-59]

14.10.2 Draft: von Weizsäcker formula

The binding energy of nuclei can be approximated by the "von Weizsäcker formula," or "Bethe-von Weizsäcker formula:"

$$\boxed{E_{\mathrm{B,vW}} = C_v A - C_s A^{2/3} - C_c \frac{Z(Z-C_z)}{A^{1/3}} - C_d \frac{(Z-N)^2}{A} - C_p \frac{o_Z + o_N - 1}{A^{C_e}}} \tag{14.10}$$

where the $C_\cdot$ are constants, while o_Z is 1 if the number of protons is odd and zero if it is even, and similar for o_N for neutrons. This book uses values given by [37] for the constants:

$$C_v = 15.409 \text{ MeV} \quad C_s = 16.873 \text{ MeV} \quad C_c = 0.695 \text{ MeV} \quad C_z = 1$$

$$C_d = 22.435 \text{ MeV} \quad C_p = 11.155 \text{ MeV} \quad C_e = 0.5$$

where a MeV (mega electron volt) is $1.602\,18\ 10^{-13}$ J, equal to the energy that an electron picks up in a one million volt electric field.

Plugged into the mass-energy relation, the von Weizsäcker formula produces the so-called "semi-empirical mass formula:"

$$m_{\mathrm{nucleus,SE}} = Z m_{\mathrm{p}} + N m_{\mathrm{n}} - \frac{E_{\mathrm{B,vW}}}{c^2} \tag{14.11}$$

14.10.3 Draft: Explanation of the formula

The various terms in the von Weizsäcker formula of the previous subsection have quite straightforward explanations. The C_v term is typical for short-range attractive forces; it expresses that the energy of every nucleon is lowered the same amount by the presence of the attracting nucleons in its immediate vicinity. The classical analogue is that the energy needed to boil away a drop of liquid is proportional to its mass, hence to its number of molecules.

The C_s term expresses that nucleons near the surface are not surrounded by a complete set of attracting nucleons. It raises their energy. This affects only a number of nucleons proportional to the surface area, hence proportional to $A^{2/3}$. The effect is negligible for a classical drop of liquid, which may have a million molecules along a diameter, but not for a nucleus with maybe ten nucleons along it. (Actually, the effect is important for a classical drop too, even if it does not affect its overall energy, as it gives rise to surface tension.)

The C_c term expresses the Coulomb repulsion between protons. Like the Coulomb energy of a sphere with constant charge density, it is proportional to the square net charge, so to Z^2 and inversely proportional to the radius, so to $A^{1/3}$. However, the empirical constant C_c is somewhat different from that of a constant charge density. Also, a correction $C_z = 1$ has been thrown in to ensure that there is no Coulomb repulsion if there is just one proton.

The last two terms cheat; they try to deviously include quantum effects in a supposedly classical model. In particular, the C_d term adds an energy increasing with the square of the difference in number of protons and neutrons. It simulates the effect of the Pauli exclusion principle. Assume first that the number of protons and neutrons is equal, each $A/2$. In that case the protons will be able to occupy the lowest $A/2$ proton energy levels, and the neutrons the lowest $A/2$ neutron levels. However, if then, say, some of the protons are turned into neutrons, they will have to move to energy levels above $A/2$, because the lowest $A/2$ neutron levels are already filled with neutrons. Therefore the energy goes up if the number of protons and neutrons becomes unequal.

The last C_p term expresses that nucleons of the same type like to pair up. When both the number of protons and the number of neutrons is even, all protons can pair up, and all neutrons can, and the energy is lower than average. When both the number of protons is odd and the number of neutrons is odd, there will be an unpaired proton as well as an unpaired neutron, and the energy is higher than average.

14.10.4 Draft: Accuracy of the formula

Figure 14.9 shows the error in the von Weizsäcker formula as colors. Blue means that the actual binding energy is higher than predicted, red that it is less than predicted. For very light nuclei, the formula is obviously useless, but for the remaining nuclei it is quite good. Note that the error is in the order of MeV, to be compared to a total binding energy of about $8A$ MeV. So for heavy nuclei the *relative* error is small.

Near the magic numbers the binding energy tends to be greater than the predicted values. This can be qualitatively understood from the quantum energy levels that the nucleons occupy. When nucleons are successively added to a nucleus, those that go into energy levels just below the magic numbers have unusually large binding energy, and the total nuclear binding energy increases above that predicted by the von Weizsäcker formula. The deviation from the formula therefore tends to reach a maximum at the magic number. Just above the magic number, further nucleons have a much lower energy level, and the deviation from the von Weizsäcker value decreases again.

14.11 Draft: Alpha Decay

In alpha decay a nucleus emits an "alpha particle," later identified to be simply a helium-4 nucleus. Since the escaping alpha particle consists of two protons plus two neutrons, the atomic number Z of the nucleus decreases by two and the mass number A by four. This section explains why alpha decay occurs.

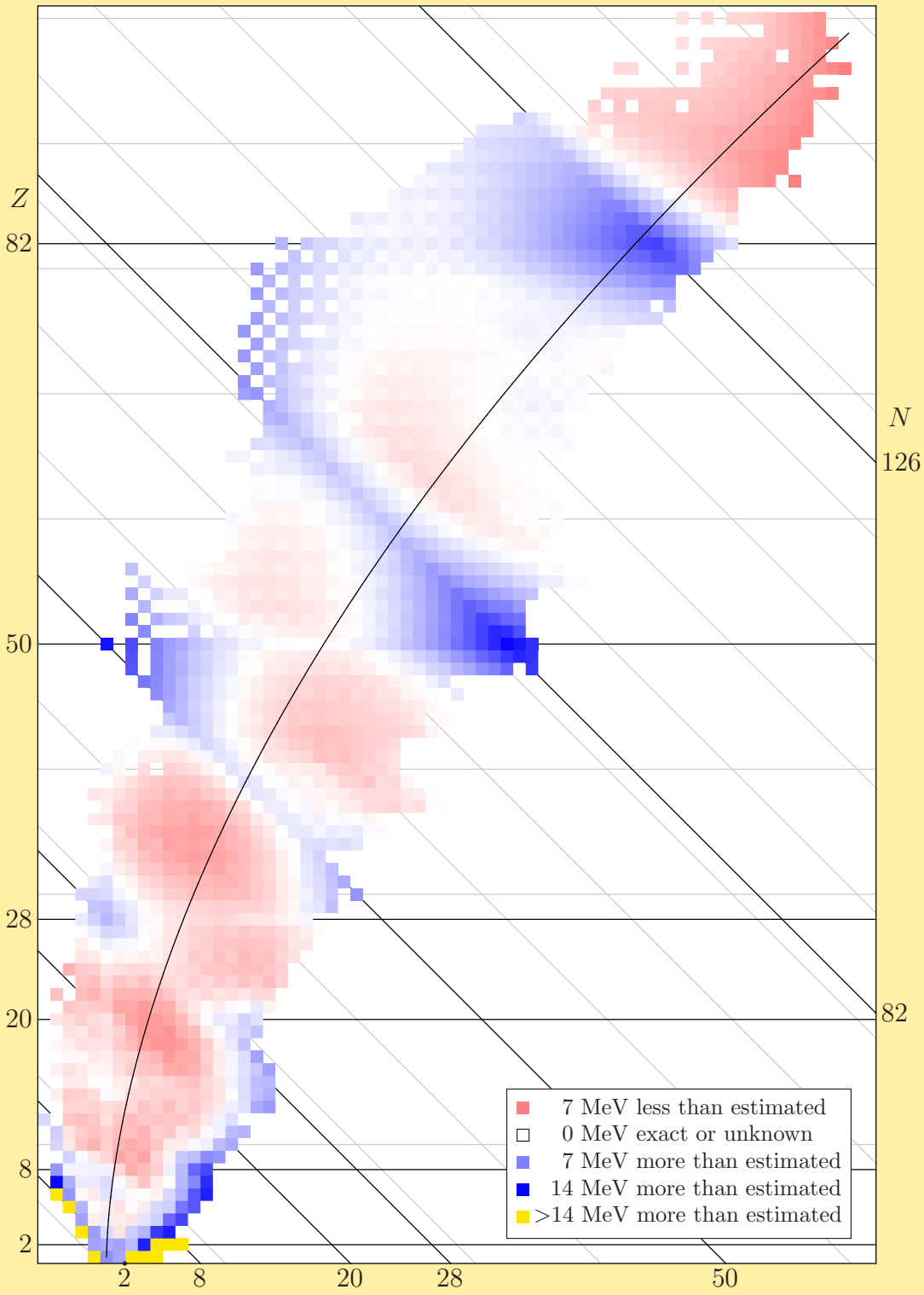


Figure 14.9: Error in the von Weizsäcker formula. [pdf][con]

14.11.1 Draft: Decay mechanism

Figure 14.10 gives decay data for the nuclei that decay exclusively through alpha decay. Nuclei are much like cherries: they have a variable size that depends mainly on their mass number A, and a charge Z that can be shown as different shades of red. You can even define a "stem" for them, as explained later. Nuclei with the same atomic number Z are joined by branches.

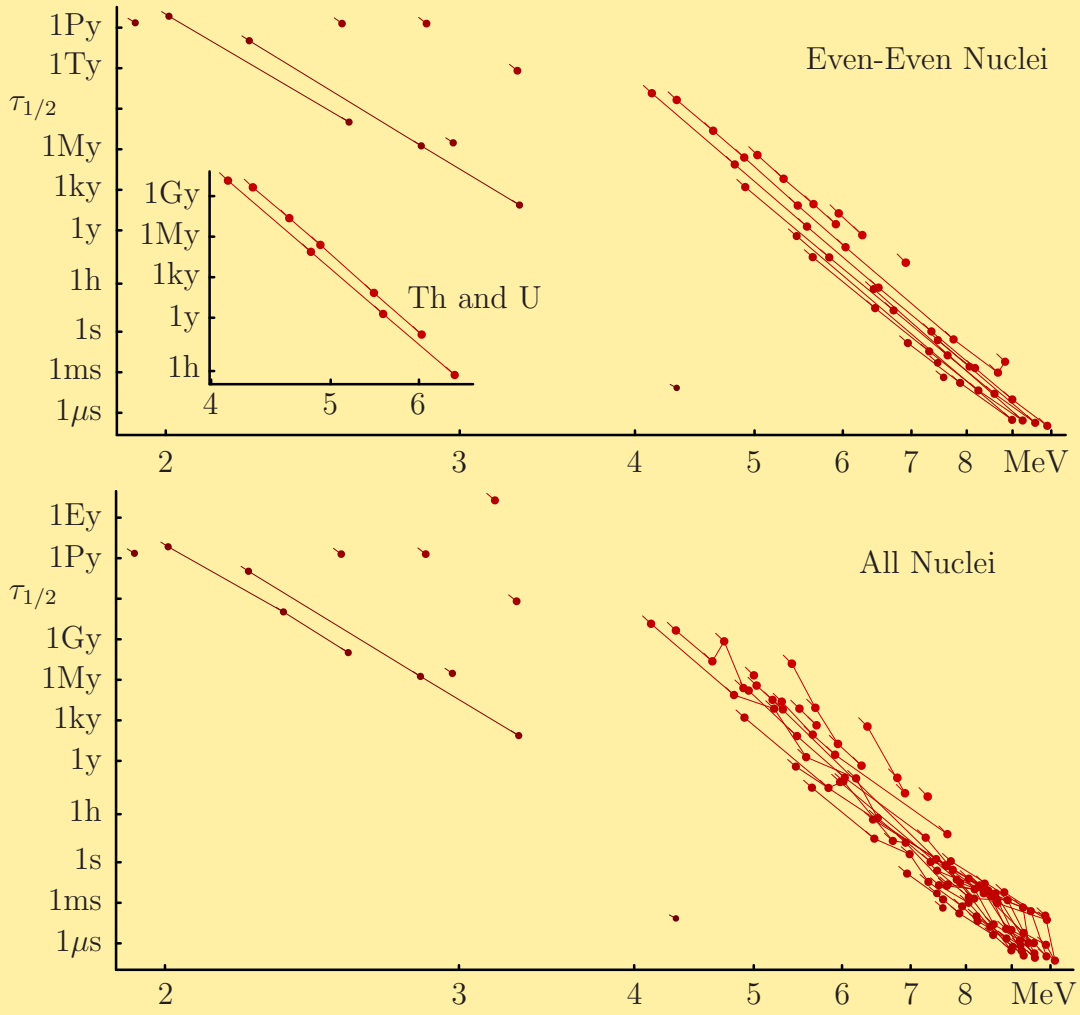


Figure 14.10: Half-life versus energy release for the atomic nuclei marked in NUBASE 2003 as showing pure alpha decay with unqualified energies. Top: only the even values of the mass and atomic numbers cherry-picked, omitting ${}^{8}_{4}\mathrm{Be}$. Inset: really cherry-picking, only a few even mass numbers for thorium and uranium! Bottom: all the nuclei except ${}^{8}_{4}\mathrm{Be}$ (67 as, 0.092 MeV). [pdf]

Not shown in figure 14.10 is the unstable beryllium isotope ${}^{8}_{4}\mathrm{Be}$, which has a half-life of only 67 as, (i.e. 67 10^{-18} s), and a decay energy of only 0.092 MeV. As you can see from the graph, these numbers are wildly different from the

other, much heavier, alpha-decay nuclei, and inclusion would make the graph very messy.

Note the tremendous range of half-lives in figure 14.10, from mere nanoseconds to quintillions of years. And that excludes beryllium's attoseconds. In the early history of alpha decay, it seemed very hard to explain how nuclei that do not seem that different with respect to their numbers of protons and neutrons could have such dramatically different half-lives. The energy that is released in the decay process does not vary that much, as figure 14.10 also shows.

To add to the mystery in those early days of quantum mechanics, if an alpha particle was shot back at the nucleus with the same energy that it came out, it would not go back in! It was reflected by the electrostatic repulsion of the positively charged nucleus. So, it had not enough energy to pass through the region of high potential energy surrounding the nucleus, yet it *did* pass through it when it came out.

Gamow, and independently Gurney & Condon, recognized that the explanation was quantum tunneling. Tunneling allows a particle to get through a potential energy barrier even if classically it does not have enough energy to do so, chapter 7.12.2.

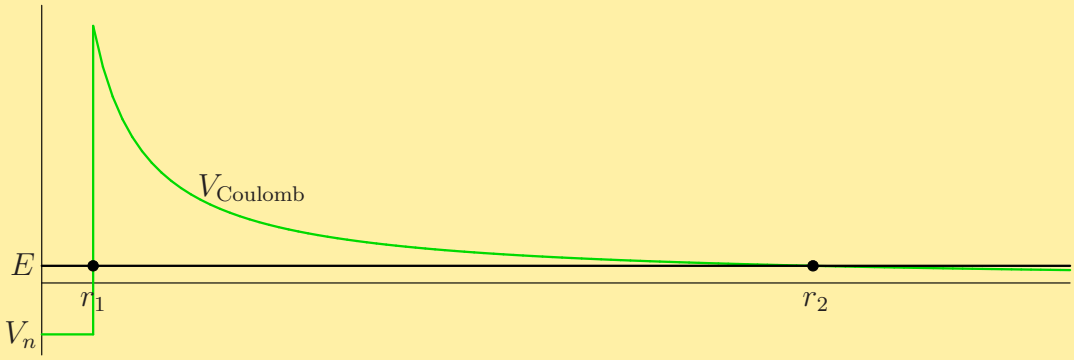


Figure 14.11: Schematic potential for an alpha particle that tunnels out of a nucleus.

Figure 14.11 gives a rough model of the barrier.The horizontal line represents the total energy of the alpha particle. Far from the nucleus, the potential energy V of the alpha particle can be defined to be zero. Closer to the nucleus, the potential energy of the alpha particle ramps up due to Coulomb repulsion. However, right at the outer edge $r = R$ of the nucleus itself, the strong but very short-range attractive nuclear force pops up, and the combined potential energy plummets almost vertically downwards to some low value V_n. In between the radial position $r_1 \approx R$ and some larger radius r_2, the potential energy exceeds the total energy that the alpha particle has available. Classically, the alpha particle cannot penetrate into this region. However, in quantum mechanics it retains a very small probability of doing so.

The region in between r_1 and r_2 is called the "Coulomb barrier." It is a poorly chosen name, because the barrier is only a Coulomb one for an alpha particle trying to get *in* the nucleus. For an alpha particle trying to get *out*, it is a nuclear force barrier; here the Coulomb force assists the tunneling particle to get through the barrier and escape. The term "nuclear barrier" would avoid this ambiguity. Therefore physicists do not use it.

Now, to get a rough picture of alpha decay, imagine an alpha particle wave packet "rattling around" inside the nucleus trying to escape. Each time it hits the barrier at r_1, it has a small chance of escaping. Eventually it gets lucky.

Assume that the alpha particle wave packet is small enough that the motion can be assumed to be one-dimensional. Then the small chance of escaping each time it hits the barrier is approximately given by the analysis of chapter 7.13 as

$$\boxed{T \approx e^{-2\gamma_{12}} \qquad \gamma_{12} = \frac{1}{\hbar}\int_{r_1}^{r_2} \sqrt{2m_\alpha(V-E)}\,\mathrm{d}r} \tag{14.12}$$

The fact that this probability involves an exponential is the basic reason for the tremendous range in half-lives: exponentials can vary greatly in magnitude for relatively modest changes in their argument.

14.11.2 Draft: Comparison with data

The previous subsection explained alpha decay in terms of an imprisoned alpha particle tunneling out of the nucleus. To verify whether that is reasonable, the next step is obviously to put in some ballpark numbers and see whether the experimental data can be explained.

First, the energy E of the alpha particle may be found from Einstein's famous expression $E = mc^2$, section 14.6. Just find the difference between the rest mass of the original nucleus and the sum of that of the final nucleus and the alpha particle, and multiply by the square speed of light. That gives the energy release. It comes out primarily as kinetic energy of the alpha particle, ignoring any excitation energy of the final nucleus. (A reduced mass can be used to allow for recoil of the nucleus.) Note that alpha decay cannot occur if E is negative; the kinetic energy of the alpha particle cannot be negative.

It may be noted that the energy release E in a nuclear process is generally called the "Q-value." The reason is that one of the most common other quantities used in nuclear physics is the so-called quadrupole moment Q. Also, the total nuclear charge is indicated by Q, as is the quality factor of radiation, while projection and rotation operators, and second points are often also indicted by Q. The underlying idea is that when you are trying to figure out some technical explanation, then if almost the only mathematical symbol used is Q, it provides a pretty strong hint that you are probably reading a book on nuclear physics.

The nuclear radius R approximately defines the start r_1 of the region that the alpha particle has to tunnel through, figure 14.11. It can be ballparked

reasonably well from the number of nucleons A; according to section 14.10,

$$R \approx R_A \sqrt[3]{A} \qquad R_A = 1.23 \text{ fm}$$

where f, femto, is 10^{-15}. That is a lot smaller than the typical Bohr radius over which electrons are spread out. Electrons are "far away" and are not really relevant.

It should be pointed out that the results are very sensitive to the assumed value of r_1. The simplest assumption would be that at r_1 the alpha particle would have its center at the nuclear radius of the remaining nucleus, computed from the above expression. But very noticeable improvements are obtained by assuming that at r_1 the center is already half the radius of the alpha particle outside. (In literature, it is often assumed that the alpha particle is a full radius outside, which means fully outside but still touching the remaining nucleus. However, half works better and is maybe somewhat less implausible.)

The good news about the sensitivity of the results on r_1 is that conversely it makes alpha decay a reasonably accurate way to deduce or verify nuclear radii, [31, p. 57]. You are hardly likely to get the nuclear radius noticeably wrong without getting into major trouble explaining alpha decay.

The number of escape attempts per unit time is also needed. If the alpha particle has a typical velocity v_α inside the original nucleus, it will take it a time of about $2r_0/v_\alpha$ to travel the $2r_0$ diameter of the nucleus. So it will bounce against the barrier about $v_\alpha/2r_0$ times per second. That is sure to be a very large number of times per second, the nucleus being so small, but each time it hits the perimeter, it only has a miniscule $e^{-2\gamma_{12}}$ chance of escaping. So it may well take trillions of years before it is successful anyway. Even so, among a very large number of nuclei a few will get out every time. Remember that a mol of atoms represents in the order of 10^{23} nuclei; among that many nuclei, a few alpha particles are likely to succeed whatever the odds against. The relative fraction of successful escape attempts per unit time is by definition the reciprocal of the lifetime τ;

$$\frac{v_\alpha}{2r_0} e^{-2\gamma_{12}} = \frac{1}{\tau} \tag{14.13}$$

Multiply the lifetime by $\ln 2$ to get the half-life.

The velocity v_α of the alpha particle can be ballparked from its kinetic energy $E - V_n$ in the nucleus as $\sqrt{2(E - V_n)/m_\alpha}$. Unfortunately, finding an accurate value for the nuclear potential V_n inside the nucleus is not trivial. But have another look at figure 14.10. Forget about engineering ideas about acceptable accuracy. A 50% error in half-life would be invisible seen on the tremendous range of figure 14.10. Being wrong by a factor 10, or even a factor 100, two orders of magnitude, is ho-hum on the scale that the half-life varies. So, the potential energy V_n inside the nucleus can be ballparked. The current results use the typical value of -35 MeV given in [31, p. 252].

That leaves the value of γ_{12} to be found from the integral over the barrier in (14.12). Because the nuclear forces are so short-range, they should be negligible over most of the integration range. So it seems reasonable to simply substitute the Coulomb potential everywhere for V. The Coulomb potential is inversely proportional to the radial position r, and it equals E at r_2, so V can be written as $V = Er_2/r$. Substituting this in, and doing the integral by making a change of integration variable to u with $r = r_2 \sin^2 u$, produces

$$\gamma_{12} = \frac{\sqrt{2m_\alpha E}}{\hbar} r_2 \left[\frac{\pi}{2} - \sqrt{\frac{r_1}{r_2}\left(1 - \frac{r_1}{r_2}\right)} - \arcsin\sqrt{\frac{r_1}{r_2}} \right]$$

The last two terms within the square brackets are typically relatively small compared to the first one, because r_1 is usually fairly small compared to r_2. Then γ_{12} is about proportional to $\sqrt{E} r_2$. But r_2 itself is inversely proportional to E, because the total energy of the alpha particle equals its potential energy at r_2;

$$E = \frac{(Z - Z_\alpha)e\, Z_\alpha e}{4\pi\epsilon_0 r_2} \qquad Z_\alpha = 2$$

That makes γ_{12} about proportional to $1/\sqrt{E}$ for a given atomic number Z.

So if you plot the half-life on a logarithmic scale, and the energy E on an reciprocal square root scale, as done in figure 14.10, they should vary linearly with each other for a given atomic number. This does assume that the variations in number of escape attempts are also ignored. The predicted slope of linear variation is indicated by the "stems" on the cherries in figure 14.10. Ideally, all cherries connected by branches should fall on a single line with this slope. The figure shows that this is quite reasonable for even-even nuclei, considering the rough approximations made. For nuclei that are not even-even, the deviations from the predicted slope are more significant. The next subsection discusses the major sources of error.

The bottom line question is whether the theory, rough as it may be, can produce meaningful values for the experimental half-lives, within reason. Figure 14.12 shows predicted half-lives versus the actual ones. Cherries on the black line indicate that the correct value is predicted. It is clear that there is no real accuracy to the predictions in any normal sense; they are easily off by several orders of magnitude. What can you expect without an accurate model of the nucleus itself? However, the predictions do successfully reproduce the tremendous range of half-lives and they do not deviate from the correct values that much compared to that tremendous range. It is hard to imagine any other theory besides tunneling that could do the same.

The worst performance of the theory is for the ${}^{209}_{83}$Bi bismuth isotope indicated by the rightmost dot in figure 14.12. Its true half-life of 19 Ey, 19 10^{18} years, is grossly underestimated to be just 9 Py, 9 10^{15} years. Then again, since the universe has only existed about 14 10^9 years, who is going to live long

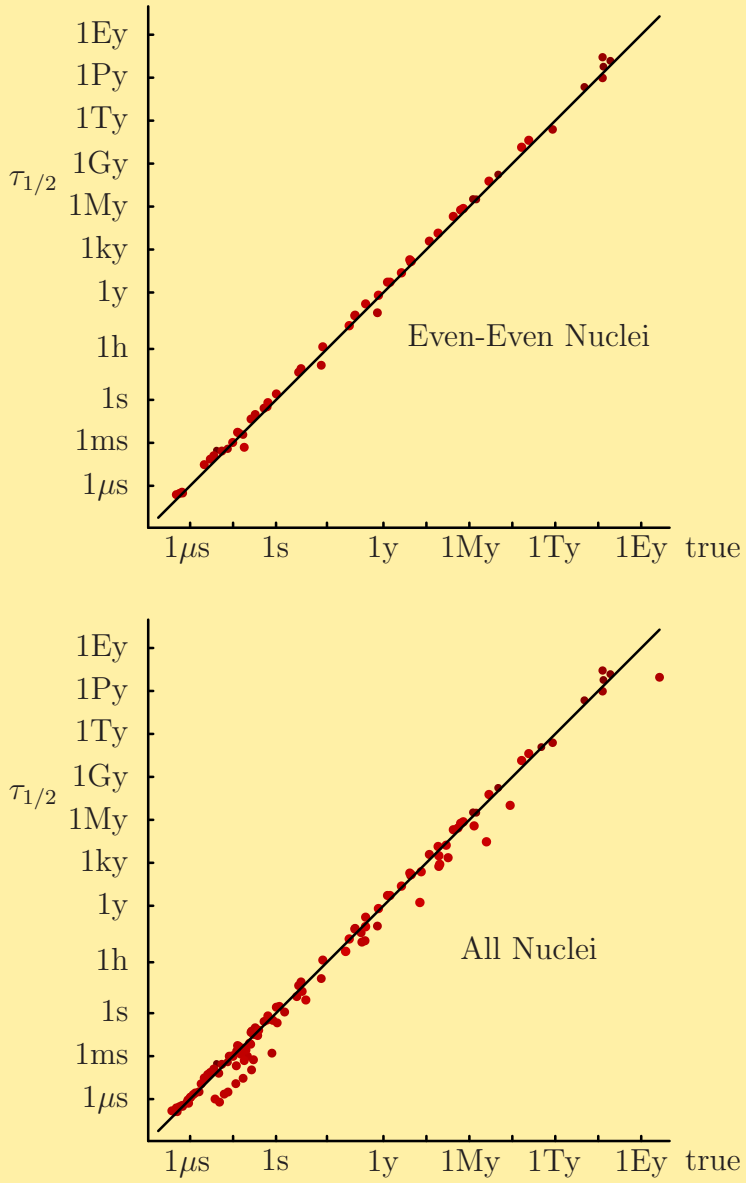


Figure 14.12: Half-life predicted by the Gamow / Gurney & Condon theory versus the true value. Top: even-even nuclei only. Bottom: all the nuclei except $^{8}_{4}$Be (55 as versus 67 as). [pdf]

enough to complain about it? Essentially none of the bismuth-209 that has ever been created in the universe has decayed. It took until 2003 for physicists to observe that bismuth-209 actually did decay; it is still listed as stable in many references. For ${}^{8}_{4}\mathrm{Be}$, which is not shown in the figure, the predicted half-life is 55 as ($55\ 10^{-18}$ s), versus a true value of 67 as.

14.11.3 Draft: Forbidden decays

You may wonder why there is so much error in the theoretical predictions of the half life. Or why the theory seems to work so much better for even-even nuclei than for others. A deviation by a factor 2 000 like for bismuth-209 seems an awful lot, rough as the theory may be.

Some of the sources of inaccuracy are self-evident from the theoretical description as given. In particular, there is the already mentioned effect of the value of r_1. It is certainly possible to correct for deviations from the Coulomb potential near the nucleus by a suitable choice of the value of r_1. However, the precise value that kills off the error is unknown, and unfortunately the results strongly depend on that value. To fix this would require an accurate evaluation of the nuclear force potential, and that is very difficult. Also, the potential of the electrons would have to be included. The alpha particle does reach a distance of the order of a tenth of a Bohr radius from the nucleus at the end of tunneling. The Bohr radius is here taken to be based on the actual nuclear charge, not the hydrogen one.

Also, the picture of a relatively compact wave packet of the alpha particle "rattling around" assumes that that the size of that wave packet is small compared to the nucleus. That spatial localization is associated with increased uncertainty in momentum, which implies increased energy. And the kinetic energy of the alpha particle is not really known anyway, without an accurate value for the nuclear force potential.

A very major other problem is the assumption that the final alpha particle and nucleus end up in their ground states. If either ends up in an excited state, the energy that the alpha particle has available for escape will be correspondingly reduced. Now the alpha particle will most certainly come out in its ground state; it takes over 20 MeV to excite an alpha particle. But for most nuclei, the remaining nucleus *cannot* be in its ground state if the mechanism is as described.

The main reason is angular momentum conservation. The alpha particle has no net internal angular angular momentum. Also, it was assumed that the alpha particle comes out radially, which means that there is no orbital angular momentum either. So the angular momentum of the nucleus after emission must be the same as that of the nucleus before the emission. That is no problem for even-even nuclei, because it is the same; even-even nuclei all have zero internal

angular momentum in their ground state. So even-even nuclei do not suffer from this problem.

However, almost all other nuclei do. All even-odd and odd-even nuclei and almost all odd-odd ones have nonzero angular momentum in their ground state. Usually the initial and final nuclei have different values. That means that alpha decay that leaves the final nucleus in its ground state violates conservation of angular momentum. The decay process is called "forbidden." The final nucleus must be excited if the process is as described. That energy subtracts from that of the alpha particle. Therefore the alpha particle has less energy to tunnel through, and the true half-life is much longer than computed.

Note in the bottom half of figure 14.12 how many nuclei that are not even-even do indeed have half-lifes that are orders of magnitude larger than predicted by theory. Consider the example of bismuth-209, with a half-life 2 000 times longer than predicted. Bismuth-209 has a spin, i.e. an azimuthal quantum number, of $\frac{9}{2}$. However, the decay product thallium-205 has spin $\frac{1}{2}$ in its ground state. If you check out the excited states of thallium-205, there is an excited state with spin $\frac{9}{2}$, but its excitation energy would reduce the energy of the alpha particle from 3.2 MeV to 1.7 MeV, making the tunneling process very much slower.

And there is another problem with that. The decay to the mentioned exited state is not possible either, because it violates conservation of parity, chapter 7.3 and 7.4. Saying "the alpha particle comes out radially," as done above is not really correct. The proper quantum way to say that the alpha particle comes out with no orbital angular momentum is to say that its wave function varies with angular location as the spherical harmonic Y_0^0, chapter 4.2.3. In spectroscopic terms, it "comes out in an s-wave." Now the initial bismuth atom has odd parity; its complete wave function changes sign if you everywhere replace $\vec{r}$ by $-\vec{r}$. But the alpha particle, the excited thallium state, and the Y_0^0 orbital motion all have even parity; there is no change of sign. That means that the total final parity is even too, so the final parity is not the same as the initial parity. That violates conservation of parity so the process cannot occur.

Thallium-205 does not have excited states below 3.2 MeV that have been solidly established to have spin $\frac{9}{2}$ and odd parity, so you may start to wonder whether alpha decay for bismuth-209 is possible at all. However, the alpha particle could of course come out with orbital angular momentum. In other words it could come out with a wave function that has an angular dependence according to Y_l^m with the azimuthal quantum number l equal to one or more. These states have even parity if l is even and odd parity when l is odd. Quantum mechanics then allows the thallium-205 excited state to have any spin j in the range $|\frac{9}{2} - l| \leqslant j \leqslant \frac{9}{2} + l$ as long as its parity is odd or even whenever l is even or odd.

For example, bismuth-209 could decay to the ground state of thallium-205 if the orbital angular momentum of the alpha particle is $l = 5$. Or it could decay

to an excited $^7/_2^+$ state with an excitation energy of 0.9 MeV if $l = 1$. The problem is that the kinetic energy in the angular motion subtracts from that available for the radial motion, making the tunneling, once again, much slower. In terms of the radial motion, the angular momentum introduces an additional effective potential $l(l+1)\hbar^2/2m_\alpha r^2$, compare the analysis of the hydrogen atom in chapter 4.3.2. Note that this effect increases rapidly with l. However, the decay of bismuth-209 appears to be to the ground state anyway; the measured energy of the alpha particle turns out to be 3.14 MeV. The predicted half-life including the effective potential is found to be 4.6 Ey, much better than the one computed in the previous section.

One final source of error should be mentioned. Often alpha decay can proceed in a number of ways and to different final excitation energies. In that case, the specific decay rates must be added together. This effect can make the true half-life shorter than the one computed in the previous subsection. But clearly, this effect should be minor on the scale of half-lifes of figure 14.12. Indeed, while the predicted half-lifes of many nuclei are way below the true value in the figure, few are significantly above it.

14.11.4 Draft: Why alpha decay?

The final question that begs an answer is why do so many nuclei so specifically want to eject an helium-4 nucleus? Why none of the other nuclei? Why not the less tightly bound, but lighter deuteron, or the more tightly bound, but heavier carbon-12 nucleus? The answer is subtle.

To understand the reason, reconsider the analysis of the previous subsection for a more general ejected nucleus. Assume that the ejected particle has an atomic number Z_1 and mass m_1. As mentioned, the precise number of escape attempts is not really that important for the half life; almost all the variation in half-life is through the quantity γ_{12}. Also, to a first approximation the ratio of start to end of the tunneling domain, r_1/r_2, can be ignored. Under those conditions, γ_{12} is proportional to

$$\gamma_{12} \propto \sqrt{\frac{m_1}{E}} Z_1(Z - Z_1)$$

It is pretty much all in there.

As long as the ejected particle has about the usual 8 MeV binding energy per nucleon, the square root in the expression above does not vary that much. In such cases the energy release E is about proportional to the amount of nucleons ejected. Table 14.3 gives some example numbers. That makes γ_{12} about proportional to Z_1, and the greatest chance of tunneling out then occurs by far for the lightest nuclei. It explains why the alpha particle tunnels out instead of heavier nuclei. It is not that a heavier nucleus like carbon-14 *cannot* be emitted, it is just that an alpha particle has already done so long before

carbon-14 gets the chance. In fact, for radium-223 it has been found that one carbon-14 nucleus is ejected for every billion alpha particles. That is about consistent with the computed half-lifes of the events as shown in table 14.3.

But the argument that Z_1 should be as small as possible should make protons or neutrons, not the alpha particle, the ones that can escape most easily. However, these do not have any binding energy. While protons or neutrons are indeed ejected from nuclei that have a very large proton, respectively neutron excess, normally the energy release for such emissions is negative. Therefore the emission cannot occur. Beta decay occurs instead to adjust the ratio between protons and neutrons to the optimum value. Near the optimum value, you would still think it might be better to eject a deuteron than an alpha. However, because the binding energy of the deuteron is only a single MeV per nucleon, the energy release is again negative. Among the light nuclei, the alpha is unique in having almost the full 8 MeV of binding energy per nucleon. It is therefore the only one that produces a positive energy release.

The final problem is that the arguments above seem to show that spontaneous fission cannot occur. For, is the fission of say fermium-256 into two tin-128 nuclei not just ejection of a tin-128 nucleus, leaving a tin-128 nucleus? The arguments above say that alpha decay should occur much before this can happen.

The problem is that the analysis of alpha decay is inapplicable to fission. The numbers for fission-scale half-lifes in table 14.3 are all wrong. Fission is indeed a tunneling event. However, it is one in which the energy barrier is disintegrating due to a global instability of the nuclear shape. That instability mechanism strongly favors large scale division over short scale ones. The only hint of this in table 14.3 are the large values of r_1/r_2 for fission-scale events. When r_1/r_2 becomes one, the tunneling region is gone. But long before that happens, the region is so small compared to the size of the ejected nucleus that the basic ideas underlying the analysis have become meaningless. Even ignoring the fact that the nuclear shapes have been assumed spherical and they are not in fission.

Thus, unlike table 14.3 suggests, fermium-256 does fission. The two fragments are usually of different size, but not vastly so. About 92% of fermium-256 nuclei spontaneously fission, while the other 8% experience alpha decay. Uranium-238 decays for 99.999 95% through α decay, and for only 0.000 05% through spontaneous fission. Although the amount of fission is very small, it is not by far as small as the numbers in table 14.3 imply. Fission is not known to occur for radium-223; this nucleus does indeed show pure alpha decay except for the mentioned rare carbon-14 emission.

$^{238}_{92}$U with $\tau_{1/2} = 1.4\ 10^{17}$ s								
Ejected:	$^{2}_{1}$H	$^{3}_{1}$H	$^{3}_{2}$He	$^{4}_{2}$He	$^{8}_{4}$Be	$^{16}_{6}$C	$^{20}_{8}$O	$^{118}_{46}$Pd
E, MeV:	−11.2	−10.0	−11.8	4.3	7.9	17.4	35.3	193.4
E/A_1:	−5.6	−3.3	−3.9	1.1	1.0	1.1	1.8	1.6
r_1/r_2:	–	–	–	0.14	0.14	0.21	0.33	0.58
γ^*_{12}:	–	–	–	85	172	237	239	580
γ_{12}:	–	–	–	45	92	103	74	79
$\tau_{1/2}$, s:	∞	∞	∞	$4\ 10^{17}$	$9\ 10^{58}$	$2\ 10^{68}$	$8\ 10^{44}$	$2\ 10^{47}$
$^{223}_{88}$Ra with $\tau_{1/2} = 9.9\ 10^{5}$ s								
Ejected:	$^{2}_{1}$H	$^{3}_{1}$H	$^{3}_{2}$He	$^{4}_{2}$He	$^{8}_{4}$Be	$^{14}_{6}$C	$^{18}_{8}$O	$^{97}_{40}$Zr
E, MeV:	−9.2	−9.2	−8.3	6.0	12.9	31.9	40.4	172.9
E/A_1:	−4.6	−3.1	−2.8	1.5	1.6	2.3	2.2	1.8
r_1/r_2:	–	–	–	0.20	0.23	0.40	0.39	0.56
γ^*_{12}:	–	–	–	69	129	156	203	534
γ_{12}:	–	–	–	32	53	40	53	77
$\tau_{1/2}$, s:	∞	∞	∞	$9\ 10^{4}$	$4\ 10^{24}$	$9\ 10^{12}$	$2\ 10^{24}$	$2\ 10^{45}$
$^{256}_{100}$Fm with $\tau_{1/2} = 9.5\ 10^{3}$ s								
Ejected:	$^{2}_{1}$H	$^{3}_{1}$H	$^{3}_{2}$He	$^{4}_{2}$He	$^{8}_{4}$Be	$^{14}_{6}$C	$^{20}_{8}$O	$^{128}_{50}$Sn
E, MeV:	−9.6	−8.5	−8.7	7.1	13.2	27.9	39.4	252.8
E/A_1:	−4.9	−2.8	−2.9	1.8	1.6	2.0	2.0	2.0
r_1/r_2:	–	–	–	0.22	0.22	0.31	0.35	0.65
γ^*_{12}:	–	–	–	72	145	192	249	622
γ_{12}:	–	–	–	31	63	63	74	61
$\tau_{1/2}$, s:	∞	∞	∞	$2\ 10^{5}$	$8\ 10^{32}$	$6\ 10^{32}$	$6\ 10^{42}$	$2\ 10^{31}$

Table 14.3: Candidates for nuclei ejected by uranium-238, radium-223, and fermium-256.

14.12 Draft: Shell model

The liquid drop model gives a very useful description of many nuclear properties. It helps understand alpha decay quite well. Still, it has definite limitations. Quantum properties such as the stability of individual nuclei, spin, magnetic moment, and gamma decay can simply not be explained using a classical liquid model with a couple of simple fixes applied.

Historically, a major clue about a suitable quantum model came from the magic numbers. Nuclei tend to be unusually stable if the number of protons and/or neutrons is one of the

$$\text{magic numbers:} \quad 2, 8, 20, 28, 50, 82, 126, \ldots \tag{14.14}$$

The higher magic number values are quite clearly seen in proton pair and neutron pair removal graphs like figures 14.7 and 14.8 in section 14.8.

If an additional proton is added to a nucleus with a magic number of protons, or an additional neutron to a nucleus with a magic number of neutrons, then that additional nucleon is much more weakly bound.

The doubly magic ${}^4_2\mathrm{He}$ helium-4 nucleus, with 2 protons and 2 neutrons, is a good example. It has more than three times the binding energy of ${}^3_2\mathrm{He}$ helium-3, which merely has a magic number of protons. Still, if you try to add another proton or neutron to helium-4, it will not be bound at all, it will be ejected in less than 10^{-21} seconds.

That is very reminiscent of the electron structure of the helium *atom*. The two electrons in the helium atom are very tightly bound, making helium into an inert noble gas. In fact, it takes 25 eV of energy to remove an electron from a helium atom. However, for lithium, with one more electron, the third electron is very loosely bound, and readily given up in chemical reactions. It takes only 5.4 eV to remove the third electron from lithium. Similar effects appear for the other noble gasses, neon with 10 electrons, argon with 18, krypton with 36, etcetera. The numbers 2, 10, 18, 36, ..., are magic for electrons in atoms.

For atoms, the unusual stability could be explained in chapter 5.9 by ignoring the direct interactions between electrons. It was assumed that for each electron, the complicated effects of all the other electrons could be modeled by some average potential that the electron moves in. That approximation produced *single-electron* energy eigenfunctions for the electrons. They then had to occupy these single-electron states one by one on account of Pauli's exclusion principle. Noble gasses completely fill up an energy level, requiring any additional electrons to go into the next available, significantly higher energy level. That greatly decreases the binding energy of these additional electrons compared to those already there.

The similarity suggests that the protons and neutrons in nuclei might be described similarly. There are now two types of particles but in the approximation that each particle is not directly affected by the others it does not make

much of a difference. Also, antisymmetrization requirements only apply when the particles are identical, either both protons or both neutrons. Therefore, protons and neutrons can be treated completely separately. Their interactions occur only indirectly through whatever is used for the average potential that they move in. The next subsections work out a model along these lines.

14.12.1 Draft: Average potential

The first step will be to identify a suitable average potential for the nucleons. One obvious difference distinguishing nuclei from atoms is that the Coulomb potential is not going to hack it. In the electron structure of an atom the electrons repel each other, and the only reason the atom stays together is that there is a nucleus to attract the electrons. But inside a nucleus, the nucleons all attract each other and there is no additional attractive core. Indeed, a Coulomb potential like the one used for the electrons in atoms would get only the first magic number, 2, right, predicting 10, instead of 8, total particles for a filled second energy level.

A better potential is needed. Now in the center of a nucleus, the attractive forces come from all directions and the net force will be zero by symmetry. Away from the center, the net force will be directed inwards towards the center to keep the nucleons together inside the nucleus. The simplest potential that describes this is the harmonic oscillator one. For that potential, the inward force is simply proportional to the distance from the center. That makes the potential energy V proportional to the square distance from the center, as sketched in figure 14.13*a*.

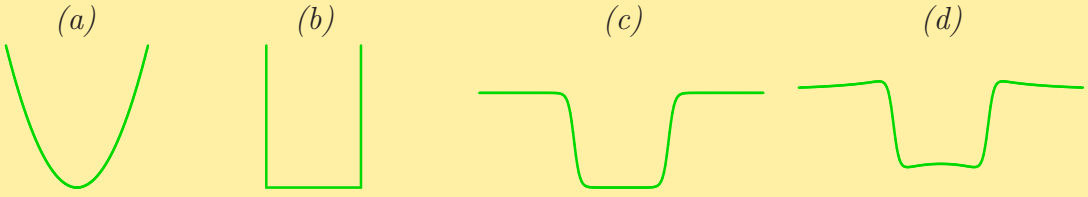


Figure 14.13: Example average nuclear potentials: (a) harmonic oscillator, (b) impenetrable surface, (c) Woods-Saxon, (d) Woods-Saxon for protons.

The energy eigenvalues of the harmonic oscillator are

$$E_n = \left(n + \tfrac{1}{2}\right)\hbar\omega \qquad n = 1, 2, 3, \ldots \tag{14.15}$$

Also, in spherical coordinates the energy eigenfunctions of the harmonic oscillator can be taken to be of the form, {D.76},

$$\psi^{\rm ho}_{nlmm_s} = R^{\rm ho}_{nl}(r) Y_l^m(\theta,\phi)\updownarrow \qquad \begin{array}{l} l = n{-}1, n{-}3, \ldots \geqslant 0 \\ m = -l, -l{+}1, \ldots, l{-}1, l \\ m_s = \pm\tfrac{1}{2} \end{array} \tag{14.16}$$

Here l is the azimuthal quantum number that gives the square orbital angular momentum of the state as $l(l+1)\hbar^2$; m is the magnetic quantum number that gives the orbital angular momentum in the direction of the arbitrarily chosen z-axis as $m\hbar$, and m_s is the spin quantum number that gives the spin angular momentum of the nucleon in the z-direction as $m_s\hbar$. The "spin-up" state with $m_s = \frac{1}{2}$ is commonly indicated by a postfix $\uparrow$, and similarly the spin-down one $m_s = -\frac{1}{2}$ by $\downarrow$. The details of the functions $R^{\rm ho}_{nl}$ and Y_l^m are of no particular interest.

(It may be noted that the above spherical eigenfunctions are different from the Cartesian ones derived in chapter 4.1, except for the ground state. However, the spherical eigenfunctions at a given energy level can be written as combinations of the Cartesian ones at that level, and vice-versa. So there is no fundamental difference between the two. It just works out that the spherical versions are much more convenient in the rest of the story.)

Compared to the Coulomb potential of the hydrogen electron as solved in chapter 4.3, the major difference is in the number of energy states at a given energy level n. While for the Coulomb potential the azimuthal quantum number l can have any value from 0 to $n-1$, for the harmonic oscillator l must be odd or even depending on whether $n-1$ is odd or even.

It does not make a difference for the lowest energy level $n = 1$; in that case only $l = 0$ is allowed for either potential. And since the number of values of the magnetic quantum number m at a given value of l is $2l+1$, there is only one possible value for m. That means that there are only two different energy states at the lowest energy level, corresponding to $m_s = \frac{1}{2}$ respectively $-\frac{1}{2}$. Those two states explain the first magic number, 2. Two nucleons of a given type can occupy the lowest energy level; any further ones of that type must go into a higher level.

In particular, helium-4 has the lowest energy level for protons completely filled with its two protons, and the lowest level for neutrons completely filled with its two neutrons. That makes helium-4 the first doubly-magic nucleus. It is just like the two electrons in the helium *atom* completely fill the lowest energy level for electrons, making helium the first noble gas.

At the second energy level $n = 2$, where the Coulomb potential allows both $l = 0$ and $l = 1$, only $l = 1$ is allowed for the harmonic oscillator. So the number of states available at energy level $n = 2$ is less than that of the Coulomb potential. In particular, the azimuthal quantum number $l = 1$ allows $2l + 1 = 3$ values of the magnetic quantum number m, times 2 values for the spin quantum number m_s. Therefore, $l = 1$ at $n = 2$ corresponds to 3 times 2, or 6 energy states. Combined with the two $l = 0$ states at energy level $n = 1$, that gives a total of 8. The second magic number 8 has been explained! It requires 8 nucleons of a given type to fill the lowest two energy levels.

It makes oxygen-16 with 8 protons and 8 neutrons the second doubly-magic nucleus. Note that for the electrons in atoms, the second energy level would

also include two $l = 0$ states. That is why the second noble gas is neon with 10 electrons, and not oxygen with 8.

Before checking the other magic numbers, first a problem with the above procedure of counting states must be addressed. It is too easy. Everybody can evaluate $2l + 1$ and multiply by 2 for the spin states! To make it more challenging, physicists adopt the so-called "spectroscopic notation" in which they do not tell you the value of l. Instead, they tell you a letter like maybe p, and you are then supposed to figure out yourself that $l = 1$. The scheme is:

$$\text{s, p, d, f, g, h, i, [j], k, } \ldots \qquad \Longrightarrow \qquad l = 0, 1, 2, 3, 4, 5, 6, 7, 8, \ldots$$

The latter part is mostly alphabetic, but by convention j is not included. However, my references on nuclear physics *do* include j; that is great because it provides additional challenge. Using spectroscopic notations, the second energy level states are renotated as

$$\psi_{21mm_s} \qquad \Longrightarrow \qquad 2p$$

where the 2 indicates the value of n giving the energy level. The additional dependence on the magnetic quantum numbers m and m_s is kept hidden from the uninitiated. (To be fair, as long as there is no preferred direction to space, these quantum numbers are physically not of importance. If an external magnetic field is applied, it provides directionality, and magnetic quantum numbers do become relevant.)

However, physicists figured that this would not provide challenge enough, since most students already practiced it for atoms. The above notation follows the one that physicists use for atoms. In this notation, the leading number is n, the energy level of the simplest theoretical model. To provide more challenge, for nuclei physicist replace the leading number with a count of states at that angular momentum. For example, physicists denote 2p above by 1p, because it is the lowest energy p states. Damn what theoretical energy level it is. For still more challenge, while most physicists start counting from one, some start from zero, making it 0p. However, since it gives the author of this book a headache to count angular momentum states upwards between shells, this book will mostly follow the atomic convention, and the leading digit will indicate n, the harmonic oscillator energy level. The "official" eigenfunction designations will be listed in the final results where appropriate. Most but not all references will follow the official designations.

In these terms, the energy levels and numbers of states for the harmonic oscillator potential are as shown in figure 14.14. The third energy level has 2 3s states and 10 3d states. Added to the 8 from the first two energy levels, that brings the total count to 20, the third magic number.

Unfortunately, this is where it stops. The fourth energy level should have only 8 states to reach the next magic number 28, but in reality the fourth

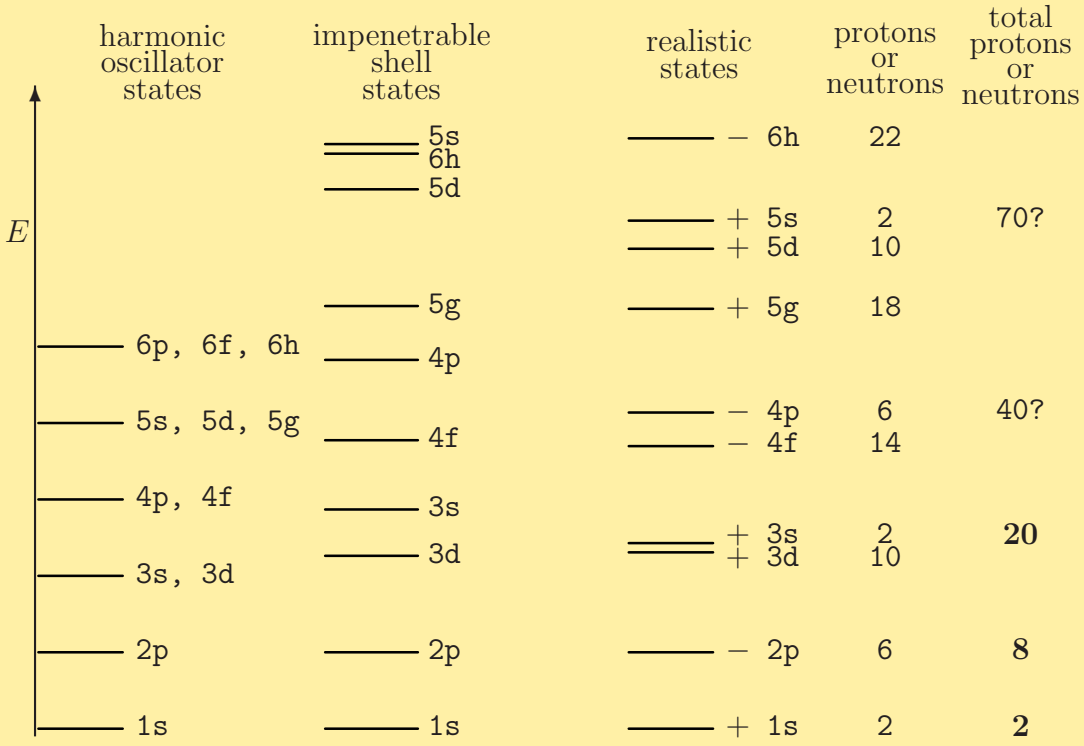


Figure 14.14: Nuclear energy levels for various assumptions about the average nuclear potential. The signs indicate the parity of the states.

harmonic oscillator level has 6 4p states and 14 4f ones. Still, getting 3 magic numbers right seems like a good start.

The logical next step is to try to improve upon the harmonic oscillator potential. In an average nucleus, it can be expected that the net force on a nucleon pretty much averages out to zero everywhere except in a very thin layer at the outer surface. The reason is that the nuclear forces are very short range; therefore the forces seem to come equally from all directions unless the nucleon is very close to the surface. Only right at the surface do the particles experience a net inward attraction because of the deficit of particles beyond the surface to provide the full compensating outward force. This suggests a picture in which the nucleons do not experience a net force within the confines of the nucleus. However, at the surface, the potential ramps up very steeply. As an idealization the potential beyond the surface can be taken infinite.

That reasoning results in the "impenetrable-shell" potential shown in figure 14.13. It too is analytically solvable, {D.77}. The energy levels are shown in figure 14.14. Unfortunately, it does not help any explaining the fourth magic number 28.

It does help understand why the shell model works at all, ⟦15⟧. That is not at all obvious; for a long time physicists really believed it would not work. For the electrons in an atom, the nucleus at least produces *some* potential that is independent of the relative positions of the electrons. In a nucleus, there is nothing: the potential experienced by the nucleons is completely dependent on *relative* nucleon positions and spins. So what reasonable justification could there possibly be to assume that the nucleons act as if they move in an average potential that is independent of the other nucleons? However, first assume that the only potential energy is the one that keeps the nucleons within the experimental nuclear radius. That is the impenetrable shell model. In that case, the energy eigenfunctions are purely kinetic energy ones, and these have a shell structure. Now restore the actual complex interactions between nucleons. You would at first guess that these should greatly change the energy eigenstates. But if they really do that, it would bring in large amounts of unoccupied kinetic energy states. That would produce a significant increase in kinetic energy, and that is not possible because the binding energy is fairly small compared to the kinetic energy. In particular, therefore, removing the last nucleon should not require an energy very different from a shell model value regardless of however complex the true potential energy really is.

Of course, the impenetrable-shell potential too is open to criticism. A nucleus has maybe ten nucleons along a diameter. Surely the thickness of the surface layer cannot reasonably be much less than the spacing between nucleons. Or much less than the range of the nuclear forces, for that matter. Also, the potential should not be infinite outside the nucleus; nucleons do escape from, or enter nuclei without infinite energy. The truth is clearly somewhere in between the harmonic oscillator and impenetrable shell potentials. A more

realistic potential along such lines is the "Woods-Saxon" potential

$$V = -\frac{V_0}{1 + e^{(r-a)/d}} + \text{constant}$$

which is sketched in figure 14.13*c*. For protons, there is an additional repulsive Coulomb potential that will be maximum at the center of the sphere and decreases to zero proportional to $1/r$ outside the nucleus. That gives a combined potential as sketched in figure 14.13*d*. Note that the Coulomb potential is not short-range like the nucleon-nucleon attractions; its nontrivial variation is not just restricted to a thin layer at the nuclear surface.

Typical energy levels are sketched in figure 14.14. As expected, they are somewhere in between the extreme cases of the harmonic oscillator and the impenetrable shell.

The signs behind the realistic energy levels in 14.14 denote the predicted "parity" of the states. Parity is a very helpful mathematical quantity for studying nuclei. The parity of a wave function is "one," or "positive," or "even," if the wave function stays the same when the positive direction of the three Cartesian axes is inverted. That replaces every $\vec{r}$ in the wave function by $-\vec{r}$. The parity is "minus one,", or "negative," or "odd," if the wave function merely changes sign under an exes inversion. Parity is uncertain when the wave function changes in any other way; however, nuclei have definite parity as long as the weak force of beta decay does not play a role. It turns out that s, d, g, ... states have positive parity while p, f, h, ... states have negative parity, {D.14} or {D.76}. Therefore, the harmonic oscillator shells have alternatingly positive and negative parity.

For the wave functions of complete nuclei, the net parity is the product of the parities, (taking them to be one or minus one), of the individual nucleons. Now physicist can experimentally deduce the parity of nuclei in various ways. It turns out that the parities of the nuclei up to the third magic number agree perfectly with the values predicted by the energy levels of figure 14.14. (Only three unstable, artificially created, nuclei disagree.) It really appears that the model is onto something.

Unfortunately, the fourth magic number remains unexplained. In fact, any reasonable spherically symmetric spatial potential will not get the fourth magic number right. There are 6 4p states and 14 4f ones; how could the additional 8 states needed for the next magic number 28 ever be extracted from that? Twiddling with the shape of a purely spatial potential is not enough.

14.12.2 Draft: Spin-orbit interaction

Eventually, Mayer in the U.S., and independently Jensen and his co-workers in Germany, concluded that spin had to be involved in explaining the magic numbers above 20. To understand why, consider the six 4p and fourteen 4f energy states at the fourth energy level of the harmonic oscillator model. Clearly,

the six 4p states cannot produce the eight states of the energy shell needed to explain the next magic number 28. And neither can the fourteen 4f states, unless for some reason they split into two different groups whose energy is no longer equal.

Why would they split? In nonquantum terms, all fourteen states have orbital and spin angular momentum vectors of exactly the same lengths. What is different between states is only the direction of these vectors. And the absolute directions cannot be relevant since the physics cannot depend on the orientation of the axis system in which it is viewed. What it can depend on is the relative alignment between the orbital and spin angular momentum vectors. This relative alignment is characterized by the dot product between the two vectors.

Therefore, the logical way to get an energy splitting between states with differently aligned orbital and spin angular momentum is to postulate an additional contribution to the Hamiltonian of the form

$$\Delta H \propto -\widehat{\vec{L}} \cdot \widehat{\vec{S}}$$

Here $\widehat{\vec{L}}$ is the orbital angular momentum vector and $\widehat{\vec{S}}$ the spin one. A contribution to the Hamiltonian of this type is called an "spin-orbit" interaction, because it couples spin with orbital angular momentum. Spin-orbit interaction was already known from improved descriptions of the energy levels of the hydrogen atom, addendum {A.39}. However, that electromagnetic effect is far too small to explain the observed spin-orbit interaction in nuclei. Also, it would get the sign of the correction wrong for neutrons.

While nuclear forces remain incompletely understood, there is no doubt that it is these much stronger forces, and not electromagnetic ones, that provide the mechanism. Still, in analogy to the electronic case, the constant of proportionality is usually taken to include the net force $\partial V/\partial r$ on the nucleon and an additional factor $1/r$ to turn orbital momentum into velocity. None of that makes a difference for the harmonic oscillator potential, for which the net effect is still just a constant. Either way, next the strength of the resulting interaction is adjusted to match the experimental energy levels.

To correctly understand the effect of spin-orbit interaction on the energy levels of nucleons is not quite trivial. Consider the fourteen ψ_{43mm_s} 4f states. They have orbital angular momentum in the chosen z-direction $m\hbar$, with $m =$ -3,-2,-1,0,1,2,3, and spin angular momentum $m_s\hbar$ with $m_s = \pm\frac{1}{2}$. Naively, you might assume that the spin-orbit interaction lowers the energy of the six states for which m and m_s have the same sign, raises it for the six where they have the opposite sign, and leaves the energy of the two states with $m = 0$ the same. That is not true. The problem is that the spin-orbit interaction $\widehat{\vec{L}} \cdot \widehat{\vec{S}}$ involves $\widehat{L}_x$ and $\widehat{L}_y$, and these do not commute with $\widehat{L}_z$ regardless of how you orient the axis system. And the same for $\widehat{S}_x$ and $\widehat{S}_y$.

With spin-orbit interaction, energy eigenfunctions of nonzero orbital angular momentum no longer have definite orbital momentum L_z in a chosen z-direction. And neither do they have definite spin S_z in such a direction.

Therefore the energy eigenfunctions can no longer be taken to be of the form $R_{nl}(r)Y_l^m(\theta,\phi)\updownarrow$. They have uncertainty in both m and m_s, so they will be combinations of states $R_{nl}(r)Y_l^m(\theta,\phi)\updownarrow$ with varying values of m and m_s.

However, consider the *net* angular momentum operator

$$\widehat{\vec{J}} \equiv \widehat{\vec{L}} + \widehat{\vec{S}}$$

If you expand its square magnitude,

$$\widehat{J}^2 = (\widehat{\vec{L}} + \widehat{\vec{S}}) \cdot (\widehat{\vec{L}} + \widehat{\vec{S}}) = \widehat{L}^2 + 2\widehat{\vec{L}} \cdot \widehat{\vec{S}} + \widehat{S}^2$$

you see that the spin-orbit term can be written in terms of the square magnitudes of orbital, spin, and net angular momentum operators:

$$-\widehat{\vec{L}} \cdot \widehat{\vec{S}} = -\tfrac{1}{2}\left[\widehat{J}^2 - \widehat{L}^2 - \widehat{S}^2\right]$$

Therefore combination states that have definite square net angular momentum J^2 remain good energy eigenfunctions even in the presence of spin-orbit interaction.

Now a quick review is needed of the weird way in which angular momenta combine into net angular momentum in quantum mechanics, chapter 12.7. In classical mechanics, the sum of an angular momentum vector with length L and one with length S could have any combined length J in the range $|L-S| \leqslant J \leqslant L+S$, depending on the angle between the vectors. However, in quantum mechanics, the length of the final vector must be quantized as $\sqrt{j(j+1)}\hbar$ where the quantum number j must satisfy $|l-s| \leqslant j \leqslant l+s$ and must change in integer amounts. In particular, since the spin is given as $s = \frac{1}{2}$, the net angular momentum quantum number j can either be $l - \frac{1}{2}$ or $l + \frac{1}{2}$. (If l is zero, the first possibility is also ruled out, since square angular momentum cannot be negative.)

For the 4f energy level $l = 3$, so the square net angular momentum quantum number j can only be $\frac{5}{2}$ or $\frac{7}{2}$. And for a given value of j, there are $2j+1$ values for the quantum number m_j giving the net angular momentum in the chosen z-direction. That means that there are six states with $j = \frac{5}{2}$ and eight states with $j = \frac{7}{2}$. The total is fourteen, still the same number of independent states at the 4f level. In fact, the fourteen states of definite net angular momentum j can be written as linear combinations of the fourteen $R_{nl}Y_l^m\updownarrow$ states. (Figure 12.5 shows such combinations up to $l = 2$; item 2 in chapter 12.8 gives a general formula.) Pictorially,

$$7\ 4f\uparrow \text{ and } 7\ 4f\downarrow \text{ states} \qquad \Longrightarrow \qquad 6\ 4f_{5/2} \text{ and } 8\ 4f_{7/2} \text{ states}$$

where the spectroscopic convention is to show the net angular momentum j as a subscript for states in which its value is unambiguous.

The spin-orbit interaction raises the energy of the six $4f_{5/2}$ states, but lowers it for the eight $4f_{7/2}$ states. In fact, from above, for any state of definite square orbital and square net angular momentum,

$$-\widehat{\vec{L}}\cdot\widehat{\vec{S}} = -\tfrac{1}{2}\hbar^2[j(j+1) - l(l+1) - s(s+1)] = \begin{cases} \tfrac{1}{2}(l+1)\hbar^2 & \text{for } j = l - \tfrac{1}{2} \\ -\tfrac{1}{2}l\hbar^2 & \text{for } j = l + \tfrac{1}{2} \end{cases}$$

The eight $4f_{7/2}$ states of lowered energy form the shell that is filled at the fourth magic number 28.

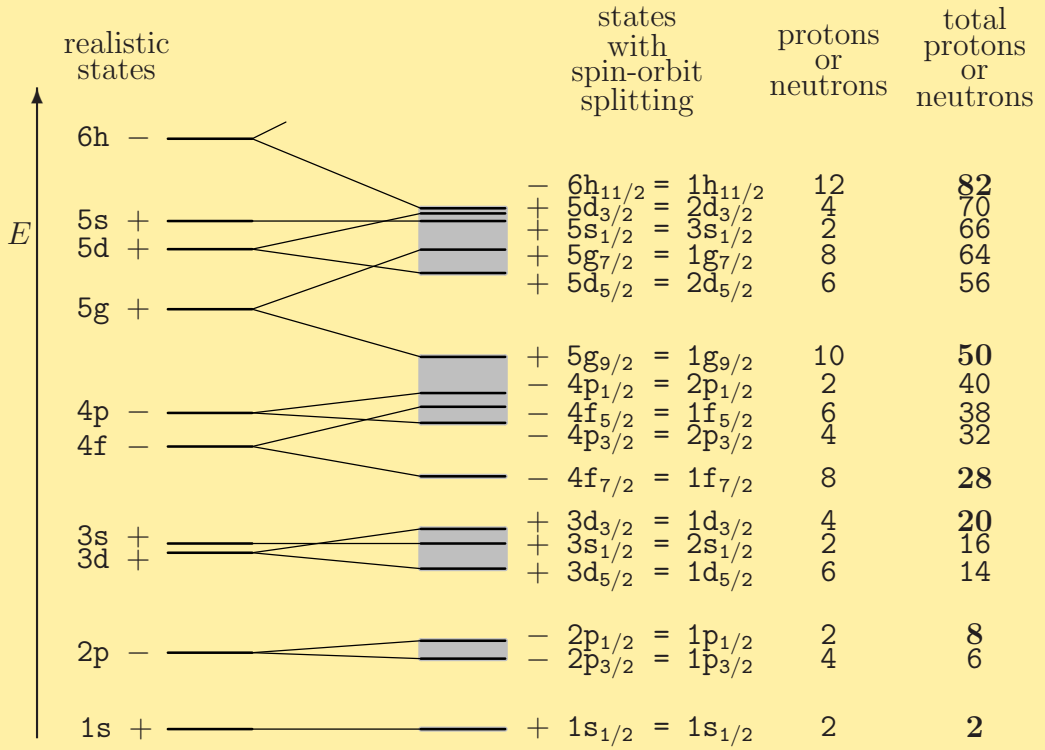


Figure 14.15: Schematic effect of spin-orbit interaction on the energy levels. The ordering within bands is realistic for neutrons. The designation behind the equals sign is the "official" one. (Assuming counting starts at 1).

Figure 14.15 shows how the spin-orbit splitting of the energy levels gives rise to the remaining magic numbers. In the figure, the coefficient of the spin orbit term was simply taken to vary linearly with the energy level n. The details depend on whether it is neutrons or protons, and may vary from nucleus to nucleus. Especially for the higher energy bands the Coulomb repulsion has an increasingly large effect on the energies of protons.

The major shells, terminated by magic numbers, are shown as grey bands. In the numbering system followed here, a subshell with a different number as

the others in the same major shell comes from a different harmonic oscillator energy level. Figure 14.15 also shows the "official" enumeration of the states. You be the judge which numbering system makes the most sense to you.

As sketched in figure 14.15, spin-orbit interaction pushes the $5g_{9/2}$ states down into the band that ends at magic number 50. However, the energy gap between between the $5g_{9/2}$ states and the $4\ldots$ states in the band is relatively large. That is why you might think of 40 as a semi-magic number if you want. For example, one good reason to consider this is figure 14.19 discussed later.

The detailed ordering of the subshells above 50 varies with author and even for a single author. There is no unique answer, because the shell model is only a simple approximation to a system that does not follow simple rules when examined closely enough. Still, a specific ordering must be adopted if the shell model is to be compared to the data. This book will use the orderings:

protons:

$$1s_{1/2}$$
$$2p_{3/2}\ 2p_{1/2}$$
$$3d_{5/2}\ 3s_{1/2}\ 3d_{3/2}$$
$$4f_{7/2}$$
$$4p_{3/2}\ 4f_{5/2}\ 4p_{1/2}\ 5g_{9/2}$$
$$5g_{7/2}\ 5d_{5/2}\ 6h_{11/2}\ 5d_{3/2}\ 5s_{1/2}$$
$$6h_{9/2}\ 6f_{7/2}\ 6f_{5/2}\ 6p_{3/2}\ 6p_{1/2}\ 7i_{13/2}$$

neutrons:

$$1s_{1/2}$$
$$2p_{3/2}\ 2p_{1/2}$$
$$3d_{5/2}\ 3s_{1/2}\ 3d_{3/2}$$
$$4f_{7/2}$$
$$4p_{3/2}\ 4f_{5/2}\ 4p_{1/2}\ 5g_{9/2}$$
$$5d_{5/2}\ 5g_{7/2}\ 5s_{1/2}\ 5d_{3/2}\ 6h_{11/2}$$
$$6f_{7/2}\ 6h_{9/2}\ 6p_{3/2}\ 6f_{5/2}\ 7i_{13/2}\ 6p_{1/2}$$
$$7g_{9/2}\ 7d_{5/2}\ 7i_{11/2}\ 7g_{7/2}\ 7s_{1/2}\ 7d_{3/2}8j_{15/2}$$

The ordering for protons follows [36, table 7-1], but not [36, p. 223], to Z=92, and then [31], whose table seems to come from Mayer and Jensen. The ordering for neutrons follows [36], with the subshells beyond 136 taken from ⟦10⟧. However, the $7i_{13/2}$ and $6p_{1/2}$ states were swapped since the shell filling [36, table 7-1] makes a lot more sense if you do. The same swap is also found in [40, p. 255], following Klinkenberg, while [31, p. 155] puts the $7i_{13/2}$ subshell even farther down below the $6p_{3/2}$ state.

14.12.3 Draft: Example occupation levels

The purpose of this section is to explore how the shell model works out for sample nuclei.

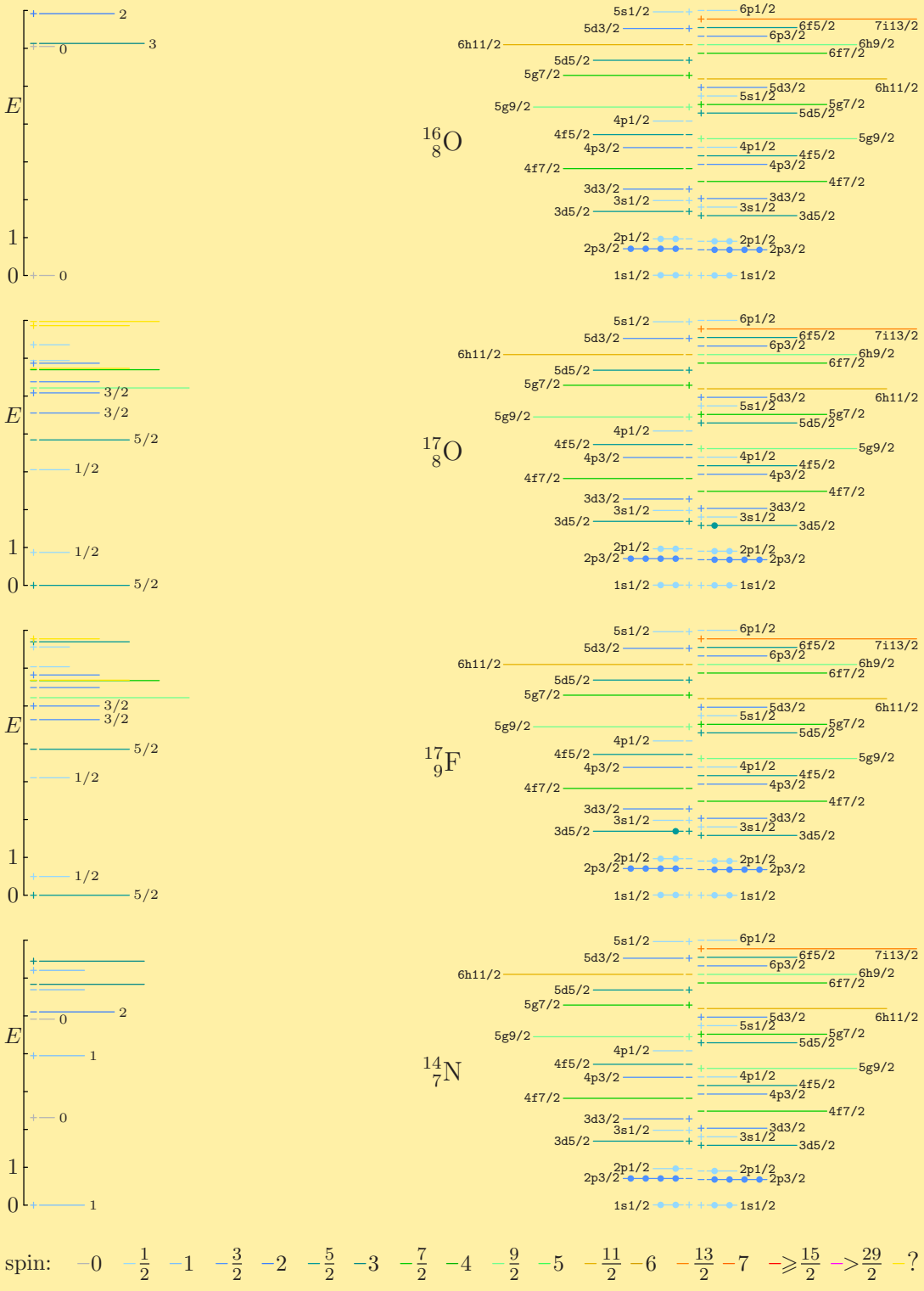


Figure 14.16: Energy levels for doubly-magic oxygen-16 and neighbors. [pdf]

Figure 14.16 shows experimental energy spectra of various nuclei at the left. The energy values are in MeV. The ground state is defined to be the zero level of energy. The length and color of the energy lines indicates the spin of the nucleus, and the parity is indicated by a plus or minus sign. Some important spin values are also listed explicitly. Yellow lines indicate states for which no unique spin and/or parity are determined or are established with reservations. At the right in the figure, a sketch of the occupation levels according to the shell model is displayed for easy reference.

The top of the figure shows data for oxygen-16, the normal oxygen that makes up 99.8% of the oxygen in the atmosphere. Oxygen-16 is a doubly-magic nucleus with 8 protons and 8 neutrons. As the right-hand diagram indicates, these completely fill up the lowest two major shells.

As the left-hand spectrum shows, the oxygen-16 nucleus has zero net spin in the ground state. That is exactly what the shell model predicts. In fact, it is a consequence of quantum mechanics that:

Completely filled subshells have zero net angular momentum.

Since the shell model says all shells are filled, the zero spin follows. The shell model got the first one right. Indeed, it passes this test with flying colors for all doubly-magic nuclei.

Next,

Subshells with an even number of nucleons have even parity.

That is just a consequence of the fact that even if the subshell is a negative parity one, negative parities multiply out pairwise to positive ones. Since all subshells of oxygen-16 contain an even number of nucleons, the combined parity of the complete oxygen-16 nucleus should be positive. It is. And it is for the other doubly-magic nuclei.

The shell model implies that a doubly-magic nucleus like oxygen-16 should be be particularly stable. So it should require a great deal of energy to excite it. Indeed it does: figure 14.16 shows that exciting oxygen-16 takes over 6 MeV of energy.

Following the shell model picture, one obvious way to excite the nucleus would be to kick a single proton or neutron out of the $2p_{1/2}$ subshell into the next higher energy $3d_{5/2}$ subshell. The net result is a nucleon with spin 5/2 in the $3d_{5/2}$ subshell and one remaining nucleon with spin 1/2 in the $2p_{1/2}$ subshell. Quantum mechanics allows these two nucleons to combine their spins into a net spin of either $\frac{5}{2}+\frac{1}{2}=3$ or $\frac{5}{2}-\frac{1}{2}=2$. In addition, since the nucleon kicked into the $3f_{5/2}$ changes parity, so should the complete nucleus. And indeed, there is an excited level a bit above 6 MeV with a spin 3 and odd parity, a 3^- level. It appears the shell model may be onto something.

Still, the exited 0^+ state suggests there may be a bit more to the story. In a shell model explanation, the parity of this state would require a pair of nucleons

to be kicked up. In the basic shell model, it would seem that this should require twice the energy of kicking up one nucleon. Not all nuclear excitations can be explained by the excitation of just one or two nucleons, especially if the mass number gets over 50 or the excitation energy high enough. This will be explored in section 14.13. However, before summarily dismissing a shell model explanation for this state, first consider the following sections on pairing and configuration mixing.

Next consider oxygen-17 and fluorine-17 in figure 14.16. These two are examples of so-called "mirror nuclei;" they have the numbers of protons and neutrons reversed. Oxygen-17 has 8 protons and 9 neutrons while its twin fluorine-17 has 9 protons and 8 neutrons. The similarity in energy levels between the two illustrates the idea of charge symmetry: nuclear forces are the same if the protons are turned into neutrons and vice versa. (Of course, this swap does mess up the Coulomb forces, but Coulomb forces are not very important for light nuclei.)

Each of these two nuclei has one more nucleon in addition to an oxygen-16 "core". Since the filled subshells of the oxygen-16 core have zero spin, the net nuclear spin should be that of the odd nucleon in the $3d_{5/2}$ subshell. And the parity should be even, since the odd nucleon is in an even parity shell. And indeed each ground state has the predicted spin of 5/2 and even parity. Chalk up another two for the shell model.

This is a big test for the shell model, because if a doubly-magic-plus-one nucleus did not have the predicted spin and parity of the final odd nucleon, there would be no reasonable way to explain it. Fortunately, all nuclei of this type pass the test.

For both oxygen-17 and fluorine-17, there is also a low-energy $\frac{1}{2}^+$ excited state, likely corresponding to kicking the odd nucleon up to the next minor shell, the $3s_{1/2}$ one. And so there is an excited $\frac{3}{2}^+$ state, for kicking up the nucleon to the $3d_{3/2}$ state instead.

However, from the shell model, in particular figure 14.15, you would expect the spacing between the $3d_{5/2}$ and $3s_{1/2}$ subshells to be more than that between the $3s_{1/2}$ and $3d_{3/2}$ ones. Clearly it is not. One consideration not in a shell model with a straightforward average potential is that a nucleon in an unusually far-out s orbit could be closer to the other nucleons in lower orbits than one in a far-out p orbit; the s orbit has larger values near the center of the nucleus, {N.8}. While the shell model gets a considerable number of things right, it is certainly not a very accurate model.

Then there are the odd parity states. These are not so easy to understand: they require a nucleon to be kicked up past a major shell boundary. That should require a lot of energy according to the ideas of the shell model. It seems to make them hard to reconcile with the much higher energy of the $\frac{3}{2}^+$ state. Some thoughts on these states will be given in the next subsection.

The fourth nucleus in figure 14.16 is nitrogen-14. This is an odd-odd nucleus,

with both an odd number of protons and of neutrons. The odd proton and odd neutron are in the $2p_{1/2}$ shell, so each has spin $\frac{1}{2}$. Quantum mechanics allows the two to combine their spins into a triplet state of net spin one, like they do in deuterium, or in a singlet state of spin zero. Indeed the ground state is a 1^+ one like deuterium. The lowest excited state is a 0^+ one.

The most obvious way to further excite the nucleus with minimal energy would be to kick up a nucleon from the $2p_{3/2}$ subshell to the $2p_{1/2}$ one. That fills the $2p_{1/2}$ shell, making its net spin zero. However, there is now a "hole," a missing particle, in the $2p_{3/2}$ shell.

> *Holes in an otherwise filled subshell have the same possible angular momentum values as particles in an otherwise empty shell.*

Therefore the hole must have the spin $\frac{3}{2}$ of a single particle. This can combine with the $\frac{1}{2}$ of the odd nucleon of the opposite type to either spin 1 or spin 2. A relatively low energy 1^+ state can be observed in the experimental spectrum.

The next higher 0^- state would require a particle to cross a major shell boundary. Then again, the energy of this excited state is quite substantial at 5 MeV. It seems simpler to assume that a $1s_{1/2}$ nucleon is kicked to the $2p_{1/2}$ shell than that a $2p_{3/2}$ nucleon is kicked to the $3d_{5/2}$ one. In the latter case, it seems harder to explain why the four odd nucleons would want to particularly combine their spins to zero. And you could give an argument based on the ideas of the next subsection that 4 odd nucleons is a lot.

14.12.4 Draft: Shell model with pairing

This section examines some nuclei with more than a single nucleon in an unfilled shell.

Consider first oxygen-18 in figure 14.17, with both an even number of protons and an even number of neutrons. As always, the filled subshells have no angular momentum. That leaves the two $3d_{5/2}$ neutrons. These could have combined integer spin from 0 to 5 if they were distinguishable particles. However, the two neutrons are identical fermions, and the wave function must be antisymmetric with respect to their exchange. It can be seen from chapter 12.8 item 3, or more simply from table 12.1, that only the 0, 2, and 4 combined spins are allowed. Still, that leaves three possibilities for the net spin of the entire nucleus.

Now the basic shell model is an "independent particle model:" there are no direct interactions between the particles. Each particle moves in a given average potential, regardless of what the others are doing. Therefore, if the shell model as covered so far would be strictly true, all three spin states 0, 2, and 4 of oxygen-18 should have equal energy. Then the ground state should be any combination of these spins. But that is untrue. The ground-state has zero spin:

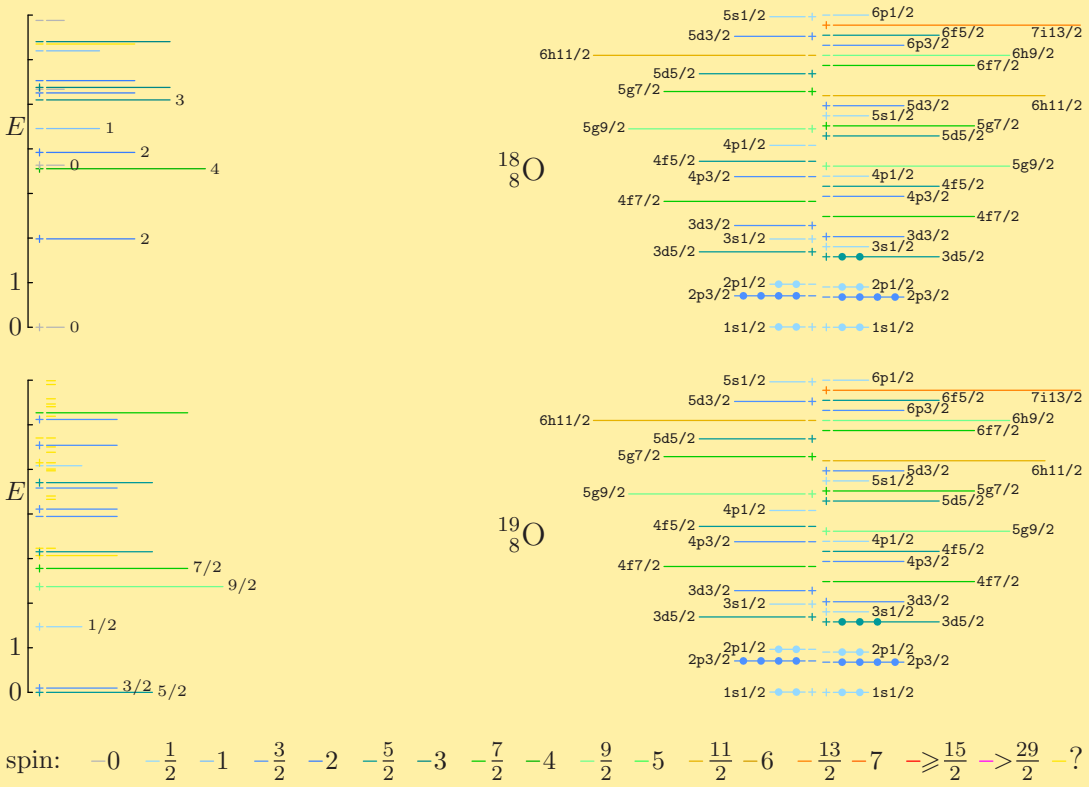


Figure 14.17: Nucleon pairing effect. [pdf]

All even-even nuclei have zero spin and even parity in the ground state.

There are zero known exceptions to this rule among either the stable or unstable nuclei.

So physicists have concluded that besides the average potential included in the shell model, there must be an additional "pairing energy" that makes nucleons of the same type want to combine pairwise into states of zero spin. In order to treat this effect mathematically without losing the basic shell model, the pairing energy must be treated as a relatively small perturbation to the shell model energy. Theories that do so are beyond the scope of this book, although the general ideas of perturbation theories can be found in addendum {A.38}. Here it must be suffice to note that the pairing effect exists and is due to interactions between nucleons not included in the basic shell model potential.

Therefore the basic shell model will from here on be referred to as the "unperturbed" shell model. The "perturbed shell model" will refer to the shell model in which additional energy corrections are assumed to exist that account for nontrivial interactions between individual nucleons. These corrections will not be explicitly discussed, but some of their effects will be demonstrated by means of experimental energy spectra.

If the pairing energy is a relatively small perturbation to the shell model, then for oxygen-18 you would expect that besides the zero spin ground state, the other possibilities of spin 2 and 4 would show up as low-lying excited states. Indeed the experimental spectrum in figure 14.17 shows 2^+ and 4^+ states of the right spin and parity, though their energy is obviously not so very low. To put it in context, the von Weizsäcker formula puts the pairing energy at $22/\sqrt{A}$ MeV, which would be of the rough order of 5 MeV for oxygen-18.

If one neutron of the pair is kicked up to the $3s_{1/2}$ state, a 2^+ or 3^+ state should result. This will require the pair to be broken up and a subshell boundary to be crossed. A potential 2^+ candidate is present in the spectrum.

Like for oxygen-16, there is again an excited 0^+ state of relatively low energy. In this case however, its energy seems rather high in view that the two $3d_{5/2}$ neutrons could simply be kicked up across the minor shell boundary to the very nearby $3s_{1/2}$ shell. An explanation can be found in the fact that physicists have concluded that:

The pairing energy increases with the angular momentum of the subshell.

When the neutron pair is kicked from the $3d_{5/2}$ shell to the $3s_{1/2}$, its pairing energy decreases. Therefore this excitation requires additional energy besides the crossing of the minor shell boundary.

It seems therefore that the perturbed shell model can give a plausible explanation for the various features of the energy spectrum. However, care must be

taken not to attach too much finality to such explanations. Section 14.13 will give a very different take on the excited states of oxygen-18. Presumably, neither explanation will be very accurate. Only additional considerations beyond mere energy levels can decide which explanation gives the better description of the excited states.

> *The purpose in this section is to examine what features seem to have a reasonable explanation within a shell model context, not how absolutely accurate that explanation really is.*

Consider again the 0^+ excited state of oxygen-16 in figure 14.16 as discussed in the previous subsection. Some of the energy needed for a pair of $2p_{1/2}$ nucleons to cross the major shell boundary to the $3d_{5/2}$ subshell will be compensated for by the higher pairing energy in the new subshell. It still seems curious that the state would end up below the 3^- one, though.

Similarly, the relatively low energy $\frac{1}{2}^-$ state in oxygen-17 and fluorine-17 can now be made a bit more plausible. To explain the negative parity, a nucleon must be kicked across the major shell boundary from the $2p_{1/2}$ subshell to the $3d_{5/2}$ one. That should require quite a bit of energy, but this will in part be compensated for by the fact that pairing now occurs at higher angular momentum.

So what to make of the next $\frac{5}{2}^-$ state? One possibility is that a $2p_{1/2}$ nucleon is kicked to the $3s_{1/2}$ subshell. The three spins could then combine into 5/2, [31, p. 131]. If true however, this would be a quite significant violation of the basic ideas of the perturbed shell model. Just consider: it requires breaking up the $2p_{1/2}$ pair and kicking one of the two neutrons across both a major shell boundary and a subshell one. That would require less energy than the $\frac{3}{2}^+$ excitation in which the odd nucleon is merely kicked over two subshell boundaries and no pair is broken up? An alternative that is more consistent with the perturbed shell model ideas would be that the $\frac{5}{2}^-$ excitation is like the $\frac{3}{2}^-$ one, but with an additional partial break up of the resulting pair. The energy seems still low.

How about nuclei with an odd number of neutrons and/or protons in a subshell that is greater than one? For these:

> *The "odd-particle shell model" predicts that even if the number of nucleons in a subshell is odd, in the ground state all nucleons except the final odd one still combine into spherically symmetric states of zero spin.*

That leaves only the final odd nucleon to provide any nonzero spin and corresponding nontrivial electromagnetic properties.

Figure 14.17 shows the example of oxygen-19, with three neutrons in the unfilled $3d_{5/2}$ subshell. The odd-particle shell model predicts that the first two

neutrons still combine into a state of zero spin like in oxygen-18. That leaves only the spin of the third neutron. And indeed, the total nuclear spin of oxygen-18 is observed to be 5/2 in the ground state, the spin of this odd neutron. The odd-particle shell model got it right.

It is important to recognize that the odd-particle shell model only applies to the ground state. This is not always sufficiently stressed. Theoretically, three $3d_{5/2}$ neutrons can combine their spins not just to spin 5/2, but also to 3/2 or 9/2 while still satisfying the antisymmetrization requirement, table 12.1. And indeed, the oxygen-19 energy spectrum in figure 14.17 shows relatively low energy $^3/_2{}^+$ and $^9/_2{}^+$ states. To explain the energies of these states would require computation using an actual perturbed shell model, rather than just the odd-particle assumption that such a model will lead to perfect pairing of even numbers of nucleons.

It is also important to recognize that the odd-particle shell model is only a prediction. It does fail for a fair number of nuclei. That is true even excluding the very heavy nuclei for which the shell model does not apply period. For example, note in figure 14.17 how close together are the $^5/_2{}^+$ and $^3/_2{}^+$ energy levels. You might guess that the order of those two states could easily be reversed for another nucleus. And so it can; there are a number of nuclei in which the spins combine into a net spin one unit less than that of the last odd nucleon. While the unperturbed shell model does not fundamentally fail for such nuclei, (because it does not predict the spin at all), the additional odd-particle assumption does.

It should be noted that different terms are used in literature for the odd-particle shell model. The term "shell model with pairing" is accurate and understandable, so that is not used. Some authors use the term "extreme independent particle model." You read that right. While the unperturbed shell model is an independent particle model, the shell model with pairing has become a *dependent* particle model: there are now postulated direct interactions between the nucleons causing them to pair. So what better way to confuse students than to call a dependent particle model an *extreme in*dependent particle model? However, this term is too blatantly wrong even for some physicists. So, some other books use instead "extreme single-particle model," and still others use "one-particle shell model." Unfortunately, it is fundamentally a multiple-particle model. You cannot have particle interactions with a single particle. Only physicists would come up with three different names for the same model and get it wrong in each single case. This book uses the term odd-particle shell model, (with odd in dictionary rather than mathematical sense), since it is not wrong and sounds much like the other names being bandied around. (The official names could be fixed up by adding the word "almost," like in "extreme almost independent particle model." This book will not go there, but you could substitute "asymptotically" for "almost" to sound more scientific.)

While the odd-particle model applies only to the ground state, *some* excited

states can still be described as purely odd-particle effects. In particular, for the oxygen-19 example, the odd $3d_{5/2}$ neutron could be kicked up to the $3s_{1/2}$ subshell with no further changes. That would leave the two remaining $3d_{5/2}$ neutrons with zero spin, and the nucleus with the new spin 1/2 of the odd neutron. Indeed a low-lying $^1/_2{}^+$ state is observed. (Because of the antisymmetrization requirement, this state cannot result from three neutrons in the $3d_{5/2}$ subshell.)

It may further be noted that "pairing" is not really the right quantum term. If two nucleons have paired into the combination of zero net spin, the next two cannot just enter the same combination without violating the antisymmetrization requirements between the pairs. What really happens is that all four as a group combine into a state of zero spin. However, everyone uses the term pairing, and so will this book.

Examples that highlight the perturbation effects of the shell model are shown in figure 14.18. These nuclei have unfilled $4d_{7/2}$ shells. Since that is a major shell with no subshells, nucleon transitions to different shells require quite a bit of energy.

First observe that all three nuclei have a final odd $4f_{7/2}$ nucleon and a corresponding ground state spin of 7/2 just like the odd-particle shell model says they should. And the net nuclear parity is negative like that of the odd nucleon. That is quite gratifying.

As far as calcium-41 is concerned, one obvious minimal-energy excitation would be that the odd neutron is kicked up from the $4f_{7/2}$ shell to the $4p_{3/2}$ shell. This will produce a $^3/_2{}^-$ excited state. Such a state does indeed exist and it has relatively high energy, as you would expect from the fact that a major shell boundary must be crossed.

Another obvious minimal-energy excitation would be that a nucleon is kicked up from the filled $3d_{3/2}$ shell to pair up with the odd nucleon already in the $4f_{7/2}$ shell. This requires again that a major shell boundary is crossed, though some energy can be recovered by the fact that the new nucleon pairing is now at higher spin. Since here a nucleon changes shells from the positive parity $3d_{3/2}$ subshell to the negative $4f_{7/2}$ one, the nuclear parity reverses and the excited state will be a $^3/_2{}^+$ one. Such a state is indeed observed.

The unstable mirror twin of calcium-41, scandium-41 has energy levels that are very much the same.

Next consider calcium-43. The odd-particle shell model correctly predicts that in the ground state, the first two $4f_{7/2}$ neutrons pair up into zero spin, leaving the $^7/_2$ spin of the third neutron as the net nuclear spin. However, even allowing for the antisymmetrization requirements, the three $4f_{7/2}$ neutrons could instead combine into spin $^3/_2$, $^5/_2$, $^9/_2$, $^{11}/_2$, or $^{15}/_2$, table 12.1. A low-energy $^5/_2{}^-$ excited state, one unit of spin less than the ground state, is indeed observed. A $^3/_2{}^-$ state is just above it. On the other hand, the lowest known $^9/_2{}^-$ state has more energy than the lowest $^{11}/_2{}^-$ one. Then again, consider the spin values that are not possible for the three neutrons if they stay in the $4f_{7/2}$ shell. The

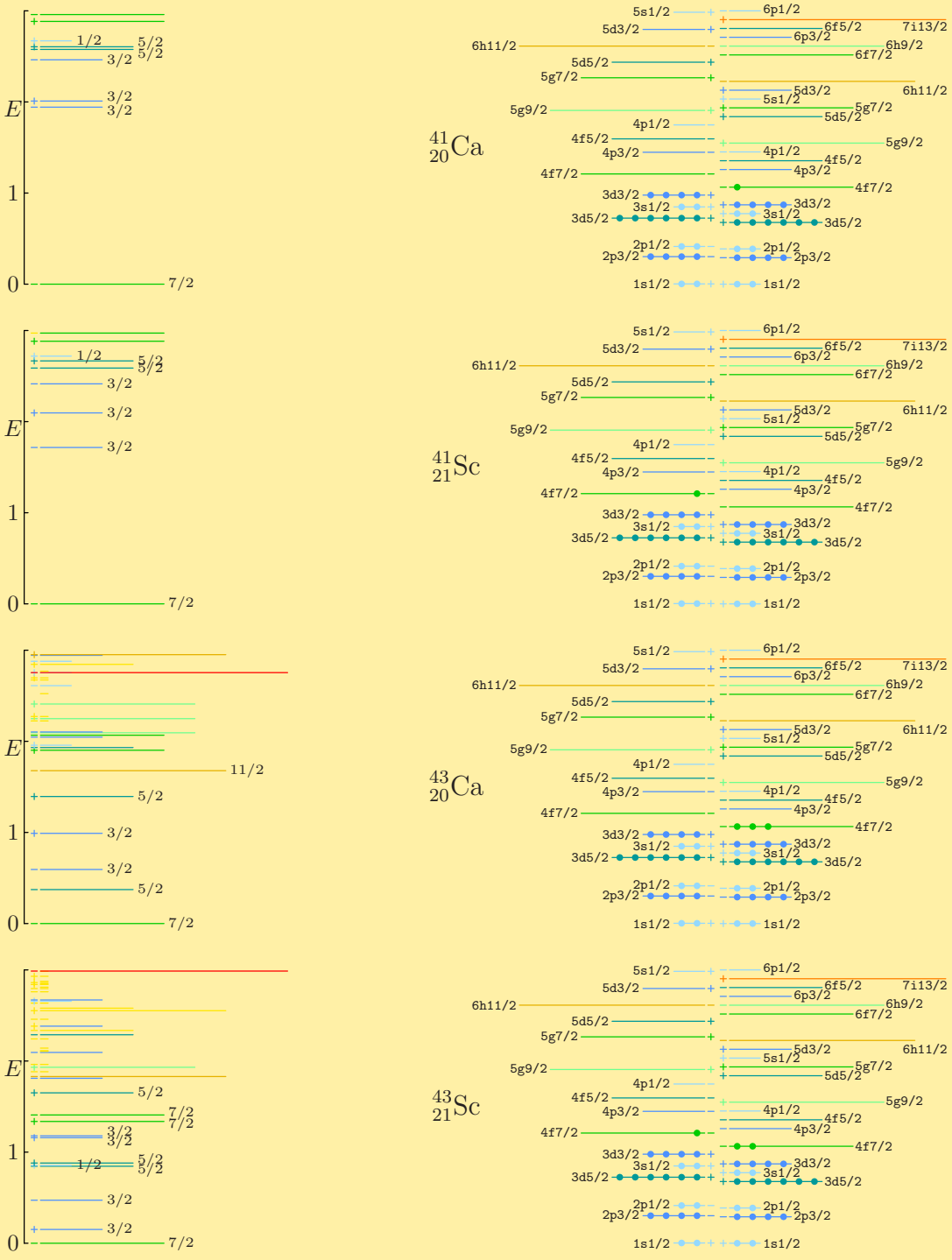


Figure 14.18: Energy levels for neighbors of doubly-magic calcium-40. [pdf]

first $^{13}/_{2}^{-}$ and $^{17}/_{2}^{-}$ states occur at energies well beyond the $^{15}/_{2}^{-}$ one, and the first $^{1}/_{2}^{-}$ state only appears at 2.6 MeV.

The lowest $^{3}/_{2}^{+}$ state energy is half that of the one for calcium-41. Apparently, the $3d_{3/2}$ neutron would rather pair up with 3 other attracting neutrons in the $4f_{7/2}$ shell than with just one. That seems reasonable enough. The overall picture seems in encouraging agreement with the perturbed shell model ideas.

Scandium-43 has one proton and two neutrons in the $4f_{7/2}$ shells. The odd-particle model predicts that in the ground state, the two neutrons combine into zero spin. However, the antisymmetrization requirement allows excited spins of 2, 4, and 6 without any nucleons changing shells. The lowest excited spin value 2 can combine with the 7/2 spin of the odd proton into excited nuclear states from $^{3}/_{2}^{-}$ up to $^{11}/_{2}^{-}$. Relatively low-lying $^{3}/_{2}^{-}$, $^{5}/_{2}^{-}$, and $^{7}/_{2}^{-}$ states, but not a $^{1}/_{2}^{-}$ one, are observed. (The lowest-lying potential $^{9}/_{2}^{-}$ state is at 1.9 MeV. The lowest lying potential $^{1}/_{2}^{-}$ state is at 3.3 MeV, though there are 4 states of unknown spin before that.)

Note how low the lowest $^{3}/_{2}^{+}$ state has sunk. That was maybe not quite unpredictable. Two protons plus two neutrons in the $4f_{7/2}$ shells have to obey less antisymmetrization requirements than four protons do, while the attractive nuclear forces between the four are about the same according to charge independence.

The difference between the energy levels of scandium-41 versus scandium-43 is dramatic. After all, the unperturbed shell model would almost completely ignore the two additional neutrons that scandium-43 has. Protons and neutrons are solved for independently in the model. It brings up a point that is often not sufficiently emphasized in other expositions of nuclear physics. The odd-particle shell model is *not* an "only the last odd particle is important" model. It is a "the last odd particle provides the ground-state spin and electromagnetic properties, because the other particles are paired up in spherically symmetric states" model. The theoretical justification for the model, which is weak enough as it is already, only applies to the second statement.

14.12.5 Draft: Configuration mixing

To better understand the shell model and its limitations, combinations of states must be considered.

Take once again the excited 0^+ state of oxygen-16 shown in figure 14.16. To create this state within the shell model picture, a pair of $2p_{1/2}$ nucleons must be kicked up to the $3d_{5/2}$ subshell. Since that requires a major shell boundary crossing by two nucleons, it should take a considerable amount of energy. Some of it will be recovered by the fact that the nucleon pairing now occurs at higher angular momentum. But there is another effect.

First of all, there are two ways to do it: either the $2p_{1/2}$ protons or the two $2p_{1/2}$ neutrons can be kicked up. One produces an excited wave function that

will be indicated by ψ_{2p} and the other by ψ_{2n}. Because of charge symmetry, and because the Coulomb force is minor for light nuclei, these two states should have very nearly the same energy.

Quantum mechanics allows for linear combinations of the two wave functions:

$$\Psi = c_1\psi_{2p} + c_2\psi_{2n}$$

Within the strict context of the unperturbed shell model, it does not make a difference. That model assumes that the nucleons do not interact directly with each other, only with an average potential. Therefore the combination should still have the same energy as the individual states.

But now consider the possibility that *both* the protons and the neutrons would be in the $3d_{5/2}$ subshell. In that case, surely you would agree that these four, mutually attracting, nucleons in the same spacial orbits would significantly interact and lower their energy. Even if the unperturbed shell model ignores that.

Of course, the four nucleons are *not* all in the $3d_{5/2}$ state; that would require four major shell crossing and make things worse. Each component state has only two nucleons in the $3d_{5/2}$ subshell. However, quantum mechanical uncertainty makes the two states interact through "twilight" terms, chapter 5.3. These act in some sense as if all four nucleons are indeed in the $3d_{5/2}$ subshell at the same time. It has the weird effect that the right combination of the states ψ_{2p} and ψ_{2n} can have significantly less energy than the lowest of the two individual states. That is particularly true if the two original states have about the same energy, as they have here.

The amount of energy lowering is hard to predict. It depends on the amount of nucleon positions that have a reasonable probability for both states and the amount of interaction of the nucleons. Intuition still suggests it should be quite considerable. And there is a more solid argument. If the strictly unperturbed shell model applies, there should be two 0^+ energy states with almost the same energy; one for protons and one for neutrons. However, if there is significant twilight interaction between the two, the energy of one of the pair will be pushed way down and the other way up. There is no known second excited 0^+ state with almost the same energy as the first one for oxygen-16.

Of course, a weird excited state at 6 MeV in a nucleus is not such a big deal. But there is more. Consider figure 14.19. It gives the excitation energy of the lowest 2^+ state for all even-even nuclei.

For all nuclei except the crossed-out ones, the 2^+ state is the lowest excited state of all. That seems curious already. Why would the lowest excited state not be a 0^+ one for a lot of even-even nuclei? Based on the shell model you would assume there are two ways to excite an even-even nucleus with minimal energy. The first way would be to kick a pair of nucleons up to the next subshell. That would create a 0^+ excited state. It could require very little energy if the subshells are close together.

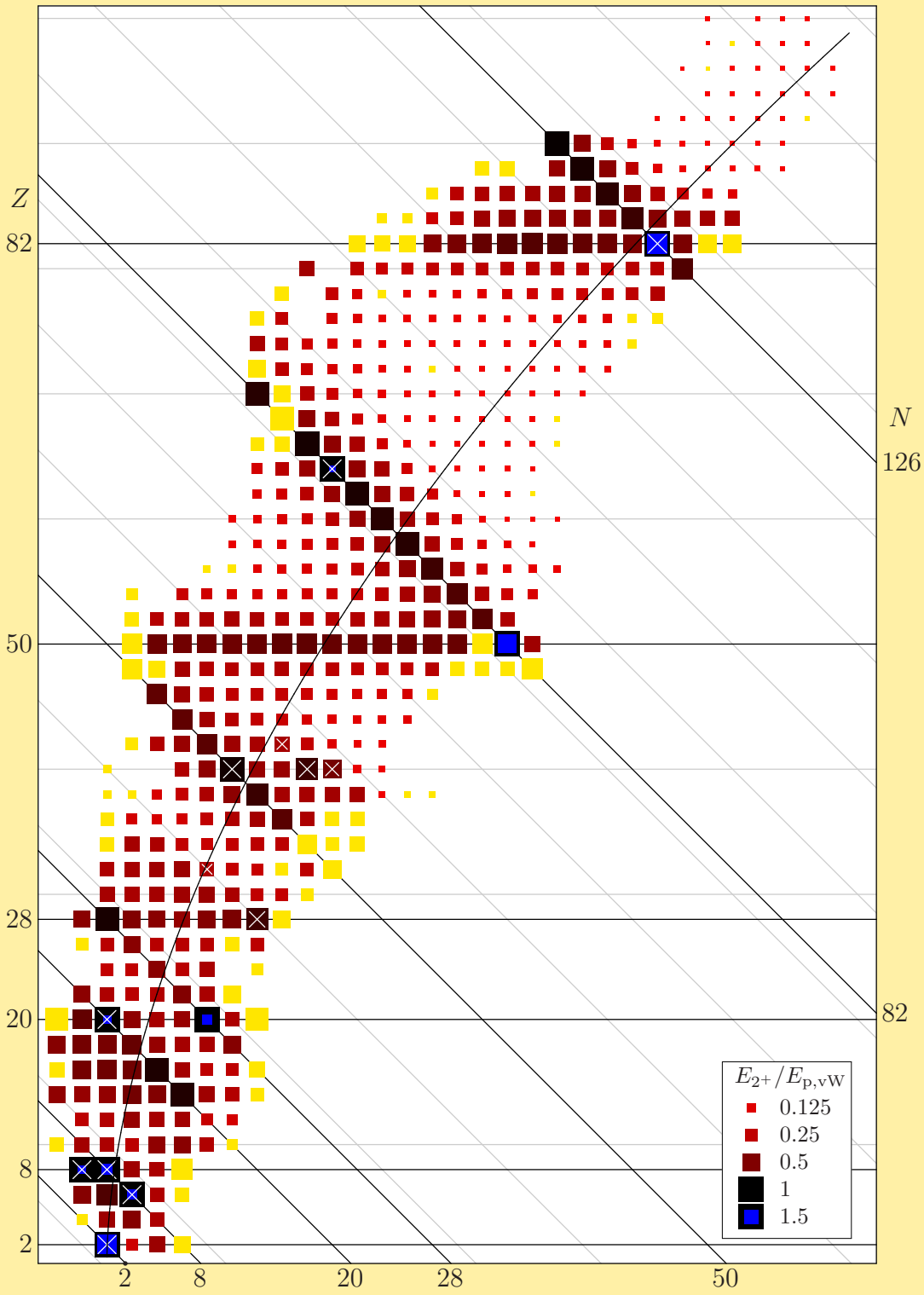


Figure 14.19: 2^+ excitation energy of even-even nuclei. [pdf][con]

The alternative way to excite an even-even nucleus with minimal energy would break up a pair, but leave them in the same subshell. This would at the minimum create a 2^+ state. (For partially filled shells of high enough angular momentum, it may also be possible to reconfigure the nucleons into a different state that still has zero angular momentum, but that does not affect the argument.) Breaking the pairing should require an appreciable amount of energy, on the MeV level. So why is the 2^+ state almost invariably the lowest energy one?

Then there is the magnitude of the 2^+ energy levels. In figure 14.19 the energies have been normalized with the von Weizsäcker value for the pairing energy,

$$\frac{2C_p}{A^{C_e}}$$

You would expect all squares to have roughly the full size, showing that it takes about the von Weizsäcker energy to break up the pair. Doubly magic nuclei are quite happy to obey. Singly magic nuclei seem a bit low, but hey, the break-up is usually only partial, you know.

But for nuclei that are not close to any magic number for either protons and neutrons all hell breaks loose. Break-up energies one to two *orders of magnitude* less than the von Weizsäcker value are common. How can the pairing energy just suddenly stop to exist?

Consider a couple of examples in figure 14.20. In case of ruthenium-104, it takes a measly 0.36 MeV to excite the 2^+ state. But there are 10 different ways to combine the four $5g_{9.2}$ protons into a 2^+ state, table 12.1. Kick up the pair of protons from the $4p_{1/2}$ shell, and there are another 10 2^+ states. The four $5g_{7/2}$ neutrons can produce another 10 of them. Kick up a pair of neutrons from the $5d_{5/2}$ subshell, and there is 10 more. Presumably, all these states will have similar energy. And there might be many other low-energy ways to create 2^+ states, [31, pp. 135-136].

Consider now the following simplistic model. Assume that the nucleus can be in any of Q different global states of the same energy,

$$\psi_1, \psi_2, \psi_3, \ldots, \psi_Q$$

Watch what happens when such states are mixed together. The energy follows from the Hamiltonian coefficients

$$\begin{array}{cccccc}
E_1 \equiv \langle\psi_1|H\psi_1\rangle & \varepsilon_{12} \equiv \langle\psi_1|H\psi_2\rangle & \varepsilon_{13} \equiv \langle\psi_1|H\psi_3\rangle & \ldots & \varepsilon_{1Q} \equiv \langle\psi_1|H\psi_Q\rangle \\
\varepsilon^*_{12} \equiv \langle\psi_2|H\psi_1\rangle & E_2 \equiv \langle\psi_2|H\psi_1\rangle & \varepsilon_{23} \equiv \langle\psi_2|H\psi_3\rangle & \ldots & \varepsilon_{2Q} \equiv \langle\psi_2|H\psi_Q\rangle \\
\varepsilon^*_{13} \equiv \langle\psi_3|H\psi_1\rangle & \varepsilon^*_{23} \equiv \langle\psi_3|H\psi_2\rangle & E_3 \equiv \langle\psi_3|H\psi_3\rangle & \ldots & \varepsilon_{3Q} \equiv \langle\psi_3|H\psi_Q\rangle \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\varepsilon^*_{1Q} \equiv \langle\psi_Q|H\psi_1\rangle & \varepsilon^*_{2Q} \equiv \langle\psi_Q|H\psi_2\rangle & \varepsilon^*_{3Q} \equiv \langle\psi_Q|H\psi_3\rangle & \ldots & E_Q \equiv \langle\psi_Q|H\psi_Q\rangle
\end{array}$$

By assumption, the energy levels $E_1, E_2, \ldots$ of the states are all about the same, and if the unperturbed shell model was exact, the perturbations $\varepsilon_{..}$ would all be

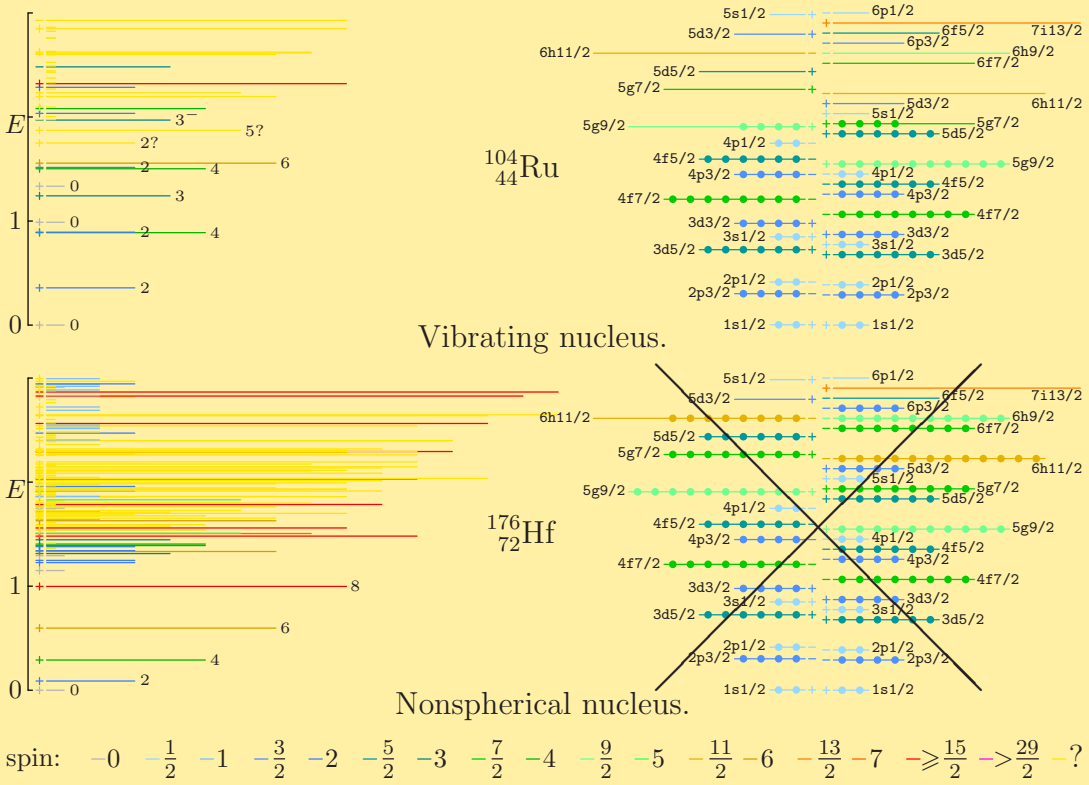


Figure 14.20: Collective motion effects. [pdf]

zero. But since the shell model is only a rough approximation of what is going on inside nuclei, the shell model states will not be true energy eigenfunctions. Therefore the coefficients $\varepsilon_{..}$ will surely not be zero, though what they will be is hard to say.

To get an idea of what can happen, assume for now that the $\varepsilon_{..}$ are all equal and negative. In that case, following similar ideas as in chapter 5.3, a state of lowered energy exists that is an equal combination of each of the Q individual excited states; its energy will be lower than the original states by an amount $(Q-1)\varepsilon$. Even if ε is relatively small, that will be a significant amount if the number Q of states with the same energy is large.

Of course, the coefficients $\varepsilon_{..}$ will not all be equal and negative. Presumably they will vary in both sign and magnitude. Interactions between states will also be limited by symmetries. (If states combine into an equivalent state that is merely rotated in space, there is no energy lowering.) Still, the lowest excitation energy will be defined by the largest negative accumulation of shell model errors that is possible.

The picture that emerges then is that the 2^+ excitation for ruthenium-104, and most other nuclei in the rough range $50 < A < 150$, is not just a matter of just one or two nucleons changing. It apparently involves the collaborative motion of a large number of nucleons. This would be quite a challenge to describe in the context of the shell model. Therefore physicists have developed different models, ones that allow for collective motion of the entire nucleus, like in section 14.13.

When the energy of the excitation hits zero, the bottom quite literally drops out of the shell model. In fact, even if the energy merely becomes low, the shell model must crash. If energy states are almost degenerate, the slightest thing will throw the nucleus from one to the other. In particular, small perturbation theory shows that originally small effects blow up as the reciprocal of the energy difference, addendum {A.38}. Physicists have found that nuclei in the rough ranges $150 < A < 190$ and $A > 220$ acquire an intrinsic nonspherical shape, fundamentally invalidating the shell model as covered here. More physically, as figure 14.19 suggests, it happens for pretty much all heavy nuclei except ones close to the magic lines. The energy spectrum of a typical nucleus in the nonspherical range, hafnium-176, is shown in figure 14.20.

14.12.6 Draft: Shell model failures

The previous subsection already indicated two cases in which the shell model has major problems with the excited states. But in a number of cases the shell model may also predict an incorrect ground state. Figure 14.21 shows some typical examples.

In case of titanium-47, the shell model predicts that there will be five neutrons in an unfilled $4f_{7/2}$ subshell. It is believed that this is indeed correct [36,

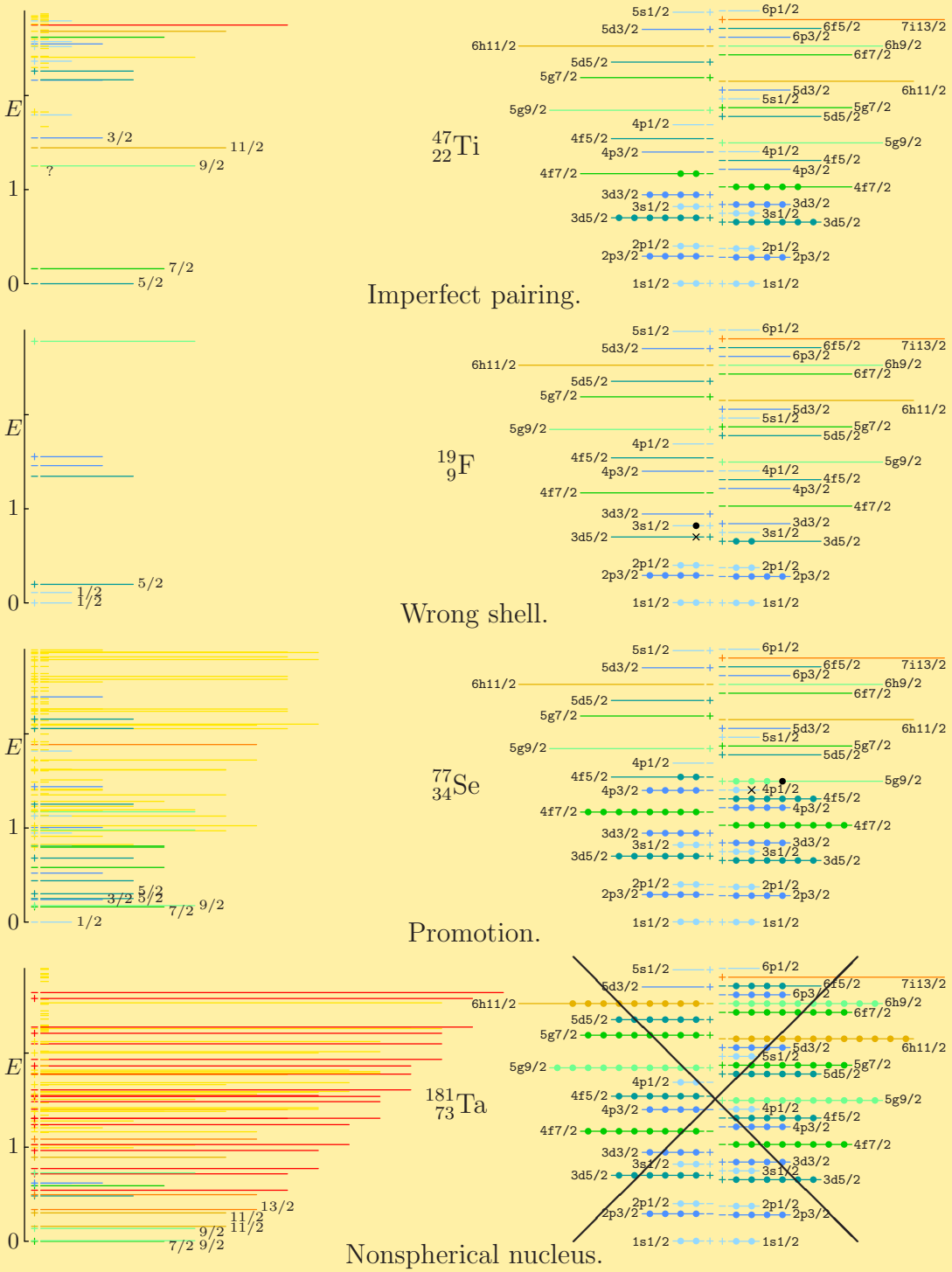


Figure 14.21: Failures of the shell model. [pdf]

p. 224]. The unperturbed shell model makes no predictions about the nuclear spin. However, the odd-particle shell model says that in the ground state the nuclear spin should be that of the odd neutron, $\frac{7}{2}$. But it is not, the spin is $\frac{5}{2}$. The pairing of the even number of neutrons in the $4f_{7/2}$ shell is not complete. While unfortunate, this is really not that surprising. The perturbation Hamiltonian used to derive the prediction of nucleon pairing is a very crude one. It is quite common to see subshells with at least three particles and three holes (three places for additional particles) end up with a unit less spin than the odd-particle model predicts. It almost happened for oxygen-19 in figure 14.17.

In fact, 5 particles in a shell in which the single-particle spin is $\frac{7}{2}$ can combine their spin into a variety of net values. Table 12.1 shows that $\frac{3}{2}$, $\frac{5}{2}$, $\frac{7}{2}$, $\frac{9}{2}$, $\frac{11}{2}$, and $\frac{15}{2}$ are all possible. Compared to that, the odd-particle prediction does not seem that bad. Note that the predicted state of spin $\frac{7}{2}$ has only slightly more energy than the ground state. On the other hand, other states that might be produced through the combined spin of the five neutrons have much more energy.

Fluorine-19 shows a more fundamental failure of the shell model. The shell model would predict that the odd proton is in the $3d_{5/2}$ state, giving the nucleus spin $\frac{5}{2}$ and even parity. In fact, it should be just like fluorine-17 in figure 14.16. For the unperturbed shell model, the additional two neutrons should not make a significant difference. But the nuclear spin is $\frac{1}{2}$, and that means that the odd proton must be in the $3s_{1/2}$ state. A look at figure 14.15 shows that the unperturbed shell model cannot qualitatively explain this swapping of the two states.

It is the theoretician's loss, but the experimentalist's gain. The fact that fluorine has spin one-half makes it a popular target for nuclear magnetic resonance studies. Spin one-half nuclei are easy to analyze and they do not have nontrivial electric fields that mess up the nice sharp signals in nuclei with larger spin.

And maybe the theoretician can take some comfort in the fact that this complete failure is rare among the light nuclei. In fact, the main other example is fluorine-19's mirror twin neon-19. Also, there is an excited state with the correct spin and parity just above the ground state. But no funny business here; if you are going to call fluorine-19 almost right, you have to call fluorine-17 almost wrong.

Note also how low the $\frac{1}{2}^-$ excited state has become. Maybe this can be somewhat understood from the fact that the kicked-up $2p_{1/2}$ proton is now in a similar spatial orbit with three other nucleons, rather than just one like in the case of fluorine-17. In any case, it would surely require a rather sophisticated perturbed shell model to describe it, one that includes nucleons of both type in the perturbation.

And note that formulating a perturbed shell model from physical principles is not easy anyway, because the basic shell model already includes the interactions

between nucleons in an average sense. The perturbations must not just identify the interactions, but more importantly, what part of these interactions is still missing from the unperturbed shell model.

For the highly unstable beryllium-11 and nitrogen-11 mirror nuclei, the shell model gets the spin right, but the parity wrong! In shell model terms, a change of parity requires the crossing of a major shell boundary. Beryllium-11 is known to be a "halo nucleus," a nucleus whose radius is noticeably larger than that predicted by the liquid drop formula (14.9). This is associated with a gross inequality between the number of protons and neutrons. Beryllium-11 has only 4 protons, but 7 neutrons; far too many for such a light nucleus. Beryllium-13 with 9 neutrons presumably starts to simply throw the bums out. Beryllium-11 does not do that, but it keeps one neutron at arms length. The halo of beryllium-11 is a single neutron one. (That of its beta-decay parent lithium-11 is a two-neutron one. Such a nucleus is called "Borromean," after the three interlocking rings in the shield of the princes of Borromeo. Like the rings, the three-body system lithium-9 plus two neutrons hangs together but if any of the three is removed, the other two fall apart too. Both lithium-10 and the dineutron are not bound.) Halo nucleons tend to prefer states of low orbital angular momentum, because in classical terms it reduces the kinetic energy they need for angular motion. The potential energy is less significant so far out. In shell model terms, the beryllium-11 neutron has the $3s_{1/2}$ state available to go to; that state does indeed have the $\frac{1}{2}$ spin and positive parity observed. Very little seems to be known about nitrogen-11 at the time of writing; no energy levels, no electric quadrupole moment (but neither is there for beryllium-11). It is hard to do experiments at your leisure on a nucleus that lives for less than 10^{-21} s.

For much heavier nuclei, the subshells are often very close together. Also, unlike for the $3d_{5/2}$ and $3s_{1/2}$ states, the shell model often does not produce an unambiguous ordering for them. In that case, it is up to you whether you want to call it a failure if a particle does not follow whatever ambiguous ordering you have adopted.

Selenium-77 illustrates a more fundamental reason why the odd particle may end up in the wrong state. The final odd neutron would normally be the third one in the $5g_{9/2}$ state. That would give the nucleus a net spin of $\frac{9}{2}$ and positive parity. There is indeed a low-lying excited state like that. (It is just above a $\frac{7}{2}$ one that might be an effect of incomplete pairing.) However, the nucleus finds that if it promotes a neutron from the $4p_{1/2}$ shell to the $5g_{9/2}$ one just above, that neutron can pair up at higher angular momentum, lowering the overall nuclear energy. That leaves the odd neutron in the $4p_{1/2}$ state, giving the nucleus a net spin of $\frac{1}{2}$ and negative parity. Promotion happens quite often if there are more than 32 nucleons of a given type and there is a state of lower spin immediately below the one being filled.

Tantalum-181 is an example nucleus that is not spherical. For it, the shell

model simply does not apply as derived here. So there is no need to worry about it. Which is a good thing, because it does not seem easy to justify a $^7/_2{}^+$ ground state based on the shell model. As noted in the previous subsection, nonspherical nuclei appear near the stable line for mass numbers of about 150 to 190 and above 220. There are also a few with mass numbers between 20 and 30.

Preston & Bhaduri [36, p. 224ff] give an extensive table of nucleons with odd mass number, listing shell occupation numbers and spin. Notable is iron-57, believed to have three neutrons in the $4p_{3/2}$ shell as the shell model says, but with a net nuclear spin of $^1/_2{}^-$. Since the three neutrons cannot produce that spin, in a shell model explanation the 6 protons in the $4f_{7/2}$ shell will need to contribute. Similarly neodymium-149 with, maybe, 7 neutrons in the $6f_{7/2}$ shell has an unexpected $^5/_2{}^-$ ground state. Palladium-101 with 5 neutrons in the $5d_{5/2}$ shell has an unexpected spin $^7/_2$ according to the table; however, the more recent data of [3] list the nucleus at the expected $^5/_2{}^+$ value. In general the table shows that the ground state spin values of spherical nuclei with odd mass numbers are almost all correctly predicted if you know the correct occupation numbers of the shells. However, predicting those numbers for heavy nuclei is often nontrivial.

14.13 Draft: Collective Structure

Some nuclear properties are difficult to explain using the shell model approach as covered here. Therefore physicists have developed different models.

For example, nuclei may have excited states with unexpectedly low energy. One example is ruthenium-104 in figure 14.20, and many other even-even nuclei with such energies may be found in figure 14.19. If you try to explain the excitation energy within a shell model context, you are led to the idea that many shell model excitations combine forces, as in section 14.12.5.

Then there are nuclei for which the normal shell model does not work at all. They are called the nonspherical or deformed nuclei. Among the line of most stable nuclei, they are roughly the "rare earth" lanthanides and the extremely heavy actinides that are deformed. In terms of the mass number, the ranges are about $150 < A < 190$ and $220 < A$. (However, various very unstable lighter nuclei are quite nonspherical too. None of this is written in stone.) In terms of figure 14.19, they are the very small squares. Examples are hafnium-176 in figure 14.20 and tantalum-181 in figure 14.21.

It seems clear that many or all nuclei participate in these effects. Trying to explain such organized massive nucleon participation based on a perturbed basic shell model alone would be very difficult, and mathematically unsound in the case of deformed nuclei. A completely different approach is desirable.

Nuclei with many nucleons and densely spaced energy levels bear some similarity to macroscopic systems. Based on that idea, physicists had another look at the classical liquid drop model for nuclei. That model was quite successful in explaining the size and ground state energy levels of nuclei in section 14.10.

But liquid drops are not necessarily static; they can vibrate. Vibrating states provide a model for low-energy excited states in which the nucleons as a group participate nontrivially. Furthermore, the vibrations can become unstable, providing a model for permanent nuclear deformation or nuclear fission. Deformed nuclei can display effects of rotation of the nuclei. This section will give a basic description of these effects.

14.13.1 Draft: Classical liquid drop

This section reviews the mechanics of a classical liquid drop, like say a droplet of water. However, there will be one additional effect included that you would be unlikely to see in a drop of water: it will be assumed that the liquid contains distributed positively charged ions. This is needed to allow for the very important destabilizing effect of the Coulomb forces in a nucleus.

It will be assumed that the nuclear "liquid" is homogeneous throughout. That is a somewhat doubtful assumption for a model of a nucleus; there is no a priori reason to assume that the proton and neutron motions are the same. But a two-liquid model, such as found in [40, p. 183ff], is beyond the current coverage.

It will further be assumed that the nuclear liquid preserves its volume. This assumption is consistent with the formula (14.9) for the nuclear radius, and it greatly simplifies the classical analysis.

The von Weizsäcker formula showed that the nuclear potential energy increases with the surface area. The reason is that nucleons near the surface of the nucleus are not surrounded by a full set of attracting neighboring nucleons. Macroscopically, this effect is explained as "surface tension." Surface tension is defined as increased potential energy per unit surface area. (The work in expanding the length of a rectangular surface area must equal the increased potential energy of the surface molecules. From that it is seen that the surface tension is also the tension force at the perimeter of the surface per unit length.)

Using the surface term in the von Weizsäcker formula (14.10) and (14.9), the nuclear equivalent of the surface tension is

$$\sigma = \frac{C_s}{4\pi R_A^2} \tag{14.17}$$

The C_d term in the von Weizsäcker formula might also be affected by the nuclear surface area because of its unavoidable effect on the nuclear shape, but to simplify things this will be ignored.

The surface tension wants to make the surface of the drop as small as possible. It can do so by making the drop spherical. However, this also crowds the protons together the closest, and the Coulomb repulsions resist that. So the Coulomb term fights the trend towards a spherical shape. This can cause heavy nuclei, for which the Coulomb term is big, to fission into pieces. It also makes lighter nuclei less resistant to deformation, promoting nuclear vibrations or even permanent deformations. To include the Coulomb term in the analysis of a classical drop of liquid, it can be assumed that the liquid is charged, with total charge Ze.

Infinitesimal vibrations of such a liquid drop can be analyzed, {A.43}. It is then seen that the drop can vibrate around the spherical shape with different natural frequencies. For a single mode of vibration, the radial displacement of the surface of the drop away from the spherical value takes the form

$$\delta = \varepsilon l \sin(\omega t - \varphi) \bar{Y}_l^m(\theta, \phi) \tag{14.18}$$

Here εl is the infinitesimal amplitude of vibration, ω the frequency, and φ a phase angle. Also θ and ϕ are the coordinate angles of a spherical coordinate system with its origin at the center of the drop, N.3. The $\bar{Y}_l^m$ are essentially the spherical harmonics of orbital angular momentum fame, chapter 4.2.3. However, in the classical analysis it is more convenient to use the real version of the Y_l^m. For $m = 0$, there is no change, and for $m \neq 0$ they can be obtained from the complex version by taking $Y_l^m \pm Y_l^{-m}$ and dividing by $\sqrt{2}$ or $\sqrt{2}$i as appropriate.

Vibration with $l = 0$ is not possible, because it would mean that the radius increased or decreased everywhere, (Y_0^0 is a constant), which would change the volume of the liquid. Motion with $l = 1$ is possible, but it can be seen from the spherical harmonics that this corresponds to translation of the drop at constant velocity, not to vibration.

Vibration occurs only for $l \geqslant 2$, and the frequency of vibration is then, {A.43}:

$$\omega = \sqrt{\frac{E_{s,l}^2}{\hbar^2}\frac{1}{A} - \frac{E_{c,l}^2}{\hbar^2}\frac{Z^2}{A^2}} \tag{14.19}$$

The constants $E_{s,l}$ and $E_{c,l}$ express the relative strengths of the surface tension and Coulomb repulsions, respectively. The values of these constants are, expressed in energy units,

$$E_{s,l} = \frac{\hbar c}{R_A}\sqrt{\frac{(l-1)l(l+2)}{3}\frac{C_s}{m_{\rm p}c^2}} \qquad E_{c,l} = \frac{\hbar c}{R_A}\sqrt{\frac{2(l-1)l}{2l+1}\frac{e^2}{4\pi\epsilon_0 R_A m_{\rm p}c^2}} \tag{14.20}$$

The most important mode of vibration is the one at the lowest frequency, which means $l = 2$. In that case the numerical values of the constants are

$$E_{s,2} \approx 35 \text{ MeV} \qquad E_{c,2} \approx 5.1 \text{ MeV} \tag{14.21}$$

Of course a nucleus with a limited number of nucleons and energy levels is not a classical system with countless molecules and energy levels. The best you may hope for that there will be some reasonable qualitative agreement between the two.

It turns out that the liquid drop model significantly overestimates the stability of nuclei with respect to relatively small deviations from spherical. However, it does much a better job of estimating the stability against the large scale deformations associated with nuclear fission.

Also, the inertia of a nucleus can be quite different from that of a liquid drop, [36, p. 345, 576]. This however affects $E_{s,l}$ and $E_{c,l}$ equally, and so it does not fundamentally change the balance between surface tension and Coulomb repulsions.

14.13.2 Draft: Nuclear vibrations

In the previous subsection, the vibrational frequencies of nuclei were derived using a classical liquid drop model. They apply to vibrations of infinitely small amplitude, hence infinitesimal energy.

However, for a quantum system like a nucleus, energy should be quantized. In particular, just like the vibrations of the electromagnetic field come in photons of energy $\hbar\omega$, you expect vibrations of matter to come in "phonons" of energy $\hbar\omega$. Plugging in the classical expression for the lowest frequency gives

$$E_{\text{vibration}} = \sqrt{E_{s,2}^2 \frac{1}{A} - E_{c,2}^2 \frac{Z^2}{A^2}} \qquad E_{s,2} \approx 35 \text{ MeV} \quad E_{c,2} \approx 5.1 \text{ MeV} \quad (14.22)$$

That is in the ballpark of excitation energies for nuclei, suggesting that collective motion of the nucleons is something that must be considered.

In particular, for light nuclei, the predicted energy is about $35/\sqrt{A}$ MeV, comparable to the von Weizsäcker approximation for the pairing energy, $22/\sqrt{A}$ MeV. Therefore, it is in the ballpark to explain the energy of the 2^+ excitation of light even-even nuclei in figure 14.19. The predicted energies are however definitely too high. That reflects the fact mentioned in the previous subsection that the classical liquid drop overestimates the stability of nuclei with respect to small deformations. Note also the big discrete effects of the magic numbers in the figure. Such quantum effects are completely missed in the classical liquid drop model.

It should also be pointed out that now that the energy is quantized, the basic assumption that the amplitude of the vibrations is infinitesimal is violated. A quick ballpark shows the peak quantized surface deflections to be in the fm range, which is not really small compared to the nuclear radius. If the amplitude was indeed infinitesimal, the nucleus would electrically appear as a spherically symmetric charge distribution. Whether you want to call the deviations from

that prediction appreciable, [36, p. 342, 354] or small, [31, p. 152], nonzero values should certainly be expected.

As far as the spin is concerned, the classical perturbation of the surface of the drop is given in terms of the spherical harmonics Y_2^m. The overall mass distribution has the same dependence on angular position. In quantum terms you would associate such an angular dependence with an azimuthal quantum number 2 and even parity, hence with a 2^+ state. It all seems to fit together rather nicely.

There is more. You would expect the possibility of a two-phonon excitation at twice the energy. Phonons, like photons, are bosons; if you combine two of them in a set of single-particle states of square angular momentum $l = 2$, then the net square angular momentum can be either 0, 2, or 4, table 12.2. So you would expect a triplet of excited 0^+, 2^+, and 4^+ states at roughly twice the energy of the lowest 2^+ excited state.

And indeed, oxygen-18 in figure 14.17 shows a nicely compact triplet of this kind at about twice the energy of the lowest 2^+ state. Earlier, these states were seemingly satisfactorily explained using single-nucleon excitations, or combinations of a few of them. Now however, the liquid drop theory explains them, also seemingly satisfactory, as motion of the entire nucleus!

That does not necessarily mean that one theory must be wrong and one right. There is no doubt that neither theory has any real accuracy for this nucleus. The complex actual dynamics is quite likely to include nontrivial aspects of each theory. The question is whether the theories can reasonably predict correct properties of the nucleus, regardless of the approximate way that they arrive at those predictions. Moreover, a nuclear physicist would always want to look at the decay of the excited states, as well as their electromagnetic properties where available, before classifying their nature. That however is beyond the scope of this book.

Many, but by no means all, even-even nuclei show similar vibrational characteristics. That is illustrated in figure 14.22. This figure shows the ratio of the second excited energy level E_2, regardless of spin, divided by the energy E_{2^+} of the lowest 2^+ excited state. Normally E_{2^+} is the lowest excited state of all, so E_2 will be larger, making the ratio greater than one. That are the nuclei we are interested in here. Helium 4, the little black square at the $Z = N = 2$ intersection, is of no interest.

Now the theoretically-ideal vibrating nucleus would have a 2^+ lowest excited state, corresponding to a single phonon of vibration. It would also have a triplet of states with two phonons, where ideally speaking all three of these states should each have an energy E_2 that is exactly twice the energy E_{2^+} of the state with one phonon. That makes the ideal energy ratio E_2/E_{2^+} equal to 2. The nuclei that do have that energy ratio are shown as bright green squares in figure 14.22. These bright green squares are likely candidates for nuclei with vibrating excited states of low energy. But in addition, for the ideal vibrating nucleus, the

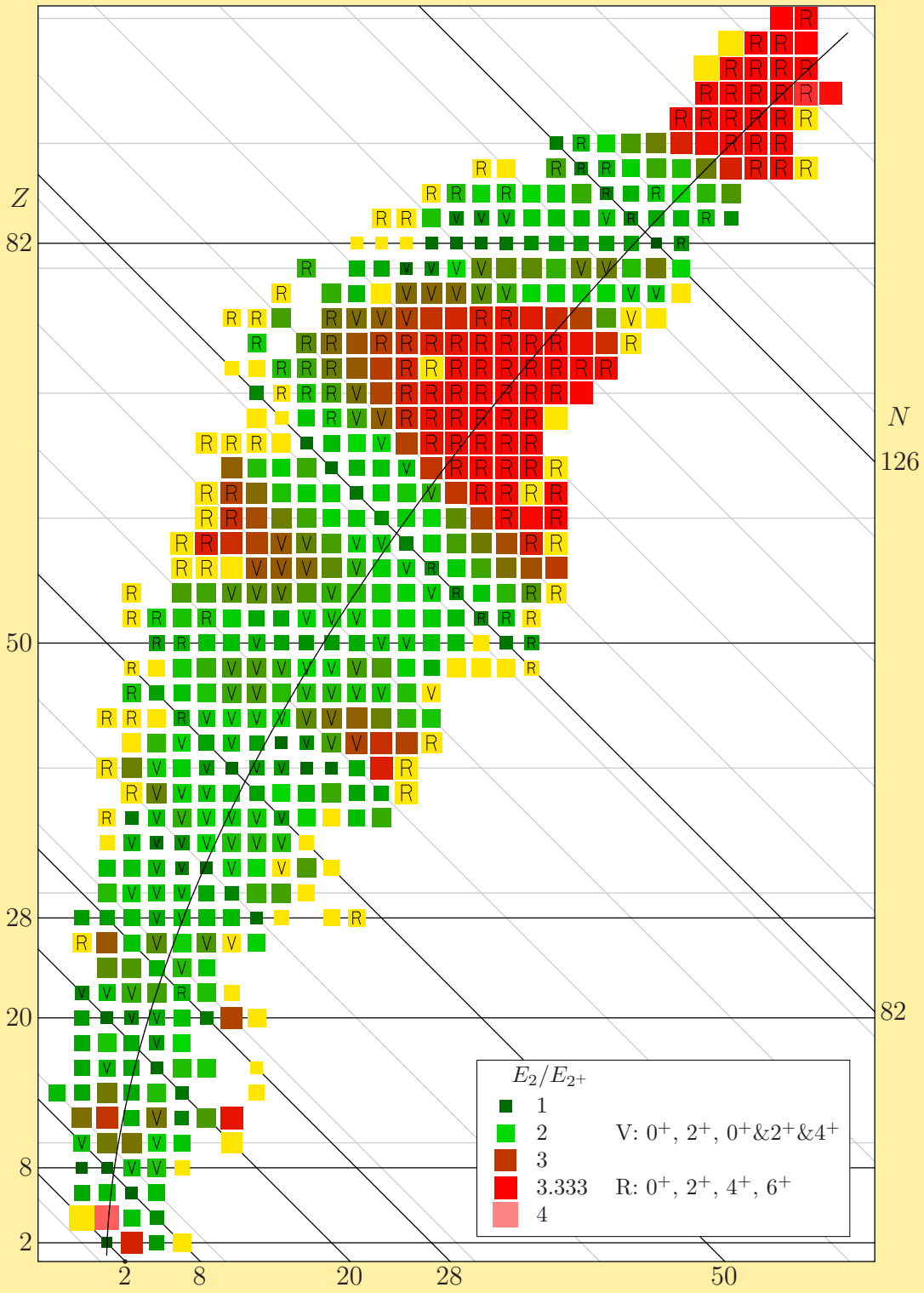


Figure 14.22: An excitation energy ratio for even-even nuclei. [pdf][con]

triplet of exited two-phonon states must consist of exactly one 0^+ state, exactly one 2^+ state, and exactly one 4^+ state, because that are the only three states that the two phonons can produce. The nuclei four which the lowest excited state is 2^+ one, and the next three states consist of one 0^+ state, one 2^+ state, and one 4^+ state, are marked with an "V" in figure 14.22. So bright green squares in figure 14.22 with an "V" in them are surely nuclei willing to vibrate. Anything else would be too much of a coincidence to believe. You can see that nuclei with vibrating excited states are quite common below $N = 82$ neutrons, or near magic numbers.

Still, many even-even nuclei do not seem to have vibrational excited states. But many of those who do not still have that unexpected lowest exited state that has spin 2^+ and very little energy. Obviously, then, liquid-drop vibrations must be only a part of the story.

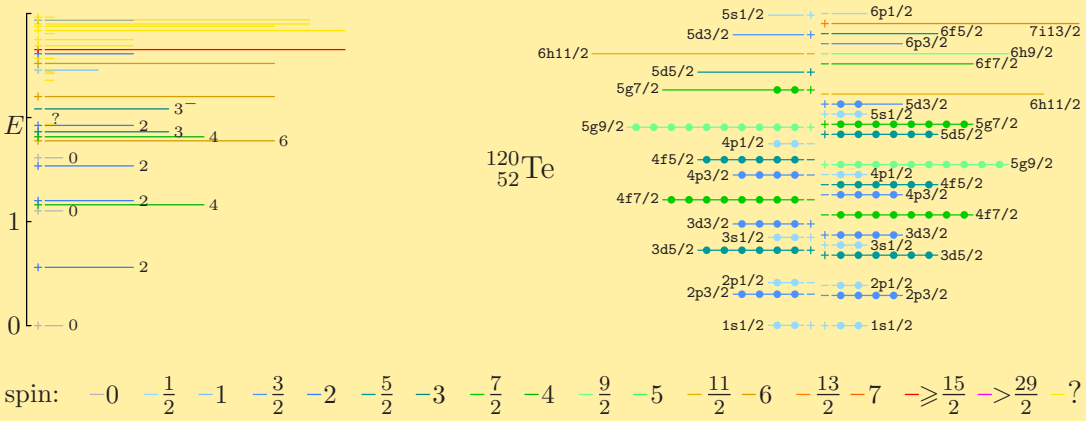


Figure 14.23: Textbook vibrating nucleus tellurium-120. [pdf]

Much heavier vibrating nuclei than oxygen-18 are tellurium-120 in figure 14.23 as well as the earlier example of ruthenium-104 in figure 14.20. Both nuclei have again a fairly compact 0^+, 2^+, 4^+ triplet at roughly twice the energy of the lowest 2^+ excited state. But these more "macroscopic" nuclei also show a nice 0^+, 2^+, 3^+, 4^+, 6^+ quintuplet at very roughly three times the energy of the lowest 2^+ state. Yes, three identical phonons of spin 2 can have a combined spin of 0, 2, 3, 4, and 6, but not 1 or 5 (c.f. 12.2).

As subsection 14.13.1 showed, liquid drops can also vibrate according to spherical harmonics Y_l^m for $l > 2$. The lowest such possibility $l = 3$ has spin 3 and negative parity. Vibration of this type is called "octupole vibration," while $l = 2$ is referred to as "quadrupole vibration." For very light nuclei, the energy of octupole vibration is about twice that of the quadrupole type. That would put the 3^- octupole vibration right in the middle of the two-phonon quadrupole triplet. However, for heavier nuclei the 3^- state will be relatively higher, since the energy reduction due to the Coulomb term is relatively smaller in the case $l = 3$. Indeed, the first 3^- states for tellurium-120 and ruthenium-104 are found

well above the quadrupole quintuplet. The lowest 3^- state for much lighter oxygen-18 is relatively lower.

14.13.3 Draft: Nonspherical nuclei

The classical liquid drop model predicts that the nucleus cannot maintain a spherical ground state if the destabilizing Coulomb energy exceeds the stabilizing nuclear surface tension. Indeed, from electromagnetic measurements, it is seen that many very heavy nuclei do acquire a permanent nonspherical shape. These are called "deformed nuclei".

They are roughly the red squares and yellow squares marked with "R" in figure 14.22. Near the stable line, their mass number ranges are from about 150 to 190 and above 220. But many unstable much lighter nuclei are deformed too.

The liquid drop model, in particular (14.19), predicts that the nuclear shape becomes unstable at

$$\frac{Z^2}{A} = \frac{E_{s,2}^2}{E_{c,2}^2} \approx 48$$

If that was true, essentially all nuclei would be spherical. A mass number of 150 corresponds to about Z^2/A equal to 26. However, as pointed out in subsection 14.13.1, the liquid drop model overestimates the stability with respect to relatively small deformations. However, it does a fairly good job of explaining the stability with respect to large ones. That explains why the deformation of the deformed nuclei does not progress until they have fissioned into pieces.

Physicists have found that most deformed nuclei can be modeled well as spheroids, i.e. ellipsoids of revolution. The nucleus is no longer assumed to be spherically symmetric, but still axially symmetric. Compared to spherical nuclei, there is now an additional nondimensional number that will affect the various properties: the ratio of the lengths of the two principal axes of the spheroid. That complicates analysis. A single theoretical number now becomes an entire set of numbers, depending on the value of the nondimensional parameter. For some nuclei furthermore, axial symmetry is insufficient and a model of an ellipsoid with three unequal axes is needed. In that case there are two nondimensional parameters. Things get much messier still then.

14.13.4 Draft: Rotational bands

Vibration is not the only semi-classical collective motion that nuclei can perform. Deformed nuclei can also rotate as a whole. This section gives a simplified semi-classical description of it.

14.13.4.1 Draft: Basic notions in nuclear rotation

Classically speaking, the kinetic energy of a solid body due to rotation around an axis is $T_{\rm R} = \frac{1}{2}\mathcal{I}_{\rm R}\omega^2$, where $\mathcal{I}_{\rm R}$ is the moment of inertia around the axis and ω the angular velocity. Quantum mechanics does not use angular velocity but angular momentum $J = \mathcal{I}_{\rm R}\omega$, and in these terms the kinetic energy is $T_{\rm R} = J^2/2\mathcal{I}_{\rm R}$. Also, the square angular momentum J^2 of a nucleus is quantized to be $\hbar^2 j(j+1)$ where j is the net "spin" of the nucleus, i.e. the azimuthal quantum number of its net angular momentum.

Therefore, the kinetic energy of a nucleus due to its overall rotation becomes:

$$\boxed{T_{\rm R} = \frac{\hbar^2}{2\mathcal{I}_{\rm R}} j(j+1) - \frac{\hbar^2}{2\mathcal{I}_{\rm R}} j_{\rm min}(j_{\rm min}+1) \qquad (j_{\rm min} \neq 1/2)} \tag{14.23}$$

Here $j_{\rm min}$ is the azimuthal quantum number of the "intrinsic state" in which the nucleus is not rotating as a whole. The angular momentum of this state is in the individual nucleons and not available for nuclear rotation, so it must be subtracted. The total energy of a state with spin j is then

$$E_j = E_{\rm min} + T_{\rm R}$$

where $E_{\rm min}$ is the energy of the intrinsic state.

Consider now first a rough ballpark of the energies involved. Since j is integer or half integer, the rotational energy comes in discrete amounts of $\hbar^2/2\mathcal{I}_{\rm R}$. The classical value for the moment of inertia $\mathcal{I}_{\rm R}$ of a rigid sphere of mass m and radius R is $\frac{2}{5}mR^2$. For a nucleus the mass m is about A times the proton mass and the nuclear radius is given by (14.9). Plugging in the numbers, the ballpark for rotational energies becomes

$$\frac{35}{A^{5/3}}\left[j(j+1) - j_{\rm min}(j_{\rm min}+1)\right] \text{ MeV}$$

For a typical nonspherical nucleus like hafnium-177 in figure 14.24, taking the intrinsic state to be the ground state with $j_{\rm min}$ equal to $7/2$, the state $9/2$ with an additional unit of spin due to nuclear rotation would have a kinetic energy of about 0.06 MeV. The lowest excited state is indeed a $9/2$ one, but its energy above the ground state is about twice 0.06 MeV. A nucleus is not at all like a rigid body in classical mechanics. It has already been pointed out in the subsection on nuclear vibrations that in many ways a nucleus is much like a classical fluid. Still, it remains true that rotational energies are small compared to typical single-nucleon excitation energies. Therefore rotational effects must be included if the low-energy excited states are to be understood.

To better understand the discrepancy in kinetic energy, drop the dubious assumption that the nuclear material is a rigid solid. Picture the nucleus instead as a spheroid shape rotating around an axis normal to the axis of symmetry. As

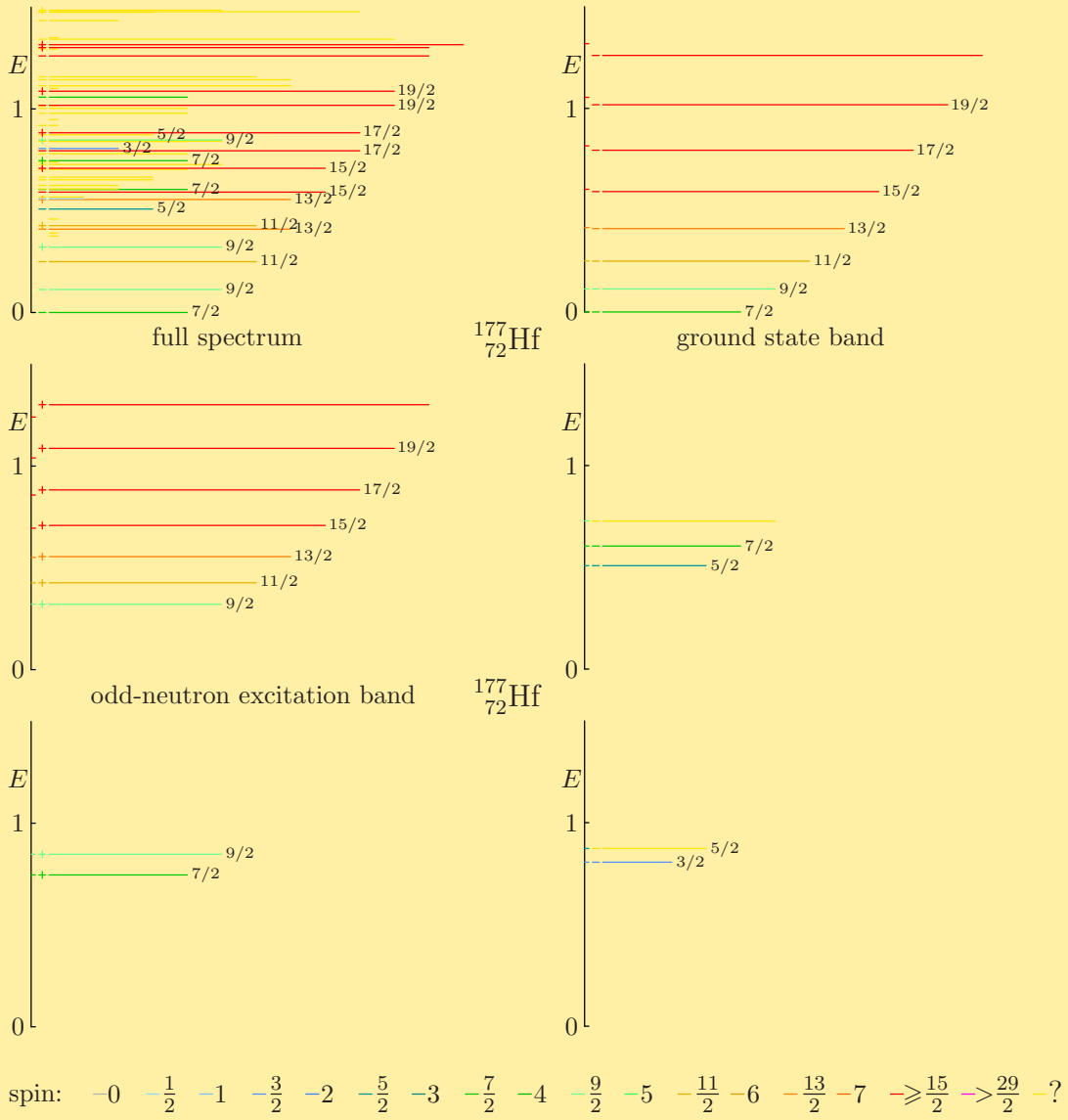


Figure 14.24: Rotational bands of hafnium-177. [pdf]

far as the individual nucleons are concerned, this shape is standing still because the nucleons are going so much faster than the nuclear shape. A typical nucleon has a kinetic energy in the order of 20 MeV, not a tenth of a MeV, and it is so much lighter than the entire nucleus to boot. Still, on the larger time scale of the nuclear rotations, the nucleons do follow the overall motion of the nuclear shape, compare chapter 7.1.5. To describe this, consider the nuclear substance to be an ideal liquid, one without internal viscosity. Without viscosity, the nuclear liquid will not pick up the overall rotation of the nuclear shape, so if the nuclear shape is spherical, the nuclear liquid will not be affected at all. This reflect the fact that

> *Nuclear rotations can only be observed in nuclei with a nonspherical equilibrium state, [31, p. 142].*

But if the rotating nuclear shape is not spherical, the nuclear liquid cannot be at rest. Then it will still have to move radially inwards or outwards to follow the changing nuclear surface radius at a given angular position. This will involve some angular motion too, but it will remain limited. (Technically speaking, the motion will remain irrotational, which means that the curl of the velocity field will remain zero.) In the liquid picture, the moment of inertia has no physical meaning and is simply *defined* by the relation $T_{\rm R} = \frac{1}{2}\mathcal{I}_{\rm R}\omega^2$, with $T_{\rm R}$ the kinetic energy of the liquid. If typical numbers are plugged into this picture, [31, p. 145], you find that the predicted rotational energies are now too high. Therefore the conclusion must be that the nuclear substance behaves like something in between a solid and an ideal liquid, at least as far as nuclear rotations are concerned. Fairly good values for the moment of inertia can be computed by modeling the nucleon pairing effect using a superfluid model, [40, pp. 493ff]

14.13.4.2 Draft: Basic rotational bands

Consider the spectrum of the deformed nucleus hafnium-177 in figure 14.24. At first the spectrum seems a mess. However, take the ground state to be an "intrinsic state" with a spin $j_{\rm min}$ equal to $^7/_2$. Then you would expect that there would also be energy levels with the nucleus still in the same state but additionally rotating as a whole. Since quantum mechanics requires that j increases in integer steps, the rotating versions of the ground state should have spin j equal to any one of $^9/_2$, $^{11}/_2$, $^{13}/_2$, ... And indeed, a sequence of such excited states can be identified in the spectrum, as shown in the top right of figure 14.24. Such a sequence of energy states is called a "rotational band." Note that all states in the band have the same parity. That is exactly what you would expect based on the classical picture of a rotating nucleus: the parity operator is a purely spatial one, so mere rotation of the nucleus in time should not change it.

How about quantitative agreement with the predicted kinetic energies of rotation (14.23)? Well, as seen in the previous subsubsection, the effective moment of inertia is hard to find theoretically. However, it can be computed from the measured energy of the $^9/_2{}^-$ rotating state relative to the $^7/_2{}^-$ ground state using (14.23). That produces a moment of inertia equal to 49% of the corresponding solid sphere value. Then that value can be used to compute the energies of the $^{11}/_2{}^-$, $^{13}/_2{}^-$, ... states, using again (14.23). The energies obtained in this way are indicated by the spin-colored tick marks on the axis in the top-right graph of figure 14.24. The lower energies agree very well with the experimental values. Of course, the agreement of the $^7/_2{}^-$ and $^9/_2{}^-$ levels is automatic, but that of the higher levels is not.

For example, the predicted energy of the $^{11}/_2{}^-$ state is 0.251 MeV, and the experimental value is 0.250 MeV. That is just a fraction of a percent error, which is very much nontrivial. For higher rotational energies, the experimental energies do gradually become somewhat lower than predicted, but nothing major. There are many effects that could explain the lowering, but an important one is "centrifugal stretching." As noted, a nucleus is not really a rigid body, and under the centrifugal forces of rapid rotation, it can stretch a bit. This increases the effective moment of inertia and hence lowers the kinetic energy, (14.23).

How about all these other excited energy levels of hafnium-177? Well, first consider the nature of the ground state. Since hafnium-177 does not have a spherical shape, the normal shell model does not apply to it. In particular, the normal shell model would have the hafnium-177's odd neutron alone in the $6f_{5/2}$ subshell; therefore it offers no logical way to explain the $^7/_2$ ground state spin. However, the Schrödinger equation can solved using a nonspherical, but still axially symmetric, potential to find suitable single particle states. Using such states, it turns out that the final odd neutron goes into a state with magnetic quantum number $^7/_2 2$ around the nuclear axis and odd parity. (For a nonspherical potential, the square angular momentum no longer commutes with the Hamiltonian and both l and j become uncertain.) With rotation, or better, with uncertainty in axial orientation, this state gives rise to the $^7/_2{}^-$ ground state of definite nuclear spin j. Increasing angular momentum then gives rise to the $^9/_2{}^-$, $^{11}/_2{}^-$, $^{13}/_2{}^-$, ... rotational band build on this ground state.

It is found that the next higher single-particle state has magnetic quantum number $^9/_2$ and even parity. If the odd neutron is kicked into that state, it produces a low-energy $^9/_2{}^+$ excited nuclear state. Adding rotational motion to this intrinsic state produces the $^9/_2{}^+$, $^{11}/_2{}^+$, $^{13}/_2{}^+$, ... rotational band shown in the middle left of figure 14.24. (Note that for this band, the experimental energies are larger than predicted. Centrifugal stretching is certainly not the only effect causing deviations from the simple theory.) In this case, the estimated moment of inertia is about 64% of the solid sphere value. There is no reason to assume that the moment of inertia remains the same if the intrinsic state of the nucleus

changes. However, clearly the value must remain sensible.

The low-lying $^5/_2^-$ state is believed to be a result of promotion, where a neutron from a $^5/_2^-$ single-particle state is kicked up to the $^7/_2^-$ state where it can pair up with the 105th neutron already there. Its rotating versions give rise to the rotational band in the middle right of figure 14.24. The moment of inertia is about 45% of the solid sphere value. The last two bands have moments of inertia of 54% and 46%, in the expected ballpark.

The general approach as outlined above has been extraordinarily successful in explaining the excited states of deformed nuclei, [31, p. 156].

14.13.4.3 Draft: Bands with intrinsic spin one-half

The semi-classical explanation of rotational bands was very simplistic. While it works fine if the intrinsic spin of the rotating nuclear state is at least one, it develops problems if it becomes one-half or zero. The most tricky case is spin one-half.

Despite less than stellar results in the past, the discussion of the problem will stick with a semi-classical approach. Recall first that angular momentum is a vector. In vector terms, the total angular momentum of the nucleus consists of rotational angular momentum and intrinsic angular momentum of the nonrotating nucleus:

$$\vec{J} = \vec{J}_{\rm rot} + \vec{J}_{\rm min}$$

Now in the expression for rotational energy, (14.23) it was implicitly assumed that the square angular momentum of the nucleus is the sum of the square angular momentum of rotation plus the square angular momentum of the intrinsic state. But classically the Pythagorean theorem shows that this is only true if the two angular momentum vectors are orthogonal.

Indeed, a more careful quantum treatment, [40, pp. 356-389], gives rise to a semi-classical picture in which the axis of the rotation is normal to the axis of symmetry of the nucleus. In terms of the inviscid liquid model of subsubsection 14.13.4.1, rotation about an axis of symmetry "does not do anything." That leaves only the intrinsic angular momentum for the component of angular momentum along the axis of symmetry. The magnetic quantum number of this component is $j_{\rm min}$, equal to the spin of the intrinsic state. Correct that: the direction of the axis of symmetry should not make a difference. Therefore, the complete wave function should be an equal combination of a state $|j_{\rm min}\rangle$ with magnetic quantum number $j_{\rm min}$ along the axis and a state $|-j_{\rm min}\rangle$ with magnetic quantum $-j_{\rm min}$.

Next, the kinetic energy of rotation is, since $\vec{J}_{\rm rot} = \vec{J} - \vec{J}_{\rm min}$,

$$\frac{1}{2\mathcal{I}_{\rm R}} \vec{J}_{\rm rot}^{\,2} = \frac{1}{2\mathcal{I}_{\rm R}} \vec{J}^{\,2} - \frac{1}{\mathcal{I}_{\rm R}} \vec{J} \cdot \vec{J}_{\rm min} + \frac{1}{2\mathcal{I}_{\rm R}} \vec{J}_{\rm min}^{\,2}$$

As long as the middle term in the right hand side averages away, the normal formula (14.23) for the energy of the rotational states is fine. This happens if there is no correlation between the angular momentum vectors $\vec{J}$ and $\vec{J}_{\rm min}$, because then opposite and parallel alignments will cancel each other.

But not too quick. There is an obvious correlation since the axial components are equal. The term can be written out in components to give

$$\frac{1}{\mathcal{I}_{\rm R}} \vec{J} \cdot \vec{J}_{\rm min} = \frac{1}{\mathcal{I}_{\rm R}} \left[J_x J_{x,\rm min} + J_y J_{y,\rm min} + J_z J_{z,\rm min} \right]$$

where the z-axis is taken as the axis of symmetry of the nucleus. Now think of these components as quantum operators. The z components are no problem: since the magnetic quantum number is constant along the axis, this term will just shift the all energy levels in the band by the same amount, leaving the spacings between energy levels the same.

However, the x and y components have the effect of turning a state $|\pm j_{\rm min}\rangle$ into some combination of states $|\pm j_{\rm min} \pm 1\rangle$, chapter 12.11. Since there is no rotational momentum in the axial direction, $\vec{J}$ and $\vec{J}_{\rm min}$ have quite similar effects on the wave function, but it is not the same, for one because $\vec{J}$ "sees" the complete nuclear momentum. If $j_{\rm min}$ is 1 or more, the effects remain inconsequential: then the produced states are part of a different vibrational band, with different energies. A bit of interaction between states of different energy is usually not a big deal, chapter 5.3. But if $j_{\rm min} = \frac{1}{2}$ then $|j_{\rm min} - 1\rangle$ and $|-j_{\rm min} + 1\rangle$ are part of the state itself. In that case, the x and y components of the $\vec{J} \cdot \vec{J}_{\rm min}$ term produces a contribution to the energy that does not average out, and the larger $\vec{J}$ is, the larger the contribution.

The expression for the kinetic energy of nuclear rotation then becomes

$$\boxed{T_{\rm R} = \frac{\hbar^2}{2\mathcal{I}_{\rm R}} \left\{ \Big[j(j+1) + a(-1)^{j+\frac{1}{2}} (j + \tfrac{1}{2}) \Big] - \left[j_{\rm min}(j_{\rm min}+1) - a \right] \right\} \quad j_{\rm min} = \tfrac{1}{2}} \tag{14.24}$$

where a is a constant. Note that the additional term is alternatingly positive and negative. Averaged over the entire rotational band, the additional term does still pretty much vanish.

As an example, consider the $\frac{1}{2}^-$ ground state rotational band of tungsten-183, figure 14.25. To compute the rotational kinetic energies in the band using (14.24) requires the values of both $\mathcal{I}_{\rm R}$ and a. The measured energies of the $\frac{3}{2}^-$ and $\frac{5}{2}^-$ states above the ground state can be used to do so. That produces a moment of inertia equal to 45% of the solid sphere value and a nondimensional constant a equal to 0.19. Next then the formula can be used to predict the energies of the remaining states in the band. As the axial tick marks in the right graph of figure 14.25 show, the prediction is quite good. Note in particular that the energy interval between the $\frac{9}{2}^-$ and $\frac{7}{2}^-$ states is *less* than that between

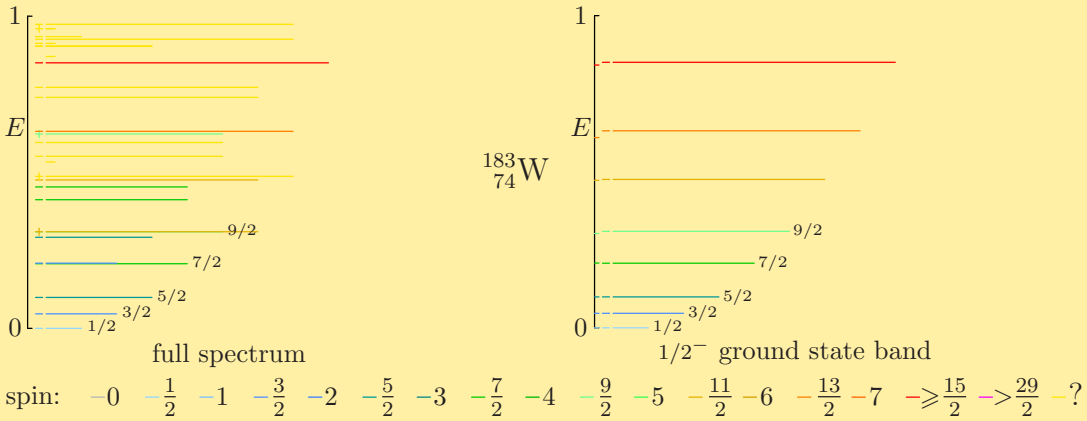


Figure 14.25: Ground state rotational band of tungsten-183. [pdf]

the $^7/_2^-$ and $^5/_2^-$ states. Without the alternating term, there would be no way to explain that.

Much larger values of a are observed for lighter nuclei. As an example, consider aluminum-25 in figure 14.26. This nucleus has been studied in detail, and a number of bands with an intrinsic spin 1/2 have been identified. Particularly interesting is the $^1/_2^-$ band in the bottom left of figure 14.26. For this band $a = -3.2$, and that is big enough to change the order of the states in the band! For this nucleus, the moments of inertia are 70%, 96%, 107%, 141% and 207% respectively of the solid sphere value.

14.13.4.4 Draft: Bands with intrinsic spin zero

The case that the intrinsic state has spin zero is particularly important, because all even-even nuclei have a 0^+ ground state. For bands build on a zero-spin intrinsic state, the thing to remember is that the only values in the band are even spin and even parity: 0^+, 2^+, 4^+, ...

This can be thought of as a consequence of the fact that the $|j_{\min}\rangle$ and $|-j_{\min}\rangle$ states of the previous subsubsection become equal. Their odd or even combinations must be constrained to prevent them from canceling each other.

As an example, consider erbium-164. The ground state band in the top right of figure 14.27 consists of the states 0^+, 2^+, 4^+, ... as expected. The energies initially agree well with the theoretical prediction (14.23) shown as tick marks. For example, the prediction for the 4^+ level has less than 2% error. Deviations from theory show up at higher angular momenta, which may have to do with centrifugal stretching.

Other bands have been identified that are build upon vibrational intrinsic states. (A β or beta vibration maintains the assumed intrinsic axial symmetry of the nucleus, a γ or gamma one does not.) Their energy spacings follow (14.23)

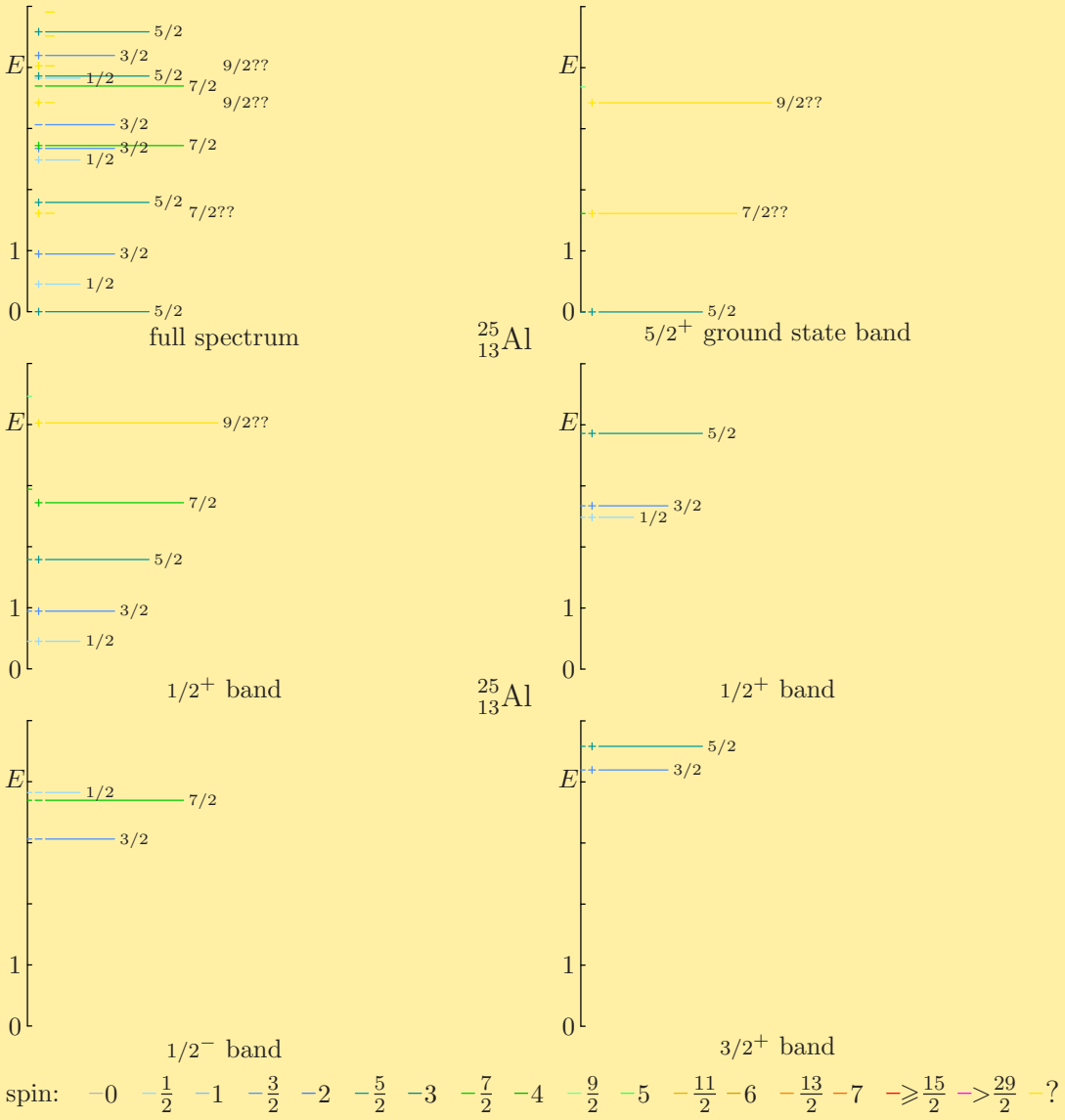


Figure 14.26: Rotational bands of aluminum-25. [pdf]

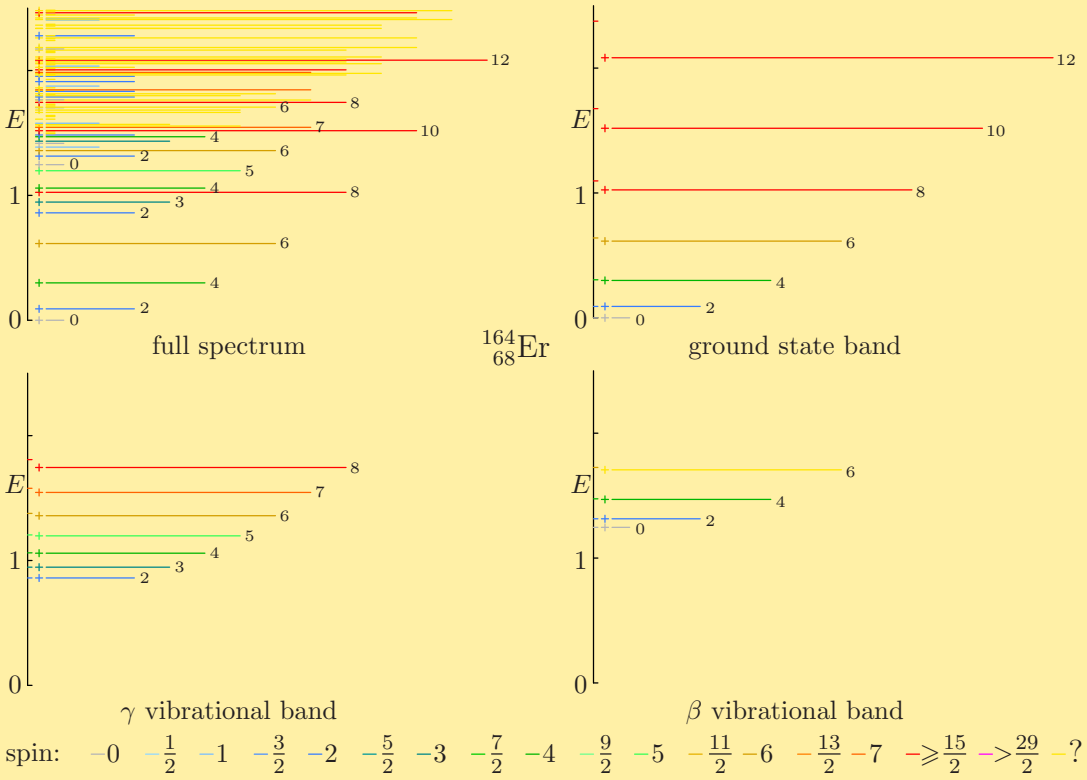


Figure 14.27: Rotational bands of erbium-164. [pdf]

again well. The moments of inertia are 46%, 49% and 61% of the solid sphere value.

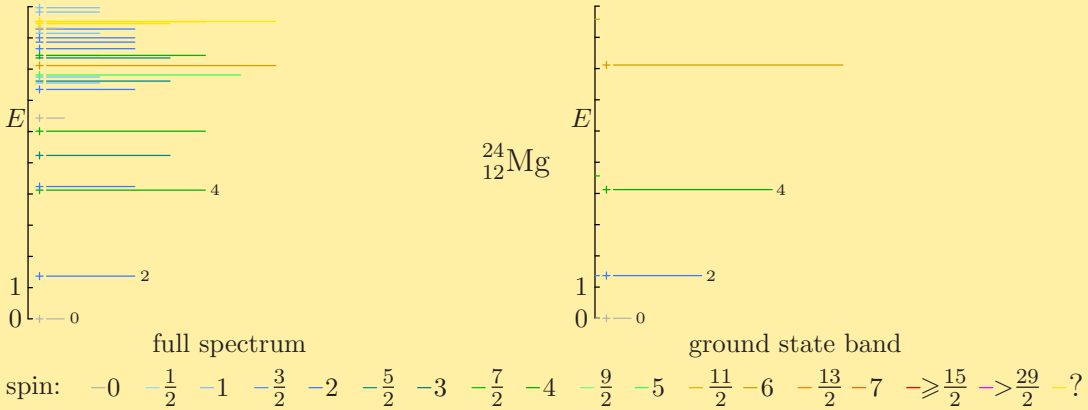


Figure 14.28: Ground state rotational band of magnesium-24. [pdf]

Light even-even nuclei can also be deformed and show rotational bands. As an example, figure 14.28 shows the ground state band of magnesium-24. The moment of inertia is 75% of the solid sphere value.

It may also be mentioned that nuclei with intrinsic spin zero combined with an octupole vibration can give rise to bands that are purely odd spin/odd parity ones, 1^-, 3^-, 5^-, ..., [40, p. 368]. The lowest odd parity states for erbium-164 are 1^- and 3^- ones, with no 2^- state in between, for a really high estimated moment of inertia of 148%, and a potential 5^- state at roughly, but not accurately, the predicted position. Anomalous bands that have the parity inverted may also be possible; hafnium-176 is believed to have a couple of excited states like that, 0^- at 1.819 MeV and 2^- at 1.857 MeV, with a moment of inertia of 98%.

Centrifugal effects can be severe enough to change the internal structure of the nucleus nontrivially. Typically, zero-spin pairings between nucleons may be broken up, allowing the nucleons to start rotating along with the nucleus. That creates a new band build on the changed intrinsic state. Physicists then define the state of lowest energy at a given angular momentum as the "yrast state." The term is not an acronym, but Swedish for "that what rotates more." For a discussion, a book for specialists will need to be consulted.

14.13.4.5 Draft: Even-even nuclei

All nuclei with even numbers of both protons and neutrons have a 0^+ ground state. For nonspherical ones, the rotational model predicts a ground state band of low-lying 2^+, 4^+, 6^+, ... states. The ratio of the energy levels of the 4^+ and

2^+ states is given by (14.23)

$$\frac{\hbar^2}{2\mathcal{I}_{\rm R}} 4(4+1) \Big/ \frac{\hbar^2}{2\mathcal{I}_{\rm R}} 2(2+1) = \frac{10}{3}$$

For spherical nuclei, the vibrational model also predicts a 2^+ lowest excited state, but the 4^+ excited state is now part of a triplet, and the triplet has only twice the energy of the 2^+ state. Therefore, if the ratio of the energy of the second excited state to the lowest 2^+ state is plotted, as done in figure 14.22, then vibrating nuclei should be indicated by a value 2 (green) and rotating nuclei by a value 3.33 (red). If the figure is examined, it may be open to some doubt whether green squares are necessarily vibrational, but the red squares quite nicely locate the rotational ones.

In the figure, nuclei marked with "V" have 0^+, 2^+, and a 0^+, 2^+, 4^+ triplet as the lowest 5 energy states, the triplet allowed to be in any order. Nuclei marked with an "R" have the sequence 0^+, 2^+, 4^+, and 6^+ as the lowest four energy states. Note that this criterion misses the light rotational nuclei like magnesium-24; for light nuclei the rotational energies are not small enough to be well separated from shell effects. Near the stable line, rotating nuclei are found in the approximate mass number ranges $20 < A < 30$, $150 < A < 190$, and $220 < A$. However, away from the stable line rotating nuclei are also found at other mass numbers.

14.13.4.6 Draft: Nonaxial nuclei

While most nuclei are well modeled as axially symmetric, some nuclei are not. For such nuclei, an ellipsoidal model can be used with the three major axes all off different length. There are now two nondimensional axis ratios that characterize the nucleus, rather than just one.

This additional nondimensional parameter makes the spectrum much more complex. In particular, in addition to the normal rotational bands, associated "anomalous" secondary bands appear. The first is a 2^+, 3^+, 4^+, ... one, the second a 4^+, 5^+, ... one, etcetera. The energies in these bands are not independent, but related to those in the primary band.

Figure 14.29 shows an example. The primary ground state band in the top right quickly develops big deviations from the axially symmetric theory (14.23) values (thin tick marks.) Computation using the ellipsoidal model for a suitable value of the deviation from axial symmetry is much better (thick tick marks.) The predicted energy levels of the first anomalous band also agree well with the predicted values. The identification of the bands was taken from [40, p. 416], but since they do not list the energies of the second anomalous band, that value was taken from [36, p. 388].

In the limit that the nuclear shape becomes axially symmetric, the anomalous bands disappear towards infinite energy. In the limit that the nuclear shape

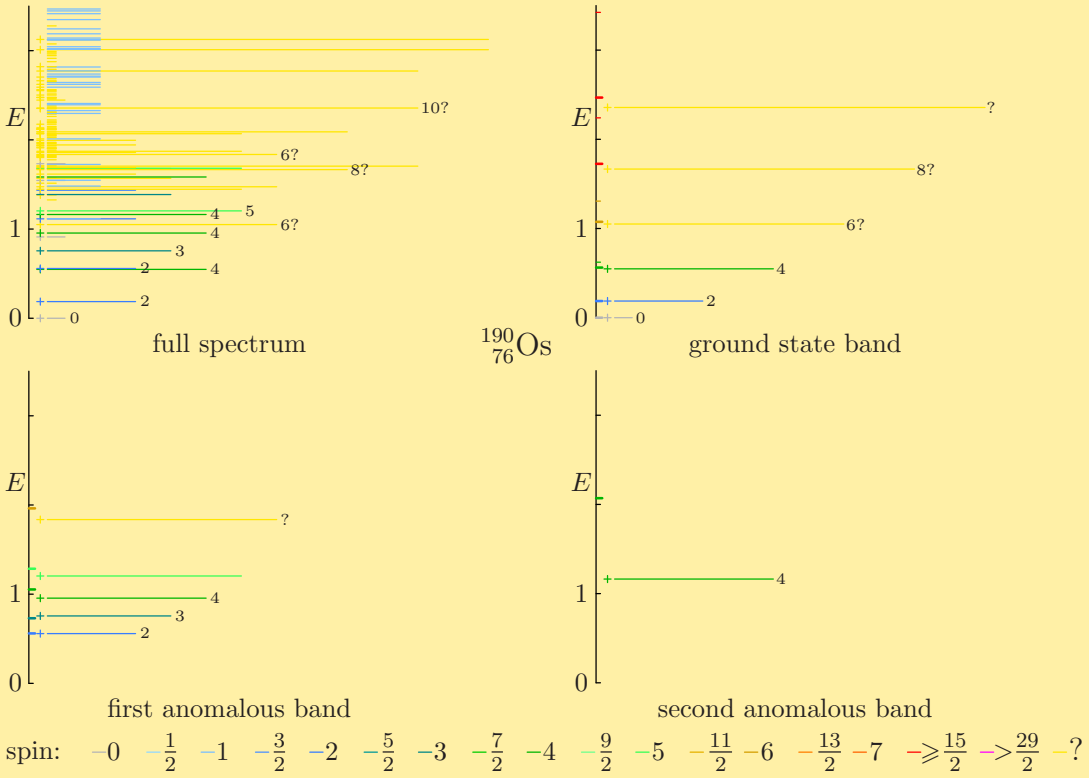


Figure 14.29: Rotational bands of osmium-190. [pdf]

becomes spherical, all states in the primary bands except the lowest one also disappear to infinity, assuming that the "moment of inertia" becomes zero as the ideal liquid model says.

14.14 Draft: Fission

In spontaneous fission, a very heavy nucleus falls apart into big fragments. If there are two fragments, it is called binary fission. In some cases, there are three fragments. That is called ternary fission; the third fragment is usually an alpha particle. This section summarizes some of the basic ideas.

14.14.1 Draft: Basic concepts

What makes fission energetically possible is that very heavy nuclei have less binding energy per nucleon than those in the nickel/iron range, as shown earlier in figure 14.4. The main culprit is the Coulomb repulsion between the protons. It has a much longer range than the nuclear attractions between nucleons. Therefore, Coulomb repulsion disproportionally increases the energy for heavy nuclei. If a nucleus like uranium-238 divides cleanly into two palladium-119 nuclei, the energy liberated is on the order of 200 MeV (200 000 000 eV). That is obviously a very large amount of energy. Chemical reactions produce maybe a few eV per atom.

The liquid drop model predicts that the nuclear shape will become unstable at Z^2/A about equal to 48. However, only the weirdest nuclei like ${}^{293}_{118}\mathrm{Ei}$ come close to that value. Below $Z = 100$ the nuclei that decay primarily through spontaneous fission are curium-250, with Z^2/A equal to 37 and a half life of 8 300 years, californium-254, 38 and two months, and fermium-256, 39 and less than 3 hours.

Indeed, while the fission products may have lower energy than the original nucleus, in taking the nucleus apart, the nuclear binding energy must be provided right up front. On the other hand the Coulomb energy gets recovered only after the fragments have been brought far apart. As a result, there is normally a energy barrier that must be crossed for the nucleus to come apart. That means that an "activation energy" must be provided in nuclear reactions, much like in most chemical reactions.

For example, uranium has an activation energy of about 6.5 MeV. By itself, uranium-235 will last a billion years or so. However, it can be made to fission by hitting it with a neutron that has only a thermal amount of energy. (Zero is enough, actually.) When hit, the nucleus will fall apart into a couple of big pieces and immediately release an average of 2.4 "prompt neutrons." These new neutrons allow the process to repeat for other uranium-235 nuclei, making a "chain reaction" possible.

In addition to prompt neutrons, fusion processes may also emit a small fraction of "delayed neutrons" neutrons somewhat later. Despite their small number, they are critically important for controlling nuclear reactors because of their slower response. If you tune the reactor so that the presence of delayed neutrons is essential to maintain the reaction, you can control it mechanically on their slower time scale.

Returning to spontaneous fission, that is possible without the need for an activation energy through quantum mechanical tunneling. Note that this makes spontaneous fission much like alpha decay. However, as section 14.11.2 showed, there are definite differences. In particular, the basic theory of alpha decay does not explain why the nucleus would want to fall apart into two big pieces, instead of one big piece and a small alpha particle. This can only be understood qualitatively in terms of the liquid drop model: a charged classical liquid drop is most unstable to large-scale deformations, not small scale ones, subsection 14.13.1.

14.14.2 Draft: Some basic features

While fission is qualitatively close to alpha decay, its actual mechanics is much more complicated. It is still an area of much research, and beyond the scope of this book. A very readable description is given by [36]. This subsection describes some of the ideas.

From a variety of experimental data and their interpretation, the following qualitative picture emerges. Visualize the nucleus before fission as a classical liquid drop. It may already be deformed, but the deformed shape is classically stable. To fission, the nucleus must deform more, which means it must tunnel through more deformed states. When the nucleus has deformed into a sufficiently elongated shape, it becomes energetically more favorable to reduce the surface area by breaking the connection between the ends of the nucleus. The connection thins and eventually breaks, leaving two separate fragments. During the messy process in which the thin connection breaks an alpha particle could well be ejected. Now typical heavy nuclei contain relatively more neutrons than lighter ones. So when the separated fragments take inventory, they find themselves overly neutron-rich. They may well find it worthwhile to eject one or two right away. This does not change the strong mutual Coulomb repulsion between the fragments, and they are propelled to increasing speed away from each other.

Consider now a very simple model in which a nucleus like fermium-256 falls cleanly apart into two smaller nuclear fragments. As a first approximation, ignore neutron and other energy emission in the process and ignore excitation of the fragments. In that case, the final kinetic energy of the fragments can be computed from the difference between their masses and the mass of the original nucleus.

In the fission process, the fragments supposedly pick up this kinetic energy from the Coulomb repulsion between the separated fragments. If it is assumed that the fragments are spherical throughout this final phase of the fission process, then its properties can be computed. In particular, it can be computed at which separation between the fragments the kinetic energy was zero. That is important because it indicates the end of the tunneling phase. Putting in the numbers, it is seen that the separation between the fragments at the end of tunneling is at least 15% more than that at which they are touching. So the model is at least reasonably self-consistent.

Figure 14.30 shows the energetics of this model. Increasing redness indicates increasing energy release in the fission. Also, the spacing between the squares indicates the spacing between the nuclei at the point where tunneling ends. Note in particular the doubly magic point of 50 protons and 82 neutrons. This point is very neutron rich, just what is needed for fission fragments. And because it is doubly magic, nuclei in this vicinity have unusually high binding energy, as seen from figure 14.9. Indeed, nuclei with 50 protons are seen to have the highest fission energy release in figure 14.30. Also, they have the smallest relative spacing between the nuclei at the end of tunneling, so likely the shortest relative distance that must be tunneled through. The conclusion is clear. The logical thing for fermium-256 to do is to come apart into two almost equal fragments with a magic number of 50 protons and about 78 neutrons, giving the fragments a mass number of 128. Less plausibly, one fragment could have the magic number of 82 neutrons, giving fragment mass numbers of 132 and 124. But the most unstable deformation for the liquid drop model is symmetric. And so is a spheroidal or ellipsoidal model for the deformed nucleus. It all seems to add up very nicely. The fragments must be about the same size, with a mass number of 128.

Except that that is all wrong.

Fermium 258 acts like that, and fermium-257 also mostly, but not fermium 256. It is rare for fermium-256 to come apart into two fragments of about equal size. Instead, the most likely mass number of the large fragment is about 140, with only a small probability of a mass number 132 or lower. A mass number of 140 clearly does not seem to make much sense based on figure 14.30.

The precise solution to this riddle is still a matter of current research, but physicists have identified quantum effects as the primary cause. The potential energy barrier that the fissioning nucleus must pass through is relatively low, on the order of say 5 MeV. That is certainly small enough to be significantly affected by quantum shell effects. Based on that idea, you would expect that mass asymmetry would decrease if the excitation energy of the nucleus is increased, and such an effect is indeed observed. Also, the separation of the fragments occurs at very low energy, and is believed to be slow enough that the fragments can develop some shell structure. Physicists have found that for many fissioning nuclei, quantum shell effects can create a relatively stable intermediate state in

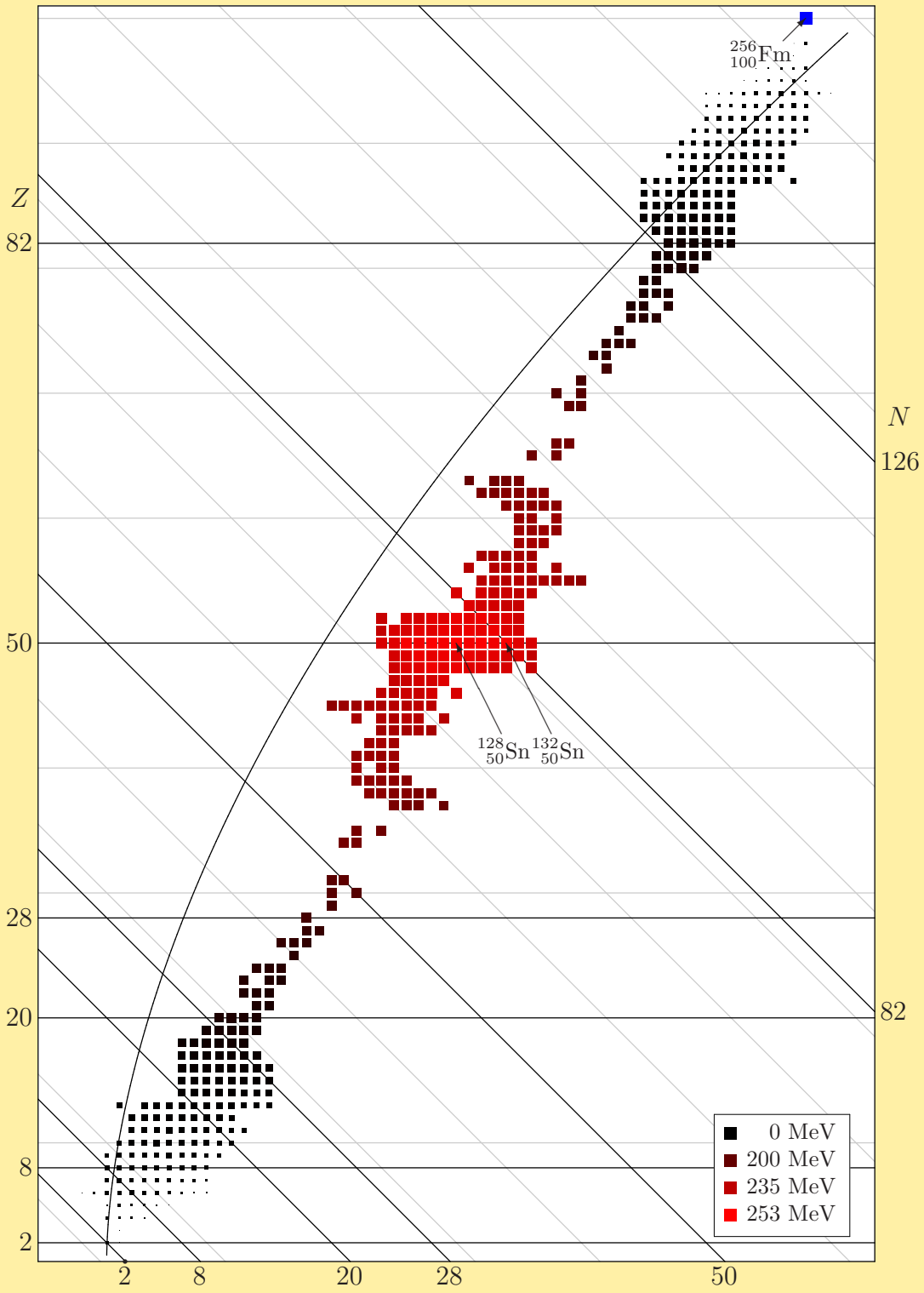


Figure 14.30: Simplified energetics for fission of fermium-256. [pdf][con]

the fission process. Such a state produces resonances in response to specific excitation energies of the nucleus. Shell corrections can also lower the energy of asymmetric nuclear fissioning shapes below those of symmetric ones, providing an explanation for the mass asymmetry.

Imagine then a very distorted stage in which a neutron-rich, doubly magic 50/82 core develops along with a smaller nuclear core, the two being connected by a cloud of neutrons and protons. Each could pick up part of the cloud in the final separation process. That picture would explain why the mass number of the large fragment exceeds 132 by a fairly constant amount while the mass number of the smaller segment varies with the initial nuclear mass. Whether or not there is much truth to this picture, at least it is a good mnemonic to remember the fragment masses for the nuclei that fission asymmetrically.

14.15 Draft: Spin Data

The net internal angular momentum of a nucleus is called the "nuclear spin." It is an important quantity for applications such as NMR and MRI, and it is also important for what nuclear decays and reactions occur and at what rate. One previous example was the categorical refusal of bismuth-209 to decay at the rate it was supposed to in section 14.11.3.

This section provides an overview of the ground-state spins of nuclei. According to the rules of quantum mechanics, the spin must be integer if the total number of nucleons is even, and half-integer if it is odd. The shell model can do a pretty good job of predicting actual values. Historically, this was one of the major reasons for physicists to accept the validity of the shell model.

14.15.1 Draft: Even-even nuclei

For nuclei with both an even number of protons and an even number of neutrons, the odd-particle shell model predicts that the spin is zero. This prediction is fully vindicated by the experimental data, figure 14.31. There are no known exceptions to this rule.

14.15.2 Draft: Odd mass number nuclei

Nuclei with an odd mass number A have either an odd number of protons or an odd number of neutrons. For such nuclei, the odd-particle shell model predicts that the nuclear spin is the net angular momentum (orbital plus spin) of the last odd nucleon. To find it, the subshell that the last particle is in must be identified. That can be done by assuming that the subshells fill in the order given in section 14.12.2. This ordering is indicated by the colored lines in figures 14.32 and 14.33. Nuclei for which the resulting nuclear spin prediction is correct

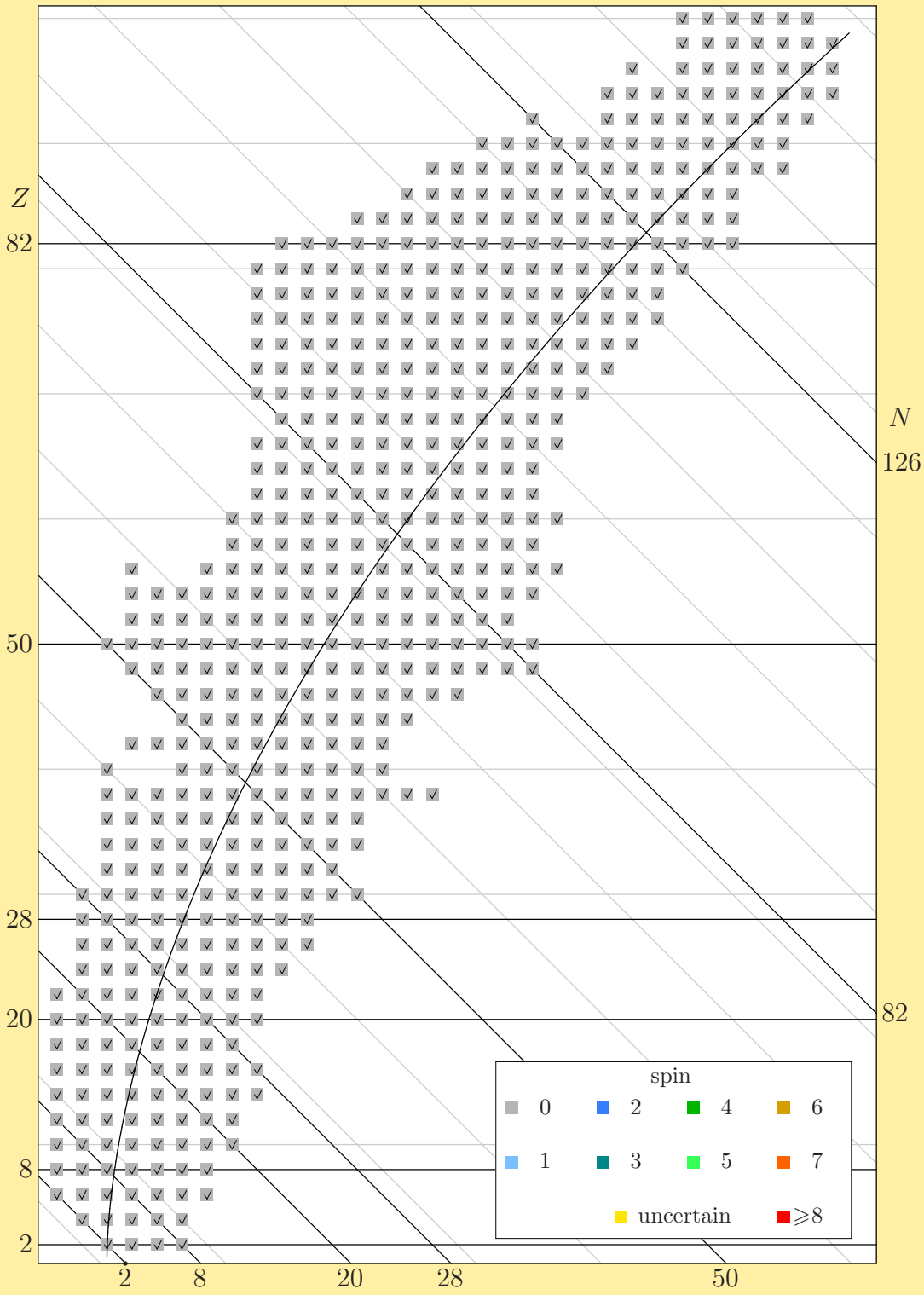


Figure 14.31: Spin of even-even nuclei. [pdf][con]

are indicated with a black check mark. (If the check mark is white rather than black, some reservation was expressed about the measured spin in NUBASE 2003.)

The prediction is correct right off the bat for a considerable number of nuclei. That is very much nontrivial. Still, there is an even larger number for which the prediction is not correct.

One major reason is that many heavy nuclei are not spherical in shape. The shell model was derived assuming a spherical nuclear shape and simply does not apply for such nuclei. These are the nonspherical, rotational nuclei, the ones that are easily excited; the nonsmall squares in 14.45, the red (or yellow) big "R" squares in figure 14.22, the small squares in figure 14.19. Their main regions are for neutron number N above 132 and in the deep interior of the $Z < 82$, $N > 82$ wedge. For these nuclei, the shell model simply does not work.

For almost all remaining nuclei near the stable line, the spin can be explained in terms of the shell model using various reasonable excuses, [36, p. 224ff]. Some excuses are marked.

In particular, if the subshell with the odd neutron is above a subshell of lower spin, a particle from the lower subshell may be promoted to the higher one. This particle can then pair up at a higher spin, which is believed to be energetically favorable. Since an odd nucleon occurs now in the lower shell, the spin of the nucleus is predicted to be the one of that shell. So the nuclear spin is lowered compared to the bare shell model.

Nuclei for which such spin lowering due to promotion can explain the observed spin are indicated with an "L" or "l" in figures 14.32 and 14.33. For the nuclei marked with "L," the odd nucleon cannot be in the normal subshell because the nucleus has the wrong parity for that. Therefore, for these nuclei there is a solid additional reason besides the spin to assume that promotion has occurred.

Promotion was only allowed for subshells immediately above one with lower spin in the same major shell, so the nucleon could only be promoted a single subshell. The idea is that the gained pairing energy should not be big enough to make major modifications to the shell model.

For some nuclei, the basic shell model is valid but the odd-particle assumption fails. The odd particle assumption implies that all nucleons except the odd one pair up in states of zero spin. But sometimes at least three nucleons combine in a state with one unit less spin than the single odd particle would have. This is only possible for subshells with at least three particles and three holes (empty spots for additional nucleons). Nuclei for which this unit spin reduction might to be the case are marked with a minus sign. It is evident, for example, for odd nucleon numbers 23 and 25.

Fluorine-19 and its mirror twin neon-19 are rare outright failures of the shell model, as already discussed in section 14.12.6.

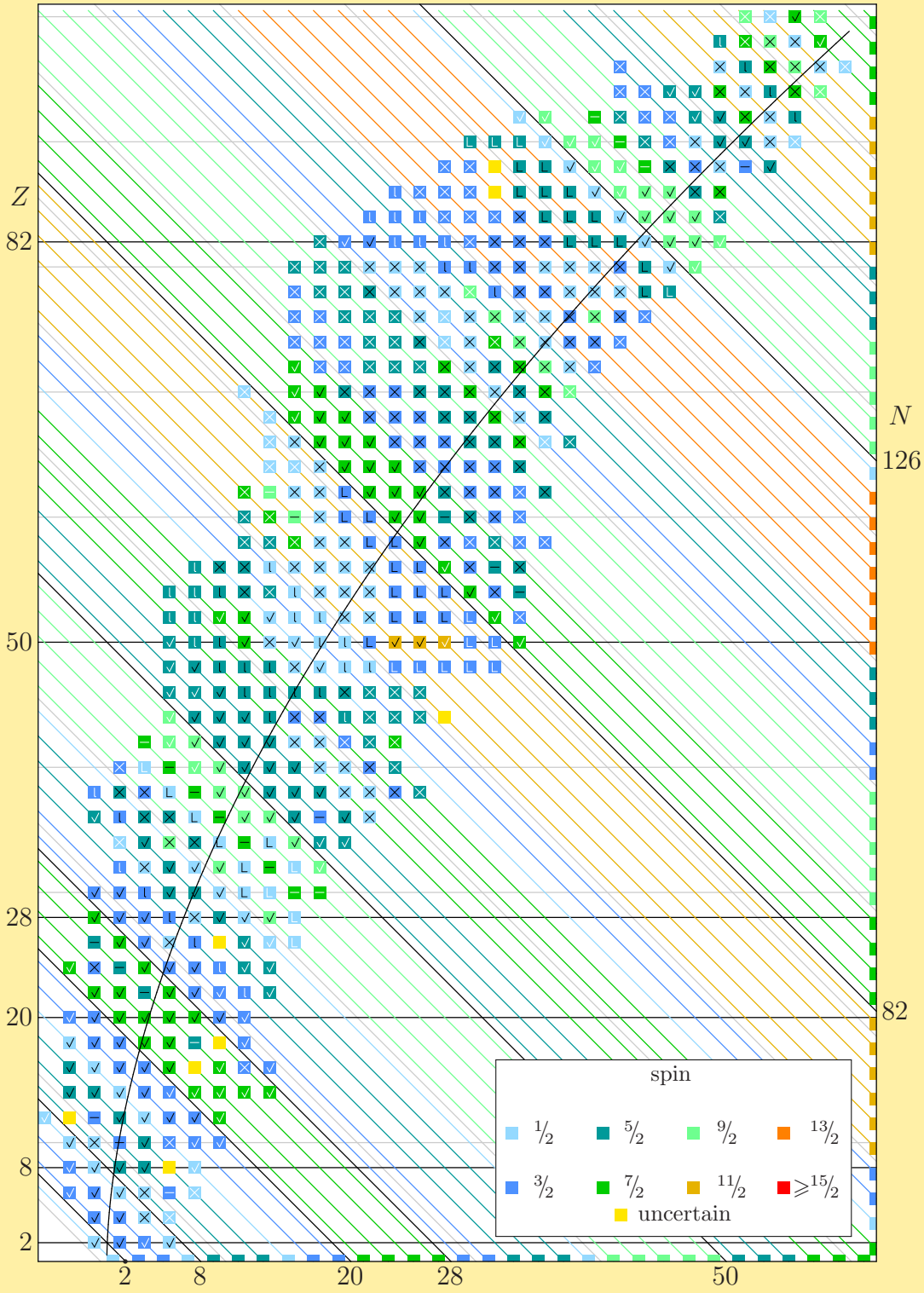


Figure 14.32: Spin of even-odd nuclei. [pdf][con]

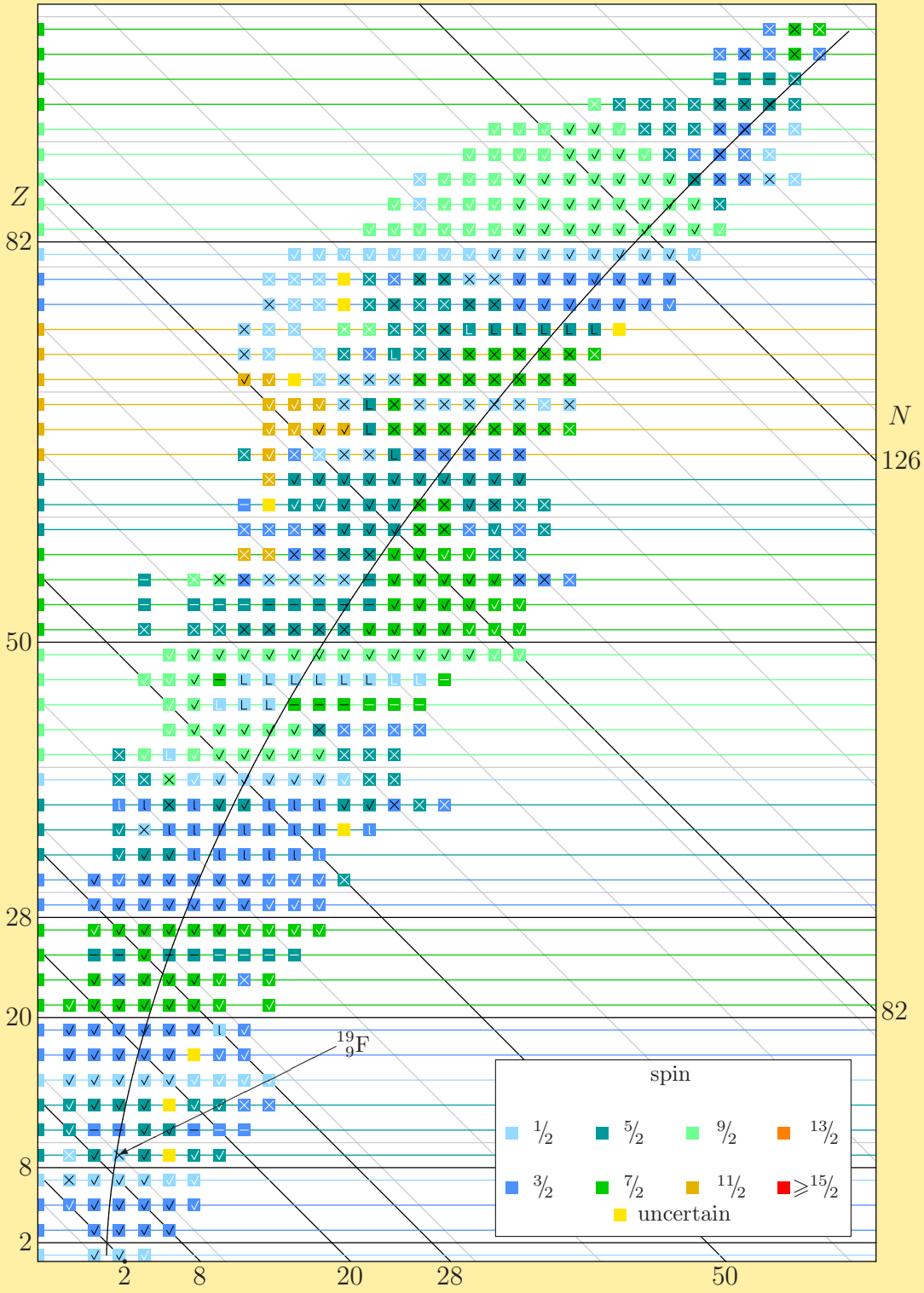


Figure 14.33: Spin of odd-even nuclei. [pdf][con]

Among the remaining failures, notable are nuclei with odd proton numbers just above 50. The $5g_{7/2}$ and $5d_{5/2}$ subshells are very close together and it depends on the details which one gets filled first.

In a few exceptional cases, like the highly unstable nitrogen-11 and beryllium-11 "halo" nuclei, the theoretical model predicted the right spin, but it was not counted as a hit because the nuclear parity was inconsistent with the predicted subshell. In all cases, it was demanded that the nuclear parity, if known, did not conflict with the proposed shell model verification (or explanation) of the spin.

For nuclei marked with a cross, no explanation for the spin using the above rules could be found. In general, you might not want to take nuclei well away from the stable line that serious. For yellow squares, NUBASE 2003 gave either no value for the spin or more than one possible value.

14.15.3 Draft: Odd-odd nuclei

If both the number of protons and the number of neutrons is odd, the nuclear spin becomes much more difficult to predict. According to the odd-particle shell model, the net nuclear spin j_N comes from combining the net angular momenta j^p of the odd proton and j^n of the odd neutron. Then according to quantum mechanics, the net nuclear spin j_N can be any integer in the range

$$|j^p - j^n| \leqslant j_N \leqslant j^p + j^n \tag{14.25}$$

That gives a total of $2\min(j^p, j^n) + 1$ different possibilities, or at least two.

That is not very satisfactory of course. You would like to get a specific prediction for the spin, not a range. The so-called "Nordheim rules" attempt to do so. The underlying idea is that nuclei like to align the spins of the odd proton and odd neutron, just like the deuterium nucleus does.

To describe the rules, forget about quantum mechanics for now. Just think of the spin and orbital angular momenta involved as simple vectors that can either point up or down. And take "up" to be the direction that the aligned proton and neutron spin vectors s^p and s^n point. Now suppose first that for both neutron and proton the orbital angular momenta l^p and l^n also point upwards (or are zero). Then for both proton and neutron, the orbital and spin angular momenta add up to a total momentum $j^p = l^p + s^p$ respectively $j^n = l^n + s^n$ that also point upwards. So the sum of the two, the total nuclear spin j_N points upwards and has magnitude $j_N = j^p + j^n$.

Conversely, if both orbital momenta point downwards (and are nonzero), then spin and orbital angular momenta are in opposite directions and subtract. Then proton and neutron have net angular momenta of magnitude $j^p = l^p - s^p$ respectively $j^n = l^n - s^n$ that point downwards. (Recall from quantum mechanics that l is at least 1 if nonzero, so is bigger than the $s = \frac{1}{2}$ pointing upwards.)

The combined nuclear angular momentum is then again $j_N = j^p + j^n$ (pointing downwards).

However, if for one nucleon the orbital angular momentum is zero or points in the direction of the spin and the other is nonzero and points in the direction opposite to the spin, then one of j^p and j^n points upwards and the other downwards. That then means that now they are in opposite directions and subtract; $j_N = |j^p - j^n|$.

That then gives the Nordheim rules as, [36, p. 239]:

1. If for both proton and neutron, $j = l + s$, or for both $j = l - s$, then the angular momenta of the two add up and $j_N = j^p + j^n$.
2. Otherwise they subtract and $j_N = |j^p - j^n|$.
3. New and improved version: if number 1 above fails, assume that the two angular momenta are opposite anyway and the spin is $j = |j^p - j^n|$ like in number 2.

Of course, the real quantum rules for angular momentum are a lot more complicated than the simplified picture above. Note in particular from the Clebsch-Gordan coefficients in figure 12.5 that if $j = l - s$, then that nucleon cannot be just in the spin-up state. The two nucleons cannot fully align then.

To check those rules is not trivial, because it requires the values of l for the odd proton and neutron. Who will say in what shells the odd proton and odd neutron really are? The simplest solution is to simply take the shells to be the ones that the shell model predicts, assuming the subshell ordering from section 14.12.2. The nuclei that satisfy the Nordheim rules under that assumption are indicated with a check mark in figure 14.34. A blue check mark means that the new and improved version has been used. (Yellow is used if there is uncertainty about the measurement.) It is seen that the rules get a number of nuclei right.

An "L" or "l" indicates that it has been assumed that the spin of at least one odd nucleon has been lowered due to promotion. The rules are the same as in the previous subsection. In case of "L," the Nordheim rules were really verified. More specifically, for these nuclei there was no possibility consistent with nuclear spin and parity to violate the rules. For nuclei with an "l" there was, and the case that satisfied the Nordheim rules was cherry-picked among other otherwise valid possibilities that did not.

A further weakening of standards applies to nuclei marked with "N" or "n." For those, one or two subshells of the odd nucleons were taken based on the spins of the immediately neighboring nuclei of odd mass number. For nuclei marked with "N" the Nordheim rules were again really verified, with no possibility of violation within the now larger context. For nuclei marked "n," other possibilities violated the rules; obviously, for these nuclei the standards have become miserably low. Note how many "correct" predictions there are in the regions of nonspherical nuclei in which the shell model is quite meaningless.

Preston and Bhaduri [36, p. 239] suggest that the proton and neutron angular momenta be taken from the neighboring pairs of nuclei of odd mass number.

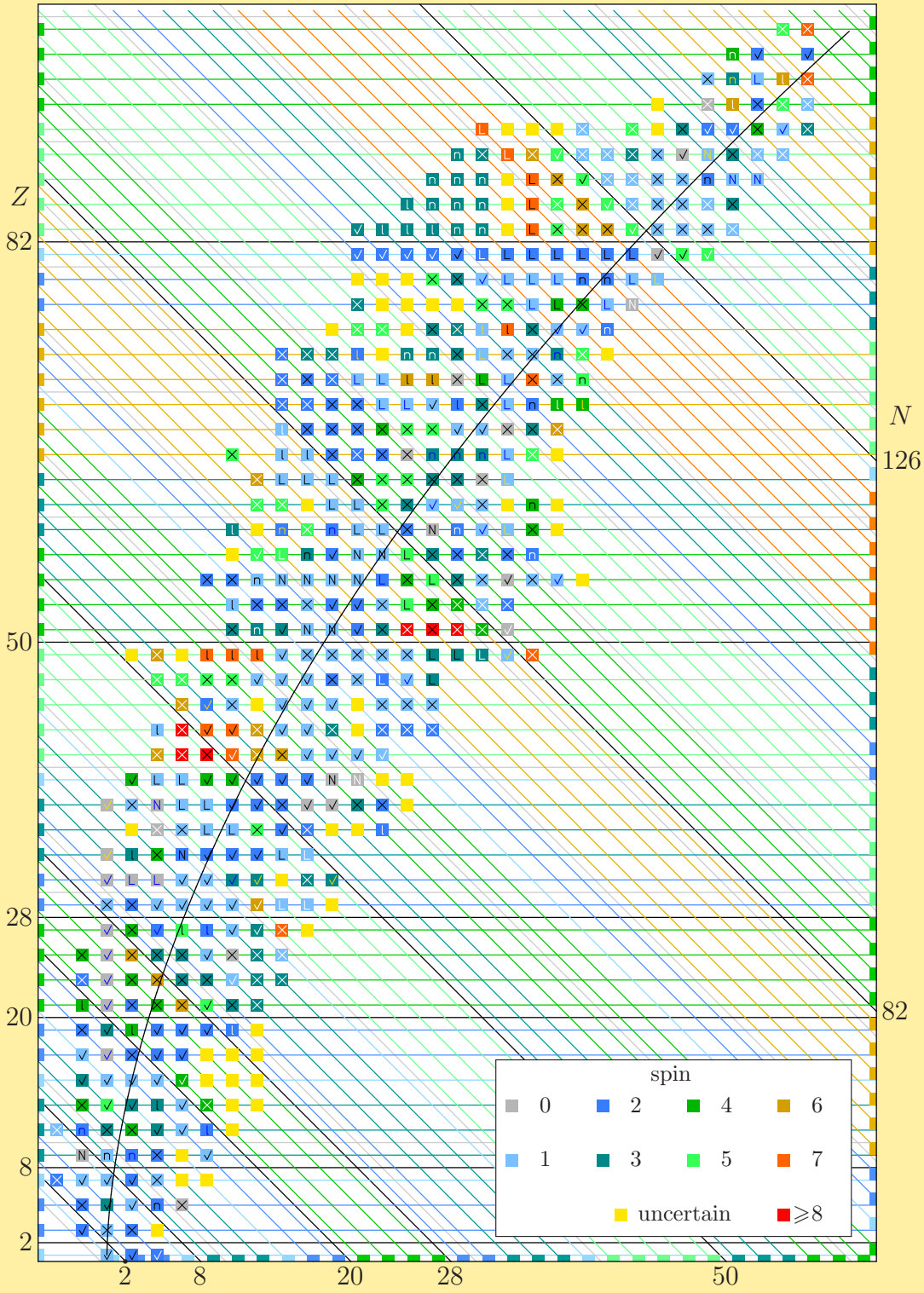


Figure 14.34: Spin of odd-odd nuclei. [pdf][con]

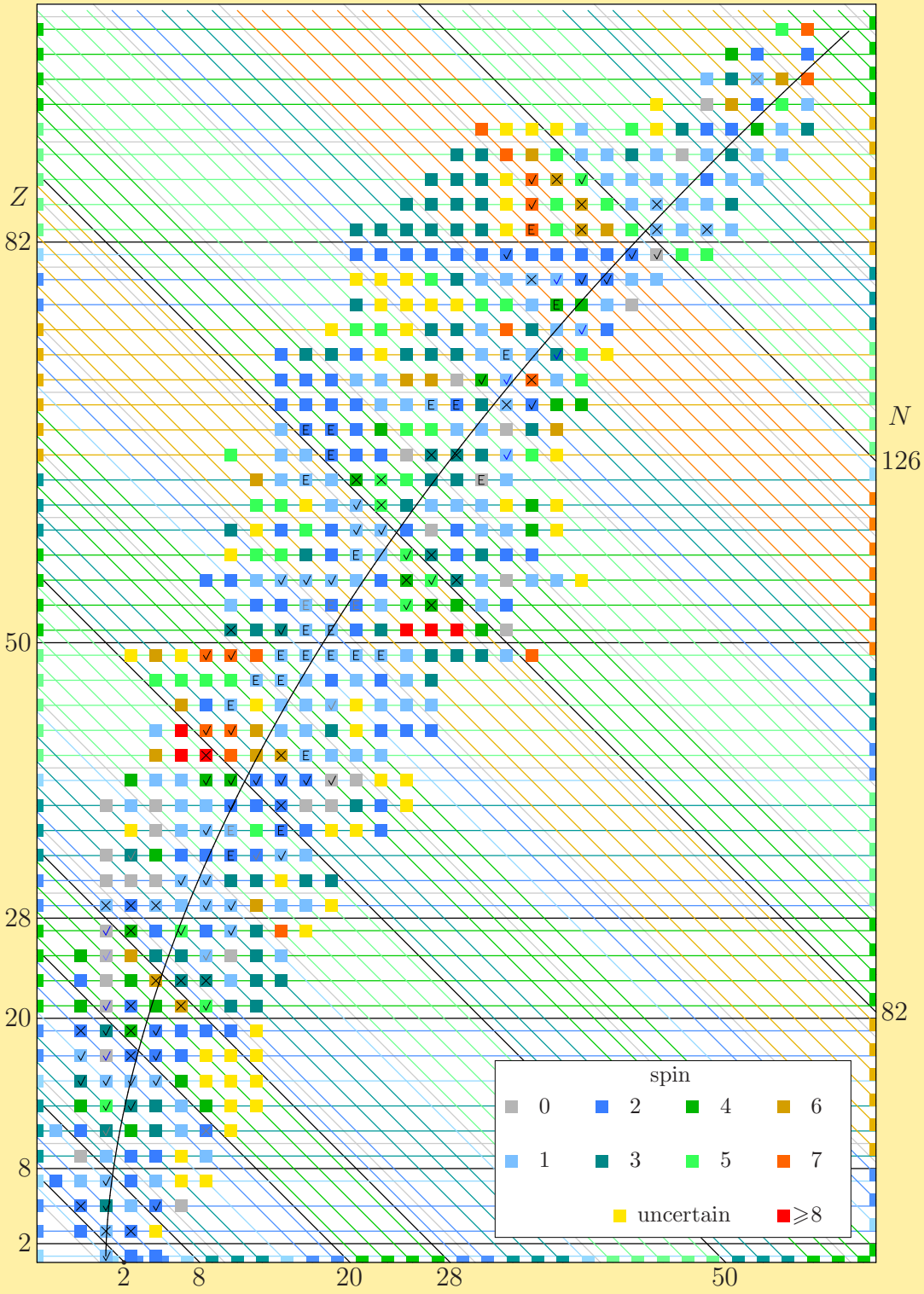


Figure 14.35: Odd-odd spins predicted using the neighbors. [pdf][con]

Figure 14.35 shows results according to that approach. To minimize failures due to other causes than the Nordheim rules, it was demanded that both spin and parity of the odd-odd nucleus were solidly established. For the two pairs of odd mass nuclei, it was demanded that both spin and parity were known, and that the two members of each pair agreed on the values. It was also demanded that the orbital momenta of the pairs could be confidently predicted from the spins and parities. Correct predictions for these superclean cases are indicated by check marks in figure 14.35, incorrect ones by an "E" or cross. Light check marks indicate cases in which the spin of a pair of odd mass nuclei is not the spin of the odd nucleon.

Preston and Bhaduri [36, p. 239] write: "When confronted with experimental data, Nordheim's rules are found to work quite well, most of the exceptions being for light nuclei." So be it. The results are definitely better than chance. Below $Z = 50$, the rules get 43 right out of 71. It may be noted that if you simply take the shells directly from theory with no promotion, like in figure 14.36, you get only 41 right, so using the spins of the neighbors seems to help. The "Nuclear Data Sheets" policies assume that the (unimproved) Nordheim rules may be helpful if there is supporting evidence.

The nuclei marked with "E" in figure 14.35 are particularly interesting. For these nuclei spin or parity show that it is impossible for the odd proton and neutron to be in the same shells as their neighbors. In four cases, the discrepancy is in parity, which is particularly clear. It shows that for an odd proton, having an odd neutron is not necessarily intermediate between having no odd neutron and having one additional neutron besides the odd one. Or vice-versa for an odd neutron. Proton and neutron shells interact nontrivially.

It may be noted that the unmodified Nordheim rules imply that there cannot be any odd-odd nuclei with 0^+ or 1^- ground states. However, some do exist, as is seen in figure 14.34 from the nuclei with spin zero (grey) and blue check marks.

14.16 Draft: Parity Data

The parity of a nucleus is even, or one, if its wave function stays the same if the positive direction of all three Cartesian axes is inverted. That replaces every $\vec{r}$ in the wave function by $-\vec{r}$. The parity is odd, or minus one, if the wave function gets multiplied by -1 under axes inversion. Nuclei have definite parity, (as long as the weak force is not an active factor), so one of the two must be the case. It is an important quantity for what nuclear decays and reactions occur and at what rate.

This section provides an overview of the ground-state spins of nuclei. It will be seen that the shell model does a pretty good job of predicting them.

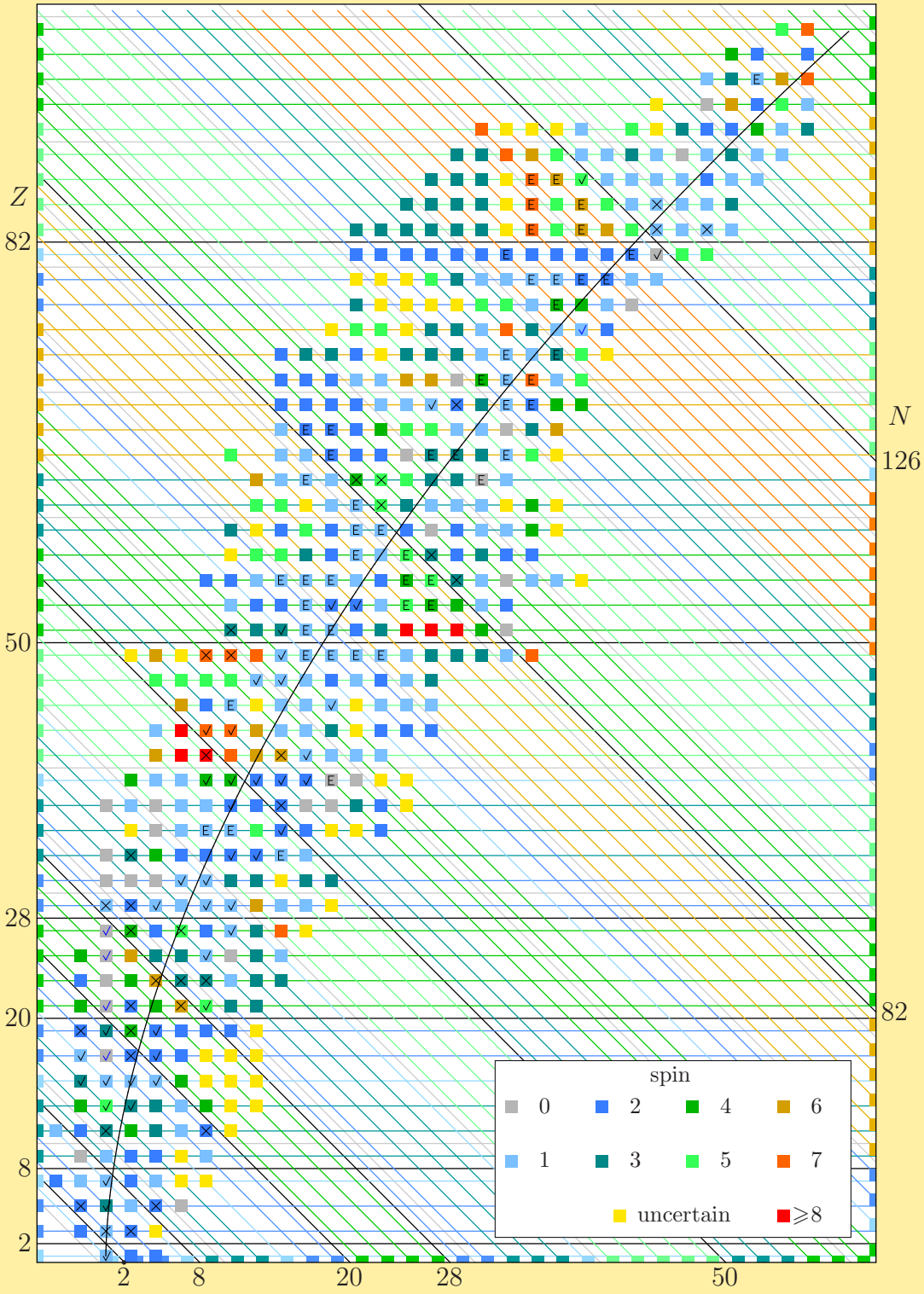


Figure 14.36: Odd-odd spins predicted from theory. [pdf][con]

14.16.1 Draft: Even-even nuclei

For nuclei with both an even number of protons and an even number of neutrons, the odd-particle shell model predicts that the parity is even. This prediction is fully vindicated by the experimental data, figure 14.37. There are no known exceptions to this rule.

14.16.2 Draft: Odd mass number nuclei

For nuclei with an odd mass number A, there is an odd proton or neutron. The odd-particle shell model says that the parity is that of the odd nucleon. To find it, the subshell that the last particle is in must be identified, section 14.12.2. This can be done with a fair amount of confidence based on the spin of the nuclei. Nuclei for which the parity is correctly predicted in this way are shown in green in figures 14.38 and 14.39. Failures are in red. Small grey signs are shell model values if the nucleons fill the shells in the normal order.

The failures above $Z = 50$ and inside the $Z < 82$, $N > 82$ wedge are expected. The shell model does not apply in these regions, because the nuclei are known to be nonspherical there. Besides that, there are very few failures. Those near the $N = 40$ and $N = 60$ lines away from the stable line are presumably also due to nonspherical nuclei. The highly unstable nitrogen-11 and beryllium-11 mirror nuclei were discussed in section 14.12.6.

14.16.3 Draft: Odd-odd nuclei

For odd-odd nuclei, the odd-particle shell model predicts that the parity is the product of those of the surrounding even-odd and odd-even nuclei. The results are shown in figure 14.40. Hits are green, failures red, and unable-to-tell black. Small grey signs are shell model values for the surrounding even-odd and odd-even nuclei. However actual even-odd and odd-even values were used in the prediction.

Failures for spherical nuclei indicate that sometimes the odd proton or neutron is in a different shell than in the corresponding odd-mass neighbors. A similar conclusion can be reached based on the spin data.

Note that the predictions also do a fairly good job in the regions in which the nuclei are not spherical. The reason is that the predictions make no assumptions about what sort of state, spherical or nonspherical, the odd nucleons are in. It merely assumes that they are in the same state as their neighbors.

14.16.4 Draft: Parity Summary

Figure 14.41 shows a summary of the parity of all nuclei together. To identify the type of nucleus more easily, the even-even nuclei have been shown as green

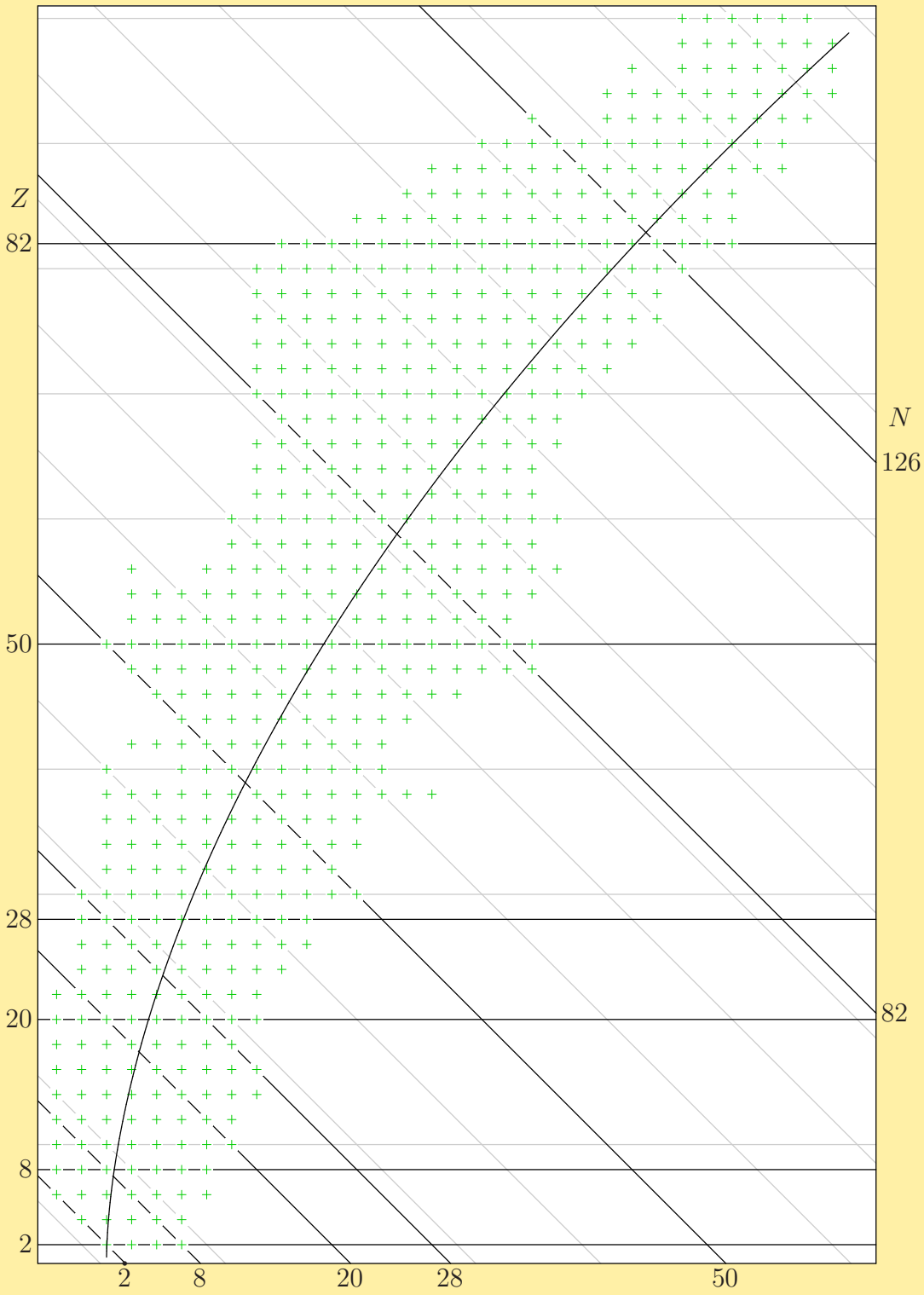


Figure 14.37: Parity of even-even nuclei. [pdf][con]

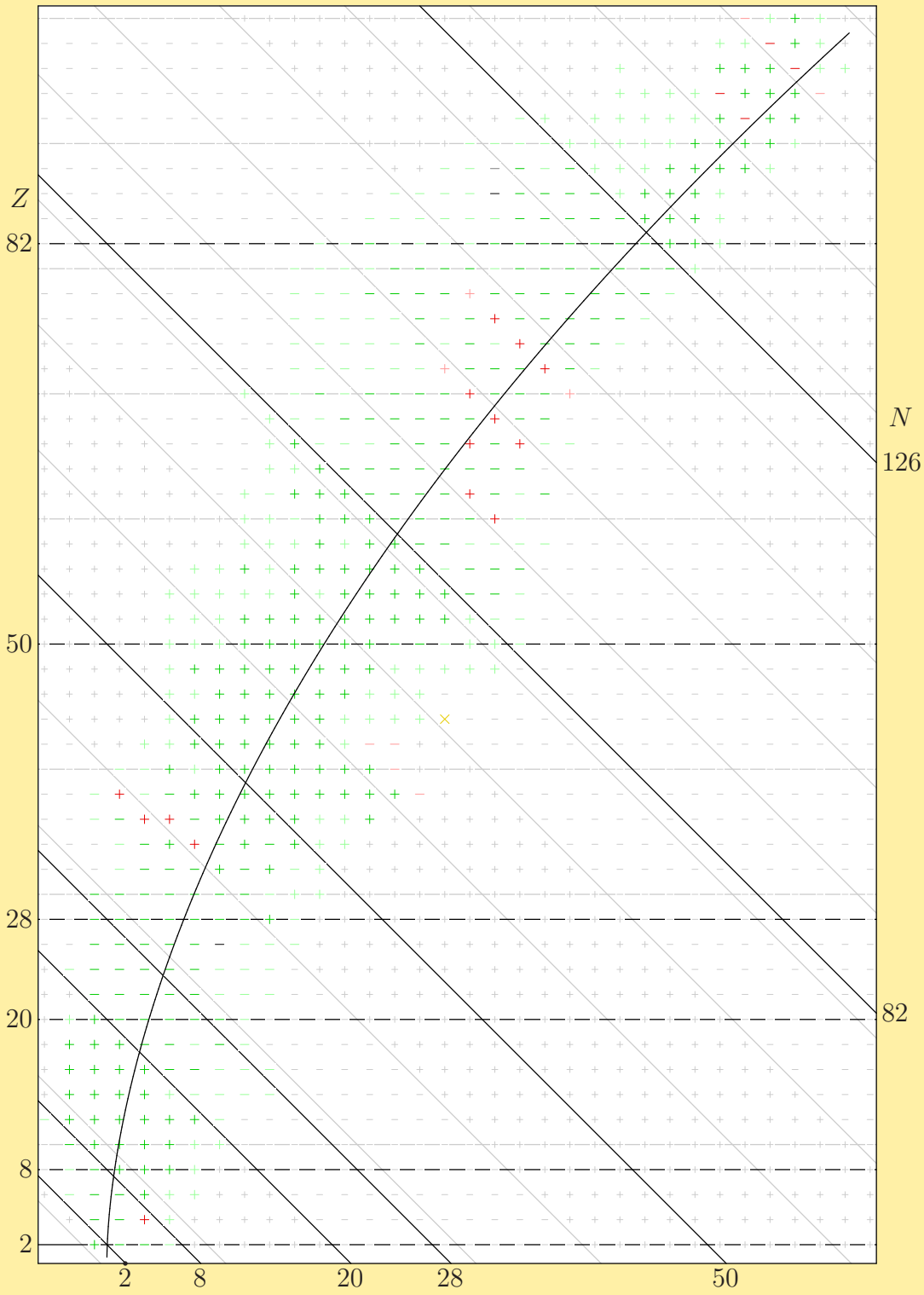


Figure 14.38: Parity of even-odd nuclei. [pdf][con]

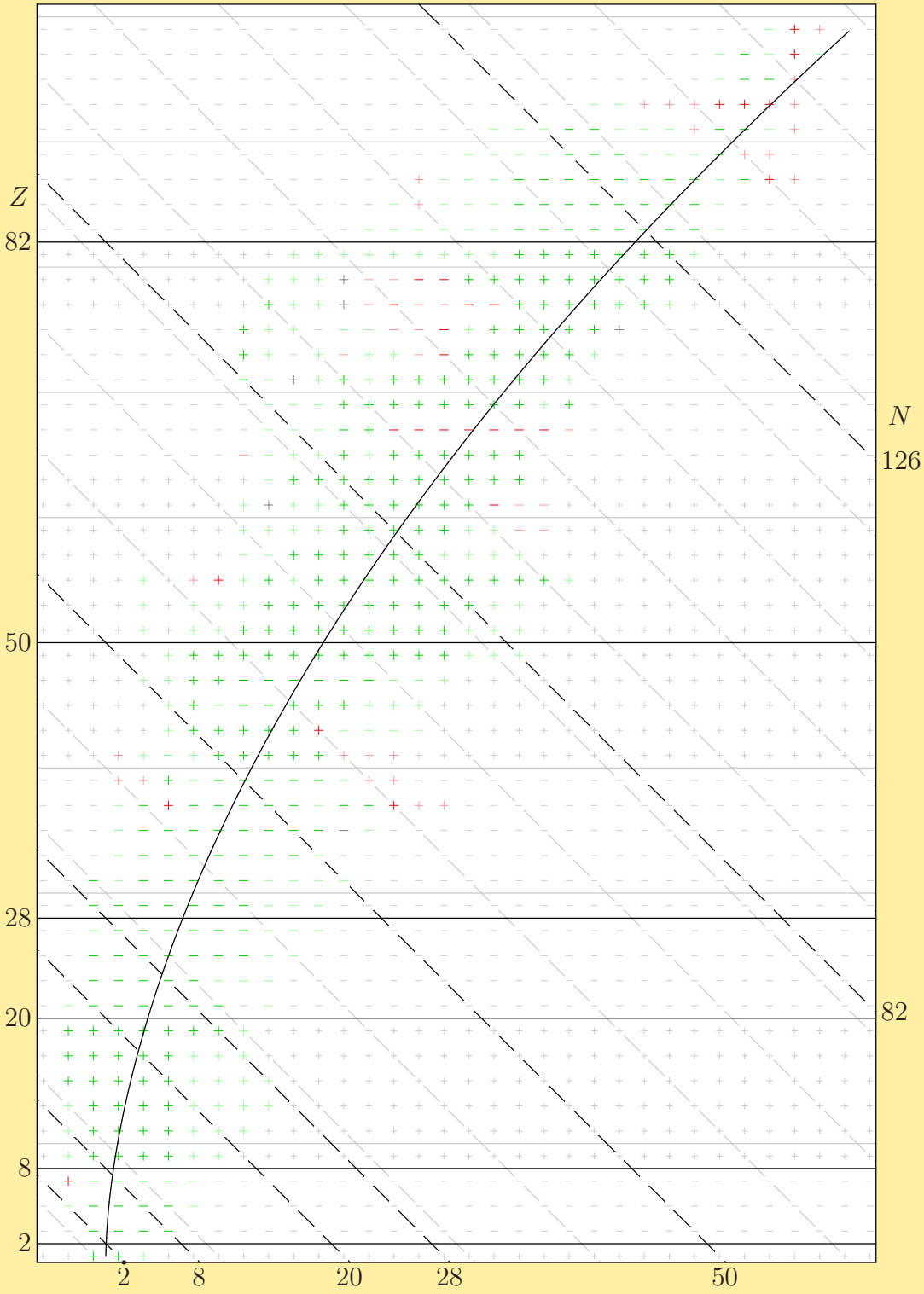


Figure 14.39: Parity of odd-even nuclei. [pdf][con]

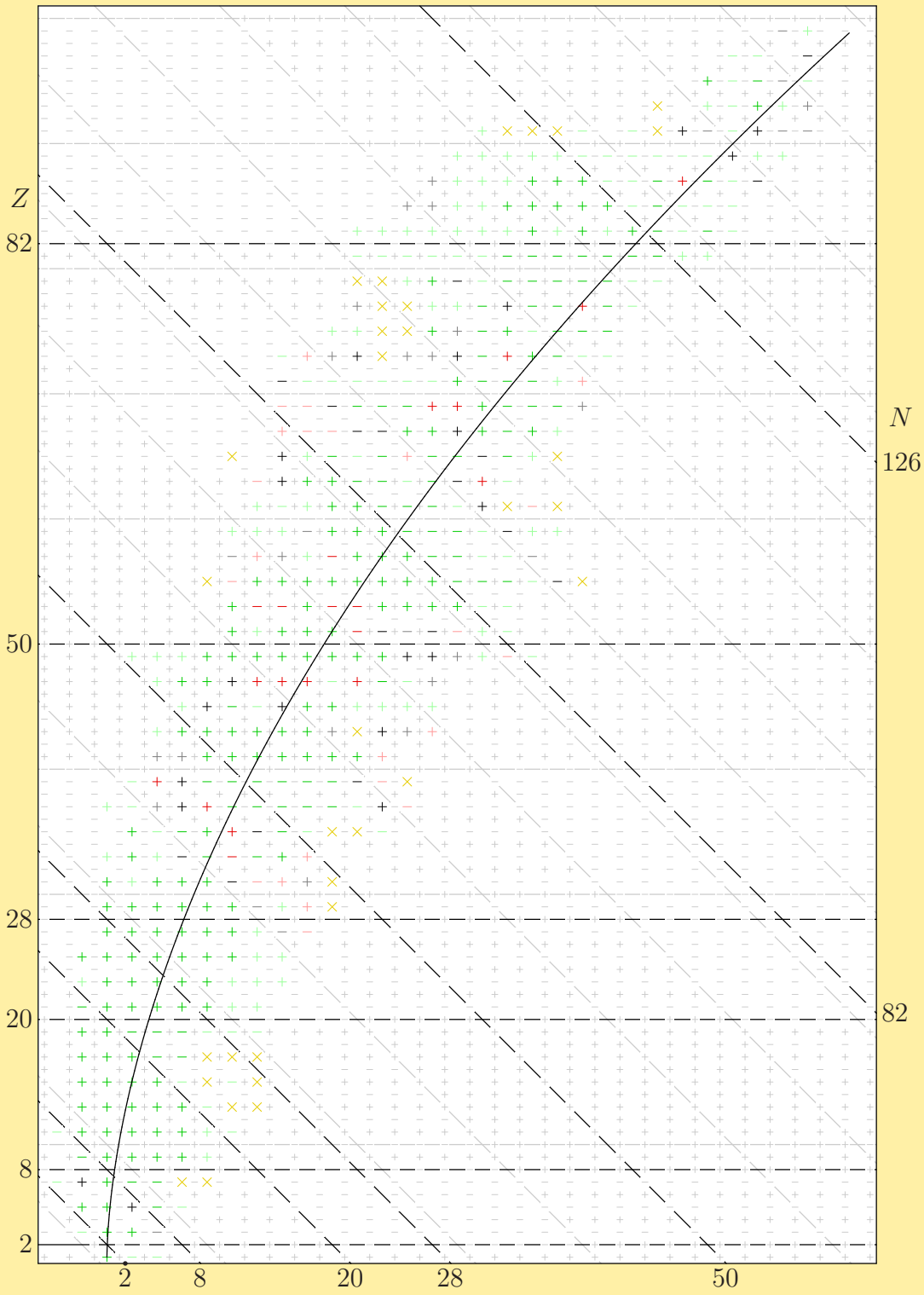


Figure 14.40: Parity of odd-odd nuclei. [pdf][con]

check marks. The odd-odd nuclei are found on the same vertical lines as the check marks. The even-odd nuclei are on the same horizontal lines as the check marks, and the odd-even ones on the same diagonal lines.

Parities that the shell model predicts correctly are in green, and those that it predicts incorrectly are in red. The parities were taken straight from section 14.12.2 with no tricks. Note that the shell model does get a large number of parities right straight off the bat. And much of the errors can be explained by promotion or nonspherical nuclei.

For parities in light green and light red, NUBASE 2003 expressed some reservation about the correct value. For parities shown as yellow crosses, no (unique) value was given.

14.17 Draft: Electromagnetic Moments

The most important electromagnetic property of nuclei is their net charge. It is what keeps the electrons in atoms and molecules together. However, nuclei are not really electric point charges. They have a small size. In a spatially varying electric field most respond somewhat different than a point charge. It is said that they have an electric quadrupole moment. Also, most nuclei act like little electromagnets. It is said that they have a "magnetic dipole moment." These properties are important for applications like NMR and MRI, and for experimentally examining nuclei.

14.17.1 Draft: Classical description

This subsection explains the magnetic dipole and electric quadrupole moments from a classical point of view.

14.17.1.1 Draft: Magnetic dipole moment

The most basic description of an electromagnet is charges going around in circles. It can be seen from either classical or quantum electromagnetics that the strength of an electromagnet is proportional to the angular momentum $\vec{L}$ of the charges times the ratio of their charge q to their mass m, chapter 13.2 or 13.4.

This leads to the definition of the magnetic dipole moment as

$$\vec{\mu} \equiv \frac{q}{2m}\vec{L}$$

In particular, a magnet wants to align itself with an external magnetic field $\vec{\mathcal{B}}_{\text{ext}}$. The energy involved in this alignment is

$$-\vec{\mu} \cdot \vec{\mathcal{B}}_{\text{ext}}$$

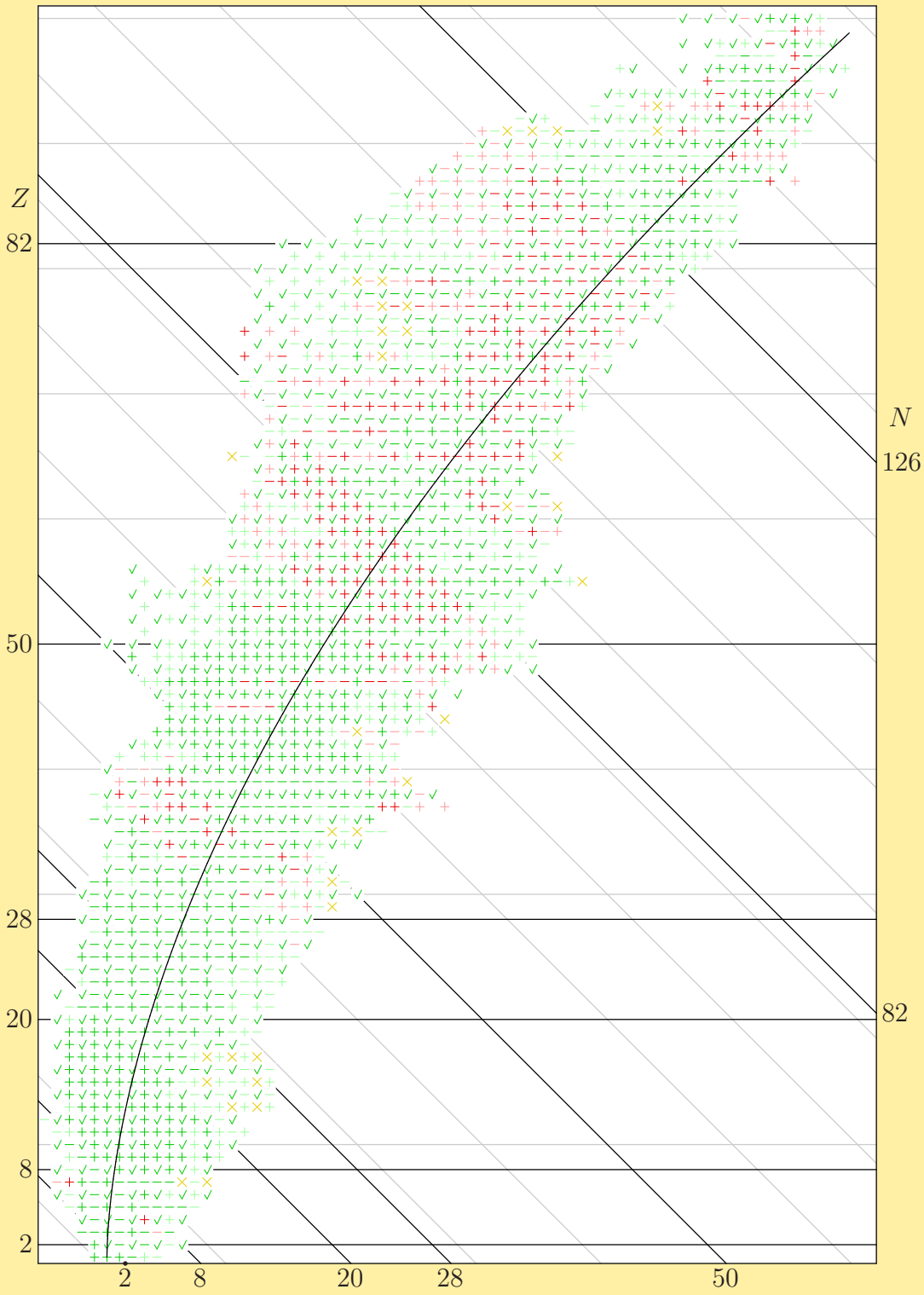


Figure 14.41: Parity versus the shell model. [pdf][con]

14.17.1.2 Draft: Electric quadrupole moment

Consider a nuclear charge distribution with charge density ρ_c placed in an external electrical potential, or "voltage" φ. The potential energy due to the external field is

$$V = \int \varphi \rho_c \, \mathrm{d}^3\vec{r}$$

It may be noted that since nuclear energies are of the order of MeV, an external field is not going to change the nuclear charge distribution ρ. It would need to have a million volt drop over a couple of femtometers to make a dent in it. Unless you shoot very high energy charged particles at the nucleus, that is not going to happen. Also, the current discussion assumes that the external field is steady or at least quasi-steady. That should be reasonable in many cases, as nuclear internal time scales are very fast.

Since nuclei are so small compared to normal external fields, the electric potential φ can be well represented by a Taylor series. That gives the potential energy as

$$V = \varphi_0 \int \rho_c \, \mathrm{d}^3\vec{r} + \sum_{i=1}^{3} \left(\frac{\partial \varphi}{\partial r_i} \right)_0 \int r_i \rho \, \mathrm{d}^3\vec{r} + \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial r_i \partial r_j} \right)_0 \int r_i r_j \rho \, \mathrm{d}^3\vec{r}$$

where $(r_1, r_2, r_3) = (x, y, z)$ are the three components of position and 0 indicates that the derivative is evaluated at the origin, the center of the nucleus.

The first integral in the expression above is just the net nuclear charge q. This makes the first term exactly the same as the potential energy of a point charge. The second integral defines the "electric dipole moment" in the i-direction. It is nonzero if on average the charge is shifted somewhat towards one side of the nucleus. But nuclei do not have nonzero electric dipole moments. The reason is that nuclei have definite parity; the wave function is either the same or the same save for a minus sign when you look at the opposite side of the nucleus. Since the probability of a proton to be found at a given position is proportional to the square magnitude of the wave function, it is just as likely to be found at one side as the other one. (That should really be put more precisely for the picky. The dipole contribution of any set of positions of the protons is canceled by an opposite contribution from the set of opposite nucleon positions.)

The last integral in the expression for the potential energy defines the quadrupole matrix or tensor. You may note a mathematical similarity with the moment of inertia matrix of a solid body in classical mechanics. Just like there, the quadrupole matrix can be simplified by rotating the coordinate system to principal axes. That rotation gets rid of the integrals $\int r_i r_j \rho \mathrm{d}^3\vec{r}$ for $i \neq j$, so what is left is

$$V = V_{\mathrm{pc}} + \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial x^2} \right)_0 \int x^2 \rho \, \mathrm{d}^3\vec{r} + \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial y^2} \right)_0 \int y^2 \rho \, \mathrm{d}^3\vec{r} + \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial z^2} \right)_0 \int z^2 \rho \, \mathrm{d}^3\vec{r}$$

where the first term is the potential of the point charge.

Note that the average of x^2, y^2, and z^2 is $\frac{1}{3}r^2$. It is convenient to subtract that average in each integral. The subtraction does not change the value of the potential energy. The reason is that the sum of the three second order derivatives of the external field φ is zero due to Maxwell"s first equation, chapter 13.2. All that then leads to a definition of an electric quadrupole moment for a single axis, taken to be the z-axis, as

$$Q \equiv \frac{1}{e} \int (3z^2 - r^2) \rho \, \mathrm{d}^3\vec{r}$$

For simplicity, the nasty fractions have been excluded from the definition of Q. Also, it has been scaled with the charge e of a single proton.

That gives Q units of square length, which is easy to put in context. Recall that nuclear sizes are of the order of a few femtometer. So the SI unit square femtometer, fm^2 or 10^{-30} m^2, works quite nicely for the quadrupole moment Q as defined. It is therefore needless to say that most sources do not use it. They use the "barn," a non-SI unit equal to 10^{-28} m^2. The reason is historical; during the second world war some physicists figured that the word "barn" would hide the fact that work was being done on nuclear bombs from the Germans. Of course, that did not work since so few memos and reports are one-word ones. However, physicists discovered that it did help confuse students, so the term has become very widely used in the half century since then. Also, unlike a square femtometer, the barn is much too large compared to a typical nuclear cross section, producing all these sophisticated looking tiny decimal fractions.

To better understand the likely values of the quadrupole moment, consider the effect of the charge distribution of a single proton. If the charge distribution is spherically symmetric, the averages of x^2, y^2 and z^2 are equal, making Q zero. However, consider the possibility that the charge distribution is not spherical, but an ellipsoid of revolution, a "spheroid.". If the axis of symmetry is the z-axis, and the charge distribution hugs closely to that axis, the spheroid will look like a cigar or zeppelin. Such a spheroid is called"prolate." The value of Q is then about $\frac{2}{5}$ of the square nuclear radius R. If the charge distribution stays close to the xy-plane, the spheroid will look like a flattened sphere. Such a spheroid is called "oblate." In that case the value of Q is about $-\frac{2}{5}$ of the square nuclear radius. Either way, the values of Q are noticeably less than the square nuclear radius.

It may be noted that the quadrupole integrals also pop up in the description of the electric field of the nucleus itself. Far from the nucleus, the deviations in its electric field from that of a point charge are proportional to the same integrals, compare chapter 13.3.3.

14.17.2 Draft: Quantum description

Quantum mechanics makes for some changes to the classical description of the electromagnetic moments. Angular momentum is quantized, and spin must be included.

14.17.2.1 Draft: Magnetic dipole moment

As the classical description showed, the strength of an electromagnet is essentially the angular momentum of the charges going around, times the ratio of their charge to their mass. In quantum mechanics angular momentum comes in units of $\hbar$. Also, for nuclei the charged particles are protons with charge e and mass $m_{\rm p}$. Therefore, a good unit to describe magnetic strengths in terms of is the so-called "nuclear magneton"

$$\boxed{\mu_{\rm N} \equiv \frac{e\hbar}{2m_{\rm p}}} \tag{14.26}$$

In those terms, the magnetic magnetic dipole moment operator of a single proton is

$$\frac{1}{\hbar}\widehat{\vec{L}}_{\rm p}\mu_{\rm N}$$

But quantum mechanics brings in a complication, chapter 13.4. Protons have intrinsic angular momentum, called spin. That also acts as an electromagnet. In addition the magnetic strength per unit of angular momentum is different for spin than for orbital angular momentum. The factor that it is different is called the proton g-factor g_p. That then makes the total magnetic dipole moment operator of a single proton equal to

$$\widehat{\vec{\mu}}_{\rm p} = \frac{1}{\hbar}\left(\widehat{\vec{L}} + g_{\rm p}\widehat{\vec{S}}\right)\mu_{\rm N} \qquad g_{\rm p} \approx 5.59 \tag{14.27}$$

The above value of the proton g-factor is experimental.

Neutrons do not have charge and therefore their orbital motion creates no magnetic moment. However, neutrons do create a magnetic moment through their spin:

$$\widehat{\vec{\mu}}_{\rm n} = \frac{1}{\hbar}g_{\rm n}\widehat{\vec{S}}\mu_{\rm N} \qquad g_{\rm n} \approx -3.83 \tag{14.28}$$

The reason is that the neutron consists of three charged quarks; they produce a net magnetic moment even if they do not produce a net charge.

The net magnetic dipole moment operator of the complete nucleus is

$$\widehat{\vec{\mu}} = \frac{1}{\hbar}\left[\sum_{i=1}^{Z}\left(\widehat{\vec{L}}_i + g_{\rm p}\widehat{\vec{S}}_i\right) + \sum_{i=Z+1}^{A} g_{\rm n}\widehat{\vec{S}}_i\right]\mu_{\rm N} \tag{14.29}$$

where i is the nucleon number, the first Z being protons and the rest neutrons.

Now assume that the nucleus is placed in an external magnetic field $\mathcal{B}$ and take the z-axis in the direction of the field. Because nuclear energies are so large, external electromagnetic fields are far too weak to change the quantum structure of the nucleus; its wave function remains unchanged to a very good approximation. However, the field does produce a tiny change in the energy levels of the quantum states. These may be found using expectation values:

$$\Delta E = \langle \Psi | -\widehat{\mu}_z \mathcal{B} | \Psi \rangle$$

The fact that that is possible is a consequence of small perturbation theory, as covered in addendum {A.38}.

However, it is not immediately clear what nuclear wave function Ψ to use in the expectation value above. Because of the large values of nuclear energies, a nucleus is affected very little by its surroundings. It behaves essentially as if it is isolated in empty space. That means that while the nuclear energy may depend on the magnitude of the nuclear spin $\widehat{\vec{J}}$, (i.e. the net nuclear angular momentum), it does not depend on its direction. In quantum terms, the energy does not depend on the component $\widehat{J}_z$ in the chosen z-direction. So, what should be used in the above expectation value to find the change in the energy of a nucleus in a state of spin j? States with definite values of J_z? Linear combinations of such states? You get a difference answer depending on what you choose.

Now a nucleus is a composite structure, consisting of protons or neutrons, each contributing to the net magnetic moment. However, the protons and neutrons themselves are composite structures too, each consisting of three quarks. Yet at normal energy levels protons and neutrons act as elementary particles, whose magnetic dipole moment is a scalar multiple $g\mu_{\rm N}$ of their spin. Their energies in a magnetic field split into two values, one for the state with $J_z = \frac{1}{2}\hbar$ and the other with $J_z = -\frac{1}{2}\hbar$. One state corresponds to magnetic quantum number $m_j = \frac{1}{2}$, the other to $m_j = -\frac{1}{2}$.

The same turns out to be true for nuclei; they too behave as elementary particles as long as their wave functions stay intact. In a magnetic field, the original energy level of a nucleus with spin j splits into equally spaced levels corresponding to nuclear magnetic quantum numbers $m_j = j, j-1, \ldots, -j$. The numerical value of the magnetic dipole moment μ is therefore *defined* to be the expectation value of $\widehat{\mu}_z$ in the nuclear state in which m_j has its largest value j, call it the $|jj\rangle$ state:

$$\boxed{\mu \equiv \langle jj | \widehat{\mu}_z | jj \rangle} \tag{14.30}$$

The fact that nuclei would behave so simple is related to the fact that nuclei are essentially in empty space. That implies that the complete wave function of a nucleus in the ground state, or another energy eigenstate, will vary in a very

simple way with angular direction. Furthermore, that variation is directly given by the angular momentum of the nucleus. A brief discussion can be found in chapter 7.3 and its note. See also the discussion of the Zeeman effect, and in particular the weak Zeeman effect, in addendum {A.38}.

The most important consequence of those ideas is that

Nuclei with spin zero do not have magnetic dipole moments.

That is not very easy to see from the general expression for the magnetic moment, cluttered as it is with g-factors. However, zero spin means on a very fundamental level that the complete wave function of a nucleus is independent of direction, chapter 4.2.3. A magnetic dipole strength requires directionality, there must be a north pole and a south pole. That cannot occur for nuclei of spin zero.

14.17.2.2 Draft: Electric quadrupole moment

The definition of the electric quadrupole moment follows the same ideas as that of the magnetic dipole moment. The numerical value of the quadrupole moment is *defined* as the expectation value of $3z^2 - r^2$, summed over all protons, in the state in which the net nuclear magnetic quantum number m_j equals the nuclear spin j:

$$\boxed{Q \equiv \langle jj| \sum_{i=1}^{Z} 3z_i^2 - r_i^2 |jj\rangle} \tag{14.31}$$

Note that there is a close relation with the spherical harmonics;

$$3z^2 - r^2 = \sqrt{\frac{16\pi}{5}} r^2 Y_2^0 \tag{14.32}$$

That is important because it implies that

Nuclei with spin zero or with spin one-half do not have electric quadrupole moments.

To see why, note that the expectation value involves the absolute square of the wave function. Now if you multiply two wave functions together that have an angular dependence corresponding to a spin j, mathematically speaking you get pretty much the angular dependence of two particles of spin j. That cannot become more than an angular dependence of spin $2j$, in other words an angular dependence with terms proportional to Y_{2j}^m. Since the spherical harmonics are mutually orthonormal, Y_{2j}^m integrates away against Y_2^0 for $j \leqslant \frac{1}{2}$.

It makes nuclei with spin ½ popular for nuclear magnetic resonance studies. Without the perturbing effects due to quadrupole interaction with the electric field, they give nice sharp signals. Also of course, analysis is easier with only two spin states and no quadrupole moment.

14.17.2.3 Draft: Shell model values

According to the odd-particle shell model, all even-even nuclei have spin zero and therefore no magnetic or electric moments. That is perfectly correct.

For nuclei with an odd mass number, the model says that all nucleons except for the last odd one are paired up in spherically symmetric states of zero spin that produce no magnetic moment. Therefore, the magnetic moment comes from the last proton or neutron. To get it, according to the second last subsubsection, what is needed is the expectation value of the magnetic moment operator $\widehat{\mu}_z$ as given there. Assume the shell that the odd nucleon is in has single-particle net momentum j. According to the definition of magnetic moment, the magnetic quantum number must have its maximum value $m_j = j$. Call the corresponding state the ψ_{nljj} one because the spectroscopic notation is useless as always. In particular for an odd-even nucleus,

$$\mu = \frac{1}{\hbar}\langle\psi_{nljj}|L_z + g_p\widehat{S}_z|\psi_{nljj}\rangle\mu_{\rm N}$$

while for an even-odd nucleus

$$\mu = \frac{1}{\hbar}\langle\psi_{nljj}|g_n\widehat{S}_z|\psi_{nljj}\rangle\mu_{\rm N}$$

The unit $\mu_{\rm N}$ is the nuclear magneton. The expectation values can be evaluated by writing the state ψ_{nljj} in terms of the component states ψ_{nlmm_s} of definite angular momentum $\widehat{L}_z$ and spin $\widehat{S}_z$ following chapter 12.8, 2.

It is then found that for an odd proton, the magnetic moment is

$$\boxed{\begin{array}{ll} j = l - \tfrac{1}{2}: & \mu_{\rm p1} = \tfrac{1}{2}\dfrac{j}{j+1}\left(2j + 3 - g_p\right)\mu_{\rm N} \\ j = l + \tfrac{1}{2}: & \mu_{\rm p2} = \tfrac{1}{2}\left(2j - 1 + g_p\right)\mu_{\rm N} \end{array}} \tag{14.33}$$

while for an odd neutron

$$\boxed{\begin{array}{ll} j = l - \tfrac{1}{2}: & \mu_{\rm n1} = -\tfrac{1}{2}\dfrac{j}{j+1}g_n\mu_{\rm N} \\ j = l + \tfrac{1}{2}: & \mu_{\rm n2} = \tfrac{1}{2}g_n\mu_{\rm N} \end{array}} \tag{14.34}$$

These are called the "Schmidt values."

Odd-odd nuclei are too messy to be covered here, even if the Nordheim rules would be reliable.

For the quadrupole moments of nuclei of odd mass number, filled shells do not produce a quadrupole moment, because they are spherically symmetric. Consider now first the case that there is a single proton in an otherwise empty shell with single-particle momentum j. Then the magnetic moment of the nucleus can be found as the one of that proton:

$$Q = Q_{\rm p} = \langle\psi_{nljj}|3z^2 - r^2|\psi_{nljj}\rangle$$

Evaluation, {D.78}, gives

$$\boxed{Q_{\rm p} = -\frac{2j-1}{2j+2}\langle r^2\rangle} \tag{14.35}$$

where $\langle r^2\rangle$ is the expectation value of r^2 for the proton. Note that this is zero as it should if the spin $j = \frac{1}{2}$. Since the spin j must be half-integer, zero spin is not a consideration. For all other values of j, the one-proton quadrupole moment is negative.

The expectation value $\langle r^2\rangle$ can hardly be much more than the square nuclear radius, excepting maybe halo nuclei. A reasonable guess would be to assume that the proton is homogeneously distributed within the nuclear radius R. That gives a ballpark value

$$\langle r^2\rangle \approx \tfrac{3}{5}R^2$$

Next consider the case that there are not one but $I \geqslant 3$ protons in the unfilled shell. The picture of the odd-particle shell model as usually painted is: the first $I-1$ protons are pairwise combined in spherically symmetric states and the last odd proton is in a single particle state, blissfully unaware of the other protons in the shell. In that case, the quadrupole moment would self evidently be the same as for one proton in the shell. But as already pointed out in section 14.12.4, the painted picture is not really correct. For one, it does not satisfy the antisymmetrization requirement for all combinations on protons. There really are I protons in the shell sharing one wave function that produces a net spin equal to j.

In particular consider the case that there are $2j$ protons in the shell. Then the wave function takes the form of a filled shell, having no quadrupole moment, plus a "hole", a state of angular momentum j for the missing proton. Since a proton hole has minus the charge of a proton, the quadrupole moment for a single hole is opposite to that of one proton:

$$\boxed{Q_{2j\rm p} = -Q_{\rm p}} \tag{14.36}$$

In other words, the quadrupole moment for a single hole is predicted to be positive. For $j = \frac{1}{2}$, a single proton also means a single hole, so the quadrupole moment must, once more, be zero. It has been found that the quadrupole moment changes linearly with the odd number of protons, [31, p, 129]. Therefore for shells with more than one proton and more than one hole, the quadrupole moment is in between the one-proton and one-hole values. It follows that the one-proton value provides an upper bound to the magnitude of the quadrupole moment for any number of protons in the shell.

Since neutrons have no charge, even-odd nuclei would in the simplest approximation have no quadrupole moment at all. However, consider the odd neutron and the spherical remainder of the nucleus as a two-body system going

around their common center of gravity. In that picture, the charged remainder of the nucleus will create a quadrupole moment. The position vector of the remainder of the nucleus is about $1/A$ times shorter than that of the odd neutron, so quadratic lengths are a factor $1/A^2$ shorter. However, the nucleus has Z times as much charge as a single proton. Therefore you expect nuclei with an odd neutron to have about Z/A^2 times the quadrupole moment of the corresponding nucleus with an odd proton instead of an odd neutron. For heavy nuclei, that would still be very much smaller than the magnetic moment of a similar odd-even nucleus.

14.17.2.4 Draft: Values for deformed nuclei

For deformed nuclei, part of the angular momentum is due to rotation of the nucleus as a whole. In particular, for the ground state rotational band of deformed even-even nuclei, all angular momentum is in rotation of the nucleus as a whole. This is orbital angular momentum. Protons with orbital angular momentum produce a magnetic dipole moment equal to their angular momentum, provided the dipole moment is expressed in terms of the nuclear magneton μ_N. Uncharged neutrons do not produce a dipole moment from orbital angular momentum. Therefore. the magnetic dipole moment of the nucleus is about

$$\boxed{\text{even-even, ground state band:} \qquad \mu = g_R j \mu_N \quad g_R \approx \frac{Z}{A}} \tag{14.37}$$

where the g-factor reflects the relative amount of the nuclear angular momentum that belongs to the protons. This also works for vibrational nuclei, since their angular momentum too is in global motion of the nucleus.

If a rotational band has a minimum spin $j_{\min}$ that is not zero, the dipole moment is, [40, p. 392],

$$\boxed{\mu = \left[g_R j + \frac{j_{\min}^2}{j+1}(g_{\text{int}} - g_R) \right] \mu_N \quad g_R \approx \frac{Z}{A} \qquad j_{\min} \neq 1/2} \tag{14.38}$$

where $g_{\text{int}} j_{\min} \mu_N$ reflects an internal magnetic dipole strength. If $j_{\min} = \frac{1}{2}$, the top of the first ratio has an additional term that has a magnitude proportional to $2j + 1$ and alternates in sign.

The quadrupole moment of deformed nuclei is typically many times larger than that of a shell model one. According to the shell model, all protons except at most one are in spherical orbits producing no quadrupole moment. But if the nucleus is deformed, typically into about the shape of some spheroid instead of a sphere, then all protons contribute. Such a nucleus has a very large "intrinsic" quadrupole moment Q_{int}.

However, that intrinsic quadrupole moment is not the one measured. For example, many heavy even-even nuclei have very distorted *intrinsic* shapes but

all even-even nuclei have a *measured* quadrupole moment that is zero in their ground state. That is a pure quantum effect. Consider the state in which the axis of the nucleus is aligned with the z-direction. In that state a big quadrupole moment would be observed due to the directional charge distribution. But there are also states in which the nucleus is aligned with the x-direction, the y-direction, and any other direction for that matter. No big deal classically: you just grab hold of the nucleus and measure its quadrupole moment. But quantum mechanics makes the complete wave function a linear combination of all these different possible orientations; in fact an equal combination of them by symmetry. If all directions are equal, there is no directionality left; the measured quadrupole moment is zero. Also, directionality means angular momentum in quantum mechanics; if all directions are equal the spin is zero. "Grabbing hold" of the nucleus means adding directionality, adding angular momentum. That creates an excited state.

A simple known system that shows such effects is the hydrogen atom. Classically the atom is just an electron and a proton at opposite sides of their center of gravity. If they are both on the z-axis, say, that system would have a nonzero quadrupole moment. But such a state is not an exact energy eigenstate, far from it. It interacts with states in which the direction of the connecting line is different. By symmetry, the ground state is the one in which all directions have the same probability. The atom has become spherically symmetric. Still, the atom has not become *intrinsically* spherically symmetric; the wave function is not of a form like $\psi_1(r_e)\psi_2(r_p)$. The positions of electron and proton are still correlated, {A.5}.

A model of a spheroidal nucleus produces the following relationship between the intrinsic quadrupole moment and the one that is measured:

$$\boxed{Q = \frac{3j_{\min}^2 - j(j+1)}{(j+1)(2j+3)} Q_{\text{int}}} \tag{14.39}$$

where $j_{\min}$ is the angular momentum of the nucleus when it is not rotating. Derivations may be found in [40] or [36]. It can be seen that when the nucleus is not rotating, the measured quadrupole moment is much smaller than the intrinsic one unless the angular momentum is really large. When the nucleus gets additional rotational angular momentum, the measured quadrupole moment decreases even more and eventually ends up with the opposite sign.

14.17.3 Draft: Magnetic moment data

Figure 14.42 shows ground state magnetic moments in units of the nuclear magneton μ_N. Even-even nuclei do not have magnetic moments in their ground state, so they are not shown. The red and blue horizontal lines are the Schmidt values predicted by the shell model. They differ in whether spin subtracts from

or adds to the net angular momentum j to produce the orbital momentum l. Red dots should be on the red lines, blue dots on the blue lines. For black dots, no confident prediction of the orbital angular momentum could be made. The values have an error of no more than about 0.1 μ_N, based on a subjective evaluation of both reported errors as well as differences between results obtained by different studies for the same number. These differences are often much larger than the reported errors for the individual numbers.

One good thing to say about it all is that the general magnitude is well predicted. Few nuclei end up outside the Schmidt lines. (Rhodium-103, a stable odd-even $\frac{1}{2}^-$ nucleus, is a notable exception.) Also, some nuclei are actually on their line. And the others tend to at least be on the right side of the cloud. The bad news is, of course, that the agreement is only qualitatively.

The main excuses that are offered are:

1. The g-factors g_p and g_n describe the effectiveness of proton and neutron spins in generating magnetic moments in free space. They may be reduced when these nucleons are inside a nucleus. Indeed, it seems reasonable enough to assume that the motion of the quarks that make up the protons and neutrons could be affected if there are other quarks nearby. Reduction of the g-factors drives the Schmidt lines towards each other, and that can clearly reduce the average errors. Unfortunately, different nuclei would need different reductions to obtain quantitative agreement.
2. Collective motion. If some of the angular momentum is into collective motion, it tends to drift the magnetic moment towards about $\frac{1}{2}j\mu_N$, compare (14.38). To compute the effect requires the internal magnetic moment of the nucleus to be known. For some nuclei, fairly good magnetic moments can be obtained by using the Schmidt values for the internal magnetic moment, [40, p. 393].

For odd-odd nuclei, the data average out to about $0.5j$ nuclear magnetons, with a standard deviation of about one magneton. These average values are shown as yellow lines in figure 14.42. Interestingly enough, the average is like a collective rotation, (14.37).

According to the shell model, two odd particles contribute to the spin and magnetic moment of odd-odd nuclei. So they could have significantly larger spins and magnetic moments than odd mass nuclei. Note from the data in figure 14.42 that that just does not happen.

Even-even nuclei do not have magnetic moments in their ground state. Figure 14.43 shows the magnetic moments of the first exited 2^+ state of these nuclei. The values are in fairly good agreement with the prediction (14.37) of collective motion that the magnetic moment equals Zj/A nuclear magnetons. Bright green squares are correct. Big deviations occur only near magic numbers. The maximum error in the shown data is about a quarter of a nuclear magneton, subjectively evaluated.

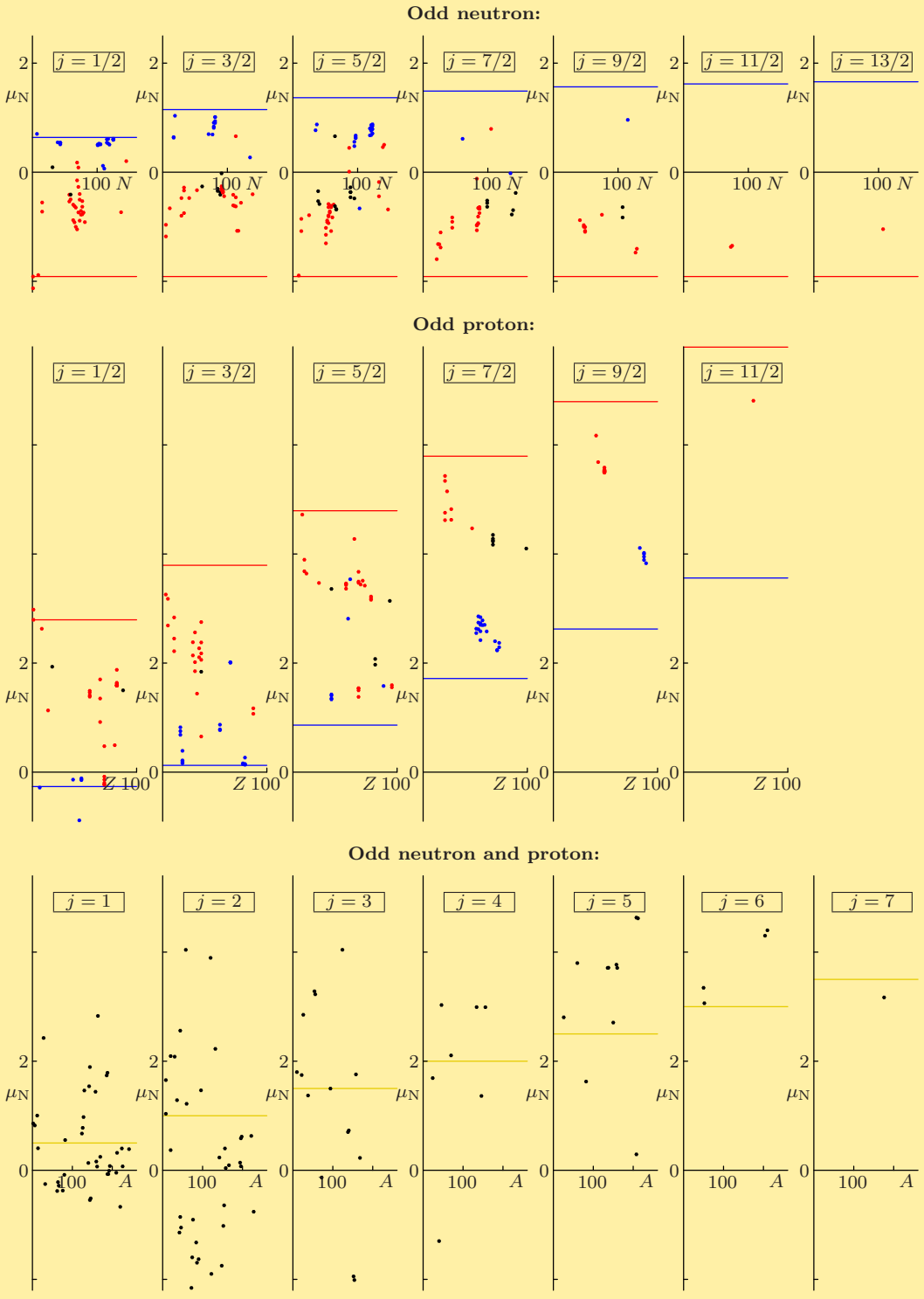


Figure 14.42: Magnetic dipole moments of the ground-state nuclei. [pdf]

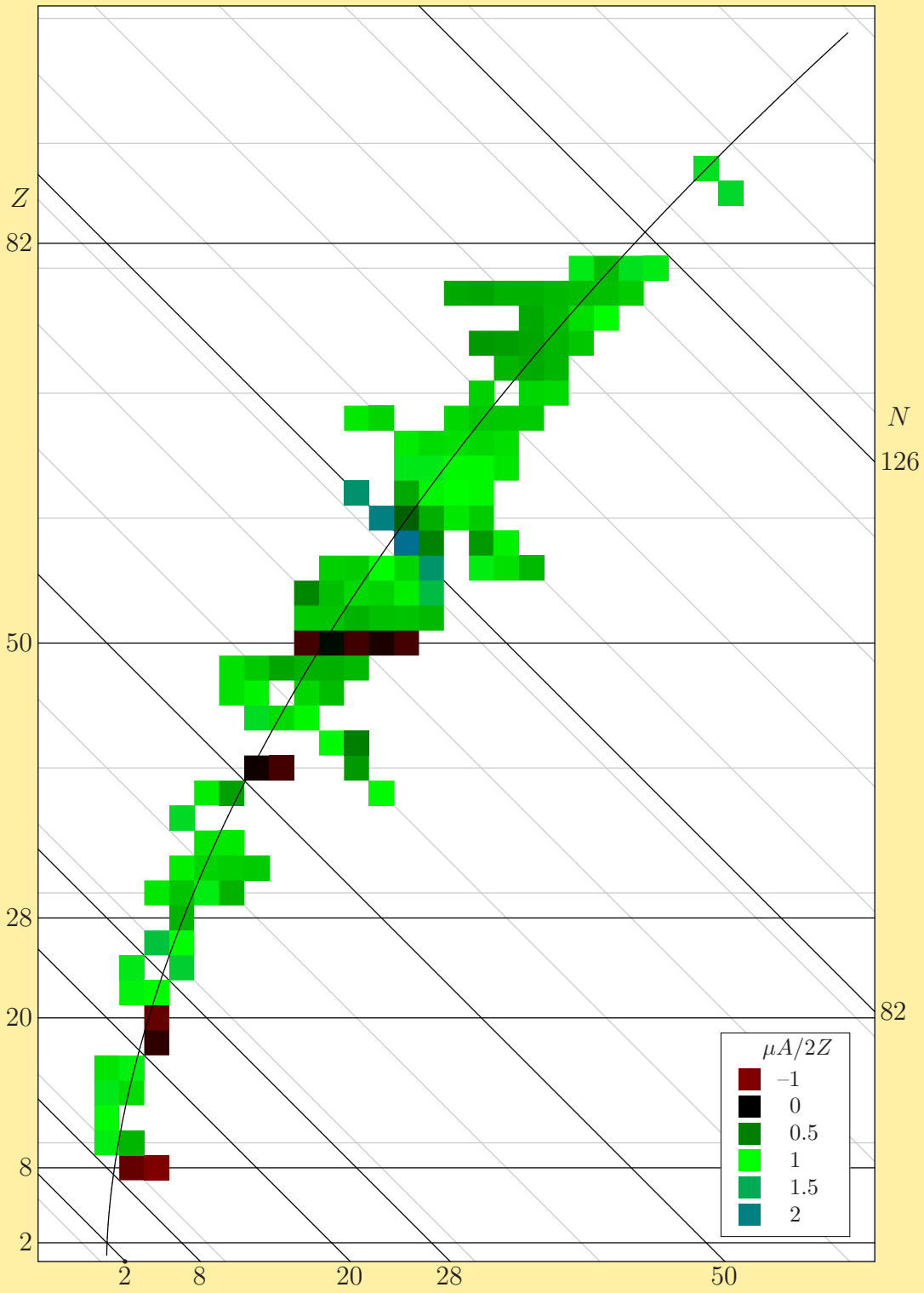


Figure 14.43: 2^+ magnetic moment of even-even nuclei. [pdf][con]

14.17.4 Draft: Quadrupole moment data

If you love the shell model, you may want to skip this subsection. It is going to get a beating.

The prediction of the shell model is relatively straightforward. The electric quadrupole moment of a single proton in an unfilled shell of high angular momentum can quite well be ballparked as

$$Q_{\text{p ballpark}} \sim \tfrac{3}{5} R^2$$

where R is the nuclear radius computed from (14.9). This value corresponds to the area of the square marked "a proton's" in the legend of figure 14.44. As discussed in subsection 14.17.2.3, if there are more protons in the shell, the magnitude is less, though the sign will eventually reverse. If the angular momentum is not very high, the magnitude is less. If there is no odd proton, the magnitude will be almost zero. So, essentially all squares in figure 14.44 must be smaller, most a lot smaller, and those on lines of even Z very much smaller, than the single proton square in the legend...

Well, you might be able to find a smaller square somewhere. For example, the square for lithium-6, straight above doubly-magic ${}^{4}_{2}\text{He}$, has about the right size and the right color, blue. The data shown have a subjectively estimated error of up to 40%, [sic], and the area of the squares gives the scaled quadrupole moment. Nitrogen-14, straight below doubly-magic ${}^{16}_{8}\text{O}$, has a suitably small square of the right color, red. So does potassium-39 with one proton less than doubly-magic ${}^{40}_{20}\text{Ca}$. Bismuth-209, with one more proton than ${}^{208}_{82}\text{Pb}$ has a relatively small square of the right color. Some nuclei on magic proton number lines have quite small scaled quadrupole moments, though hardly almost zero as they should. Nuclei one proton above magic proton numbers tend to be of the right color, blue, as long as their squares are small. Nuclei one proton below the magic proton numbers should be red; however, promotion can mess that up.

Back to reality. Note that many nuclei in the $Z < 82$, $N > 82$ wedge, and above $Z = 82$, as well as various other nuclei, especially away from the stable line, have quadrupole moments that are very many times larger than the ballpark for a single proton. That is simply not possible unless many or all protons contribute to the quadrupole moment. The odd-particle shell model picture of a spherically symmetric nuclear core plus an odd proton, and maybe a neutron, in nonspherical orbits hanging on is completely wrong for these nuclei. These nuclei have a global shape that simply is not spherical. And because the shell model was derived based on a spherical potential, its results are invalid for these nuclei. They are the deformed nuclei that also showed up in figures 14.19 and 14.22. It is the quadrupole moment that shows that it was not just an empty excuse to exclude these nuclei in shell model comparisons. The measured quadrupole moments show without a shadow of a doubt that the shell model cannot be valid.

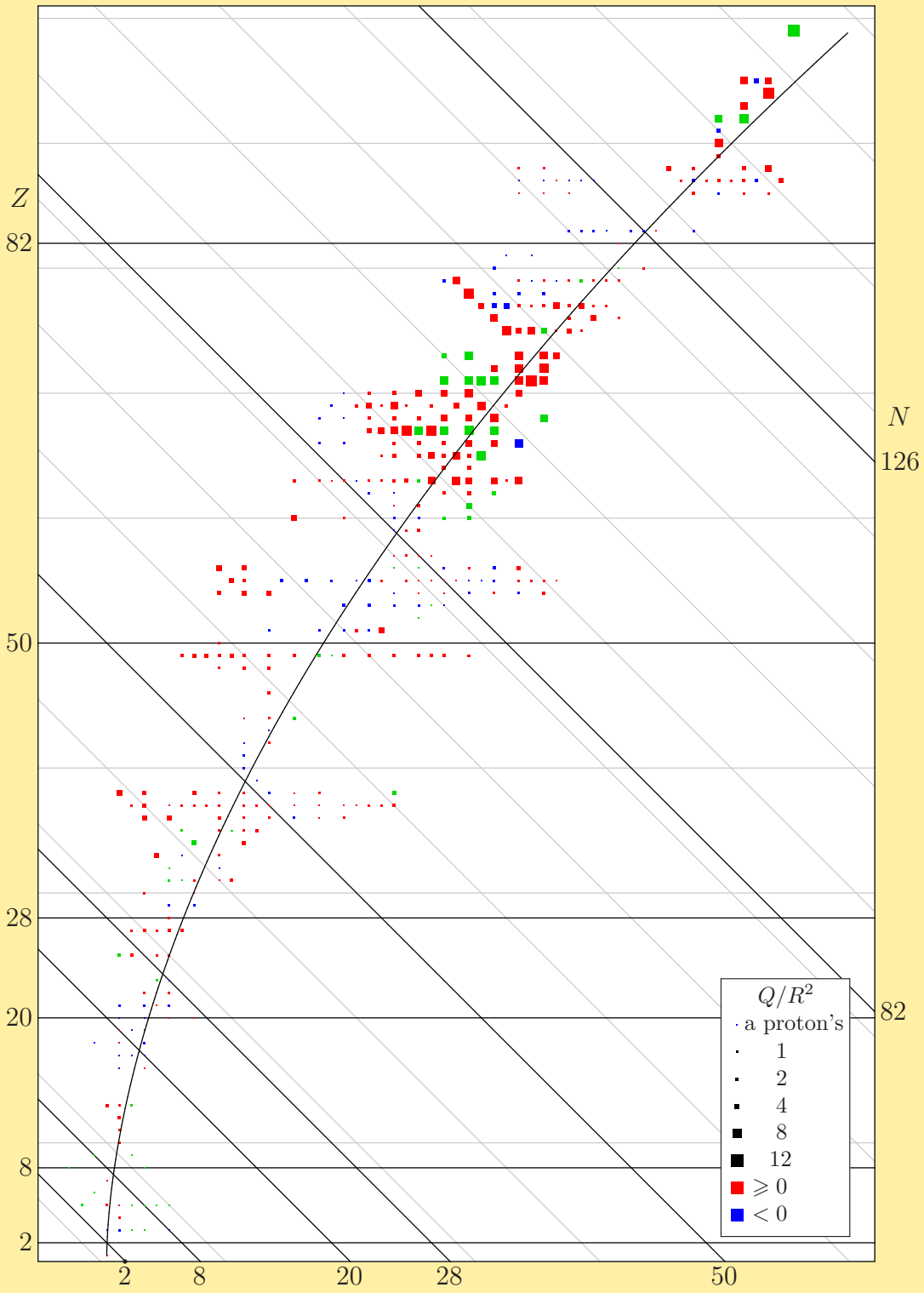


Figure 14.44: Electric quadrupole moment. [pdf][con]

You might however wonder about the apparently large amount in random scatter in the quadrupole moments of these nuclei. Does the amount of deformation vary that randomly? Before that can be answered, a correction to the data must be applied. Measured quadrupole moments of a deformed nucleus are often much too small for the actual nuclear deformation. The reason is uncertainty in the angular orientation of these nuclei. In particular, nuclei with spin zero have complete uncertainty in orientation. Such nuclei have zero measured quadrupole moment regardless how big the deformation of the nucleus is. Nuclei with spin one-half still have enough uncertainty in orientation to measure as zero.

Figure 14.45 shows what happens if you try to estimate the "intrinsic" quadrupole moment of the nuclei in absence of uncertainty in angular orientation. For nuclei whose spin is at least one, the estimate was made based on the measured value using (14.39), with both $j_{\min}$ and j equal to the spin. This assumes that the intrinsic shape is roughly spheroidal. For shell-model nuclei, this also roughly corrects for the spin effect, though it overcorrects to some extent for nuclei of low spin.

To estimate the intrinsic quadrupole moment of nuclei with zero ground state spin, including all even-even nuclei, the quadrupole moment of the lowest excited 2^+ state was used, if it had been measured. For spin one-half the lowest $^3/_2$ state was used. In either case, $j_{\min}$ was taken to be the spin of the ground state and j that of the excited state. Regrettably, these estimates do not make much sense if the nucleus is not a rotating one.

Note in figure 14.45 how much more uniform the squares in the regions of deformed nuclei have become. And that the squares of nuclei of spin zero and one-half have similar sizes. These nuclei were not really more spherical; it was just hidden from experiments.

The observed intrinsic quadrupole moments in the regions of deformed nuclei correspond to roughly 20% radial deviation from the spherical value. Clearly, that means quite a large change in shape.

It may be noted that figure 14.44 leaves out magnesium-23, whose reported quadrupole moment of 1.25 barn is far larger that that of similar nuclei. If this value is correct, clearly magnesium-23 must be a halo nucleus with two protons outside a neon-21 core.

14.18 Draft: Isospin

Isospin is another way of thinking about the two types of nucleons. It has proved quite useful in understanding nuclei, as well as elementary particles.

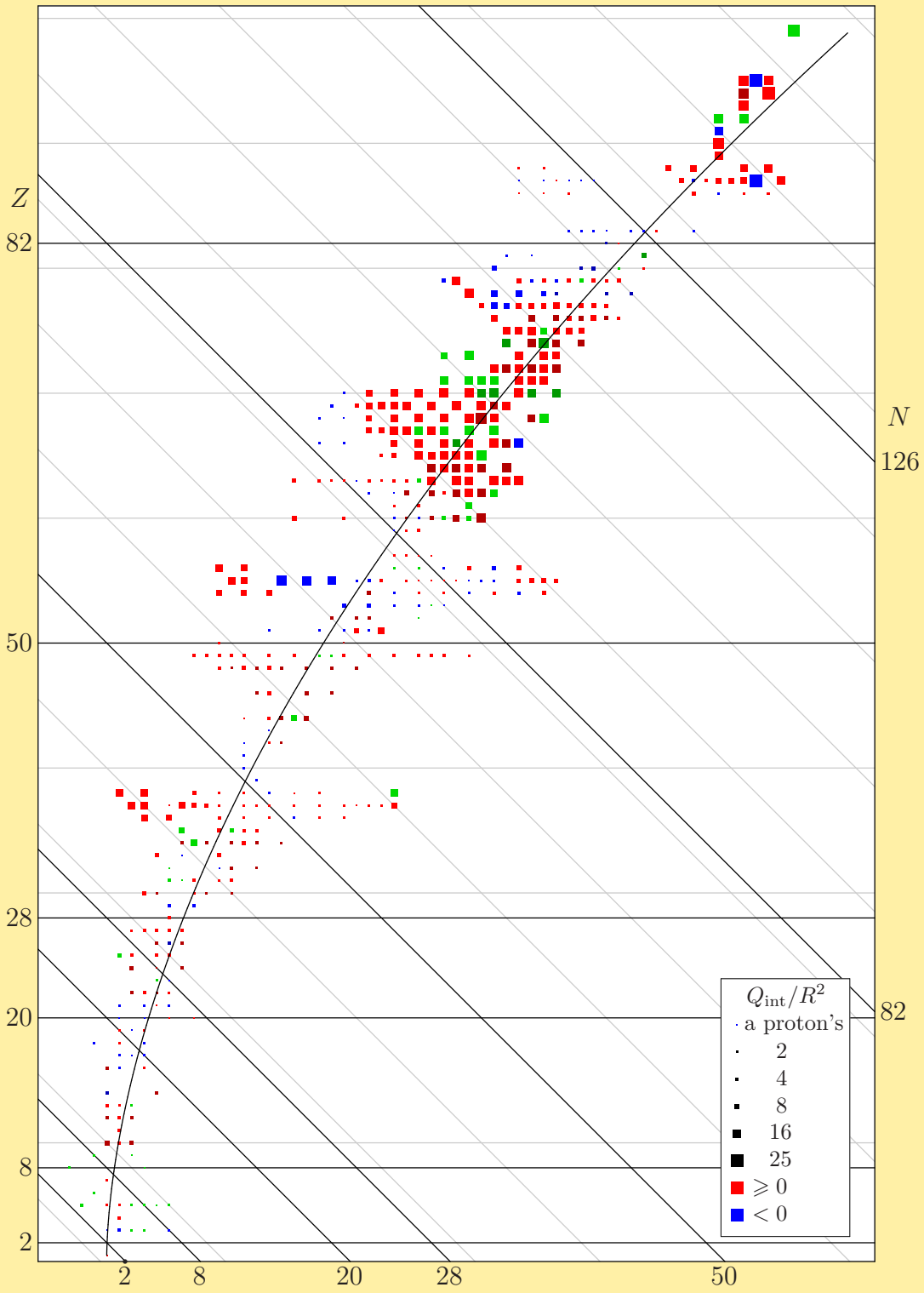


Figure 14.45: Electric quadrupole moment corrected for spin. [pdf][con]

14.18.1 Draft: Basic ideas

Normally, you think of nuclei as consisting of protons and neutrons. But protons and neutrons are very similar in properties, if you ignore the Coulomb force. They have almost the same mass. Also, according to charge independence, the nuclear force is almost the same whether it is protons or neutrons.

So suppose you define only one particle, called nucleon. Then you can give that particle an additional property called "nucleon type." If the nucleon type is $\frac{1}{2}$, it is a proton, and if the nucleon type is $-\frac{1}{2}$ it is a neutron. That makes nucleon type a property that is mathematically much like the spin S_z of a nucleon in a chosen z-direction. But of course, there is no physical "nucleon-type axis." Therefore nucleon type is conventionally indicated by the symbol T_3, not T_z, with no physical meaning attached to the 3-axis. In short, nucleon type is defined as:

$$\boxed{\text{proton: } T_3 = \tfrac{1}{2} \qquad \text{neutron: } T_3 = -\tfrac{1}{2}} \tag{14.40}$$

So far, all this it may seem like a stupid mathematical trick. And normally it would be. The purpose of mathematical analysis is to understand systems, not to make them even more incomprehensible.

But to the approximation that the nuclear force is charge-independent, nucleon type is not so stupid after all. If the nuclear force is charge-independent, and the Coulomb force is ignored, you can write down nuclear wave functions without looking at the nucleon type. Now suppose that in doing so, you find some energy eigenfunction of the form

$$\psi_{\rm A} = \psi_{\rm s}(\vec{r}_2 - \vec{r}_1)|1\ 1\rangle$$

This is a wave function for two nucleons labeled 1 and 2. Assume here that the spatial part $\psi_{\rm s}$ is unchanged under nucleon exchange (swapping the nucleons):

$$\psi_{\rm s}(\vec{r}_2 - \vec{r}_1) = \psi_{\rm s}(\vec{r}_1 - \vec{r}_2)$$

(This is equivalent to assuming that the wave function has even parity.) Further recall from chapter 5.5.6 that the triplet spin state $|1\ 1\rangle$ is unchanged under nucleon exchange too. So the total wave function $\psi_{\rm A}$ above is unchanged under nucleon exchange.

That is fine if nucleon 1 is a proton and nucleon 2 a neutron. Or vice-versa. But it is not OK if both nucleons are protons, or if they are both neutrons. Wave functions must change sign if two identical fermions are exchanged. That is the antisymmetrization requirement. The wave function above stays the same. So it is only acceptable for the deuteron, with one proton and one neutron.

Next suppose you could find a different wave function of the form

$$\psi_{\rm B} = \psi_{\rm s}(\vec{r}_2 - \vec{r}_1)|0\ 0\rangle$$

(Here ψ_s is not necessarily the same as before, but assume it still has even parity.) The singlet spin state $|0\,0\rangle$ changes sign under nucleon exchange. Then so does the entire wave function $\psi_{\rm B}$. And that then means that $\psi_{\rm B}$ is acceptable even if the nucleons are both protons or both neutrons. This wave function works not just for the deuteron, but also for the "diproton" and the "dineutron." (The prefix "di" means two.)

That would give nontrivial insight in nuclear energy levels. It would mean physically that the diproton, the dineutron, and the deuteron can be in an identical energy state. Such identical energy states, occurring for different nuclei, are called "isobaric analog (or analogue) states." Or "charge states." Or "isobaric multiplets." Or "T-multiplets." Hey, don't blame the messenger.

Disappointingly, in real life there is no bound state of the form $\psi_{\rm B}$. Still the bottom line stays:

> *Within the approximations of charge independence and negligible Coulomb effects, whether a given state applies to a given set of nucleon types depends only on the antisymmetrization requirements.*

Now for bigger systems of nucleons, the antisymmetrization requirements get much more complex. A suitable formalism for dealing with that has already been developed in the context of the spin of systems of identical fermions. It is convenient to adopt that formalism also for nucleon type.

As an example, consider the above three hypothetical isobaric analog states for the diproton, dineutron, and deuteron. They can be written out separately as, respectively,

$$\psi_{\rm s}(\vec{r}_2-\vec{r}_1)|0\,0\rangle\uparrow\!\!\!\uparrow_1\uparrow\!\!\!\uparrow_2 \qquad \psi_{\rm s}(\vec{r}_2-\vec{r}_1)|0\,0\rangle\downarrow\!\!\!\downarrow_1\downarrow\!\!\!\downarrow_2 \qquad \psi_{\rm s}(\vec{r}_2-\vec{r}_1)|0\,0\rangle\frac{\uparrow\!\!\!\uparrow_1\downarrow\!\!\!\downarrow_2+\downarrow\!\!\!\downarrow_1\uparrow\!\!\!\uparrow_2}{\sqrt{2}}$$

Here $\uparrow\!\!\!\uparrow$ means the nucleon is a proton and $\downarrow\!\!\!\downarrow$ means it is a neutron. If you want, you can write out the above wave functions explicitly in terms of the nucleon type T_3 as

$$\psi_{\rm s}(\vec{r}_2-\vec{r}_1)|0\,0\rangle(\tfrac{1}{2}+T_{31})(\tfrac{1}{2}+T_{32})$$

$$\psi_{\rm s}(\vec{r}_2-\vec{r}_1)|0\,0\rangle(\tfrac{1}{2}-T_{31})(\tfrac{1}{2}-T_{32})$$

$$\psi_{\rm s}(\vec{r}_2-\vec{r}_1)|0\,0\rangle\frac{(\frac{1}{2}+T_{31})(\frac{1}{2}-T_{32})+(\frac{1}{2}-T_{31})(\frac{1}{2}+T_{32})}{\sqrt{2}}$$

Note, for example, that the first wave function is zero if either T_{31} or T_{32} is equal to $-\frac{1}{2}$. So it is zero if either nucleon is a neutron. The only way to get something nonzero is if both nucleons are protons, with $T_{31}=T_{32}=\frac{1}{2}$. Similarly, the second wave function is only nonzero if both nucleons are neutrons, with $T_{31}=T_{32}=-\frac{1}{2}$.

The third wave function represents something of a change in thinking. It requires that one nucleon is a proton and the other a neutron. So it is a wave function for the deuteron. But the actual wave function above is a superposition of two states. In the first state, nucleon 1 is the proton and nucleon 2 the neutron. In the second state, nucleon 1 is the neutron and nucleon 2 the proton. In the combined state the nucleons have lost their identity. It is uncertain whether nucleon 1 is the proton and nucleon 2 the neutron, or vice-versa.

Within the formalism of identical nucleons that have an additional nucleon-type property, this uncertainty in nucleon types is unavoidable. The wave function would not be antisymmetric under nucleon exchange without it. But if you think about it, this may actually be an improvement in the description of the physics. Protons and neutrons do swap identities. That happens if they exchange a charged pion. Proton-neutron scattering experiments show that they can do that. For nucleons that have a probability of swapping type, assigning a fixed type in energy eigenstates is not right. Energy eigenstates must be stationary. And having a better description of the physics can affect what sort of potentials you would want to write down for the nucleons.

(You might think that without charge independence, the additional antisymmetrization requirement for identical nucleons would change the physics. But actually, it does not. The antisymmetrization requirement can be accomodated by uncertainty in which nucleon you label 1 and which 2. Consider some completely general proton-neutron wave function $\Psi(\vec{r}_{\rm p}, S_{z{\rm p}}, \vec{r}_{\rm n}, S_{z{\rm n}})$, one that would not be the same if you swap the proton and neutron. It might be a subsystem in some larger nucleus for which charge-independence is not a good approximation. The antisymmetrized identical-nucleon wave function is

$$\frac{1}{\sqrt{2}}\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2})\uparrow_1\downarrow_2 - \frac{1}{\sqrt{2}}\Psi(\vec{r}_2, S_{z2}, \vec{r}_1, S_{z1})\downarrow_1\uparrow_2 \tag{14.41}$$

This is a superposition of two parts. In the first part the proton is labeled 1 and the neutron 2. In the second part, the proton is labeled 2 and the neutron 1. The physics has stayed exactly the same. What has changed is that there is now confusion about whether the particle labeled 1 is the proton or the neutron. This illustrates that without charge independence, the mathematical trickery employed here is not really wrong. But it is "entirely useless," as Wigner was the first to point out, [50, p. 4].)

Now compare the trailing nucleon-type factors in the three hypothetical isobaric analog states above with the possible combined spin states of two spin ½ fermions. It is seen that the nucleon-type factors take the exact same form as the so-called "triplet" spin states, (5.26). So define similarly

$$\uparrow_1\uparrow_2 \equiv |1\ 1\rangle_T \qquad \downarrow_1\downarrow_2 \equiv |1\ {-1}\rangle_T \qquad \frac{\uparrow_1\downarrow_2 + \downarrow_1\uparrow_2}{\sqrt{2}} \equiv |1\ 0\rangle_T \tag{14.42}$$

In those terms, the isobaric analog states are all three of the form

$$\psi_{\rm s}(\vec{r}_2 - \vec{r}_1)|0\,0\rangle|1\,m_T\rangle_T$$

where m_T is 1, -1, or 0 for the diproton, dineutron, and deuteron respectively. Note that that is just the sum of the T_3 values of the two nucleons,

$$m_T = T_{3_1} + T_{3_2}$$

Now reconsider the wave function for two nucleons that was written down first. The one that was only acceptable for the deuteron. In the same terminology, it can be written as

$$\psi_{\rm s}(\vec{r}_2 - \vec{r}_1)|1\,1\rangle|0\,0\rangle_T \qquad |0\,0\rangle_T \equiv \frac{\Uparrow_1\Downarrow_2 - \Downarrow_1\Uparrow_2}{\sqrt{2}}$$

The formalism of identical nucleons with nucleon type forces again uncertainty in the nucleon types. But now there is a minus sign. That makes the nucleon-type state a singlet state in the terminology of spin.

Of course, all this raises the question what to make of the leading 0 in the singlet state $|0\,0\rangle_T$, and the leading 1 in the triplet states $|1\,m_T\rangle$? If this was spin angular momentum, the 0 or 1 would indicate the quantum number s of the square spin angular momentum. Square spin angular momentum is the sum of the square spin components in the x, y, and z directions. But at first that seems to make no sense for nucleon type. Nucleon type is just a simple number, not a vector. While it has been formally associated with some abstract 3-axis, there are no "T_1" and "T_2" components.

However, it is possible to define such components in complete analogy with the x and y components of spin. In quantum mechanics the components of spin are the eigenvalues of operators. And using advanced concepts of angular momentum, chapter 12, the operators of x and y angular momenta can be found without referring explicitly to their axes. The same procedure can be followed for nucleon type.

To do so, first an operator $\widehat{T}_3$ for the nucleon type T_3 is defined as

$$\boxed{\widehat{T}_3 \Uparrow = \tfrac{1}{2} \Uparrow \qquad \widehat{T}_3 \Downarrow = -\tfrac{1}{2} \Downarrow} \tag{14.43}$$

In words, the proton state is an eigenstate of this operator with eigenvalue $\frac{1}{2}$. The neutron state is an eigenstate with eigenvalue $-\frac{1}{2}$. That follows the usual rules of quantum mechanics; observable quantities, (here nucleon type), are the eigenvalues of Hermitian operators.

Next a "charge creation operator" is defined by

$$\boxed{\widehat{T}^+ \Downarrow = \Uparrow \qquad \widehat{T}^+ \Uparrow = 0} \tag{14.44}$$

In words, it turns a neutron into a proton. In effect it adds a unit of charge to it. Since a nucleon with two units of charge does not exist, the operator needs to turn a proton state into zero. Similarly a "charge annihilation operator" is defined by

$$\boxed{\widehat{T}^- \Uparrow = \Downarrow \qquad \widehat{T}^- \Downarrow = 0} \tag{14.45}$$

Operators for nucleon type in the 1 and 2 directions can now be *defined* as

$$\boxed{\widehat{T}_1 = \tfrac{1}{2}\widehat{T}^+ + \tfrac{1}{2}\widehat{T}^- \qquad \widehat{T}_2 = -\tfrac{1}{2}\mathrm{i}\widehat{T}^+ + \tfrac{1}{2}\mathrm{i}\widehat{T}^-} \tag{14.46}$$

The eigenvalues of these operators are by definition the values of T_1 respectively T_2.

With these operators, square nucleon type can be defined just like square spin. All the mathematics has been forced to be the same.

The quantum number of square nucleon type will be indicated by t_T in this book. Different sources use different notations. Many sources swap case, using lower case for the operators and upper case for the quantum numbers. Or they use lower case if it is for a single nucleon and upper case for the entire nucleus. They often do the same for angular momentum. Some sources come up with I for the square nucleon type quantum number, using J for the angular momentum one. However, this book cannot adopt completely inconsistent notations just for nuclear physics. Especially if there is no generally agreed-upon notation in the first place.

In any case there are three scaled operators whose definition and symbols are fairly standard in most sources. These are defined as

$$\boxed{\tau_1 = 2\widehat{T}_1 \quad \tau_2 = 2\widehat{T}_2 \quad \tau_3 = 2\widehat{T}_3 \qquad \tau^+ = 2\widehat{T}^+ \qquad \tau^- = 2\widehat{T}^{-1}} \tag{14.47}$$

The first three are the direct equivalents of the famous Pauli spin matrices, chapter 12.10. Note that they simply scale away the factors $\frac{1}{2}$ in the nucleon type. The Pauli spin matrices also scale away the factor $\hbar$ that appears in the values of spin.

14.18.2 Draft: Heavier nuclei

Now consider an example of isobaric analog states that actually exist. In this case the nucleons involved are carbon-14, nitrogen-14, and oxygen-14. All three have 14 nucleons, so they are isobars. However, carbon-14 has 6 protons and 8 neutrons, while oxygen-14 has 8 protons and 6 neutrons. Such pairs of nuclei, that have their numbers of protons and neutrons swapped, are called "conjugate" nuclei. Or "mirror" nuclei. Nitrogen-14 has 7 protons and 7 neutrons and is called "self-conjugate."

Since T_3 values add up, carbon-14 with 6 protons at $\frac{1}{2}$ each and 8 neutrons at $-\frac{1}{2}$ each has net $T_3 = -1$. Similarly, nitrogen-14 has $T_3 = 0$, as any self-conjugate nucleus, while oxygen-14 has $T_3 = 1$. In general,

$$\boxed{T_3 = \tfrac{1}{2}(Z - N)} \tag{14.48}$$

where Z is the number of protons and N the number of neutrons. Note that the value of T_3 is fixed for a given nucleus. It is minus half the neutron excess of the nucleus.

In general, 14 nucleons can have a maximum $T_3 = 7$, if all 14 are protons. The minimum is -7, if all 14 are neutrons.

Here is where the analogy with spin angular momentum gets interesting. Angular momentum is a vector. A given angular momentum vector can still have different directions. And different directions means different values of the z-component of its spin S_z. In particular, the quantum numbers of the possible z components are

$$m_s \equiv S_z/\hbar = -s,\ -s{+}1,\ \ldots,\ s{-}1,\ s$$

Here s is the quantum number of the square spin of the vector.

Since nucleon type has been defined to be completely equivalent to spin, essentially the same holds. A given nucleon energy state can still have different values of T_3:

$$\boxed{m_T \equiv T_3 = -t_T,\ -t_T{+}1,\ \ldots,\ t_T{-}1,\ t_T} \tag{14.49}$$

Here t_T is the square nucleon-type quantum number of the state. The different values for T_3 above correspond to isobaric analog states for different nuclei. They are the same energy state, but for different nuclei.

You could say that isobaric analog states arise because "rotating" an energy state in the abstract 1,2,3-space defined above does not make a difference. And the reason it does not make a difference is charge independence.

Based on the values of T_3 above, consider the possible values of t_T for 14 nucleons. The value of t_T cannot be greater than 7. Otherwise there would be isobaric analog states with T_3 greater than 7, and that is not possible for 14 nucleons. As far as the lowest possible value of t_T is concerned, it varies with nucleus. As the expression above shows, the value of $|T_3|$ cannot be greater than t_T. So t_T cannot be less than $|T_3|$. Since carbon-14 and oxygen-14 have $|T_3| = 1$, for these nuclei, t_T cannot be less than 1. However, nitrogen-14, with $T_3 = 0$, also allows states with $t_T = 0$.

It turns out that light nuclei in their ground state generally have the smallest value of t_T consistent with their value of T_3. One way to get some idea of why that would be so is to look at the antisymmetrization requirements. A set of states with $t_T = 7$ for 14 nucleons allows the state $T_3 = 7$, in which all 14 nucleons are protons. In that state, the wave function must be antisymmetric

when any nucleon is interchanged with any other. On the other hand, a $t_T = 0$ state only needs to satisfy the antisymmetrization requirements for 7 protons and 7 neutrons. It does not have to be antisymmetric if a proton is exchanged with any one of the 7 neutrons. So antisymmetrization is less confining. In general, a state with t_T greater than $|T_3|$ must work for more nuclei than one with $t_T = |T_3|$.

(Another argument, offered in literature, is essentially the reverse of the one that gives rise to the so-called "Hund rule" for atoms. Simply put, the Hund rule says that a couple of electrons maximize their spin, given the option between single-particle states of the same energy. The reason is that this allows electrons to stay farther apart, reducing their Coulomb repulsion, {A.34}. This argument reverses for nucleons, since they normally attract rather than repel each other. However, surely this is a relatively minor effect? Consider 3 nucleons. For these, the highest value $t_T = \frac{3}{2}$ allows the possibility that all 3 are protons. Within a single-particle-state picture, only one can go into the lowest energy state; the second must go into the second lowest energy state, and the third in the third lowest. On the other hand, for say 2 protons and 1 neutron, the neutron can go into the lowest energy state with the first proton. So the lower value $t_T = \frac{1}{2}$ should normally have significantly less energy.)

For the deuteron $T_3 = 0$, so the lowest possible value of t_T is 0. Then according to the general rule above, the ground state of the deuteron should have $t_T = 0$. That was already established above; it was the bound state not shared with the diproton and dineutron. Nitrogen-14 also has $T_3 = 0$, which means it too must have $t_T = 0$ in its ground state. This lowest energy state cannot occur for carbon-14 or oxygen-14, because they have $|T_3| = 1$. So nitrogen-14 should have less energy in its ground state than carbon-14 and oxygen-14. That seems at first surprising since nitrogen-14 has odd numbers of protons and neutrons, while carbon-14 and oxygen-14 have even numbers. Normally odd-odd nuclei are less tightly bound than even-even ones.

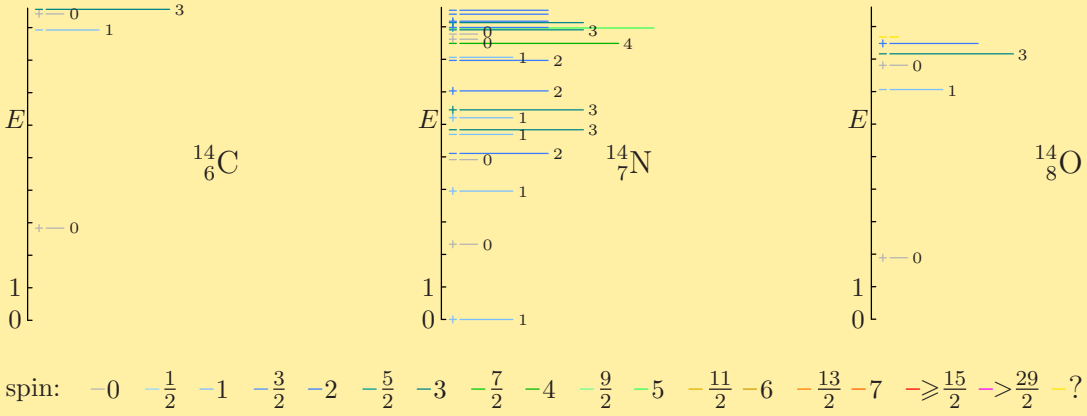


Figure 14.46: Isobaric analog states. [pdf]

But it is true. Figure 14.46 shows the energy levels of carbon-14, nitrogen-14, and oxygen-14. More precisely, it shows their binding energy, relative to the ground state value for nitrogen-14. In addition the von Weizsäcker value for the Coulomb energy has been subtracted to more clearly isolate the nuclear force effects. (The combined effect is simply to shift the normal spectra of carbon-14 and oxygen-14 up by 2.83, respectively 1.89 MeV.) It is seen that nitrogen-14 is indeed more tightly bound in its ground state than carbon-14 and oxygen-14. Qualitatively, since nitrogen-14 does not have 8 nucleons of the same kind, it has an easier job with satisfying the antisymmetrization requirements.

Traces of the lower energy of light nuclei with $T_3 = 0$ can also be detected in figures like 14.4, and 14.5 through 14.8. In these figures $T_3 = 0$ straight above the $Z = N = 2$ helium nucleus. Note in particular a distinct dark/light discontinuity in figures 14.5 through 14.8 along this vertical line. This discontinuity is quite distinct both from the magic numbers and from the average stability line that curves away from it.

Carbon-14 and oxygen-14 are mirror nuclei, so you would expect them to have pretty much the same sort of energy levels. Indeed, any oxygen-14 state, having $T_3 = 1$, must be part of a multiplet with t_T at least 1. Such a multiplet must have an equivalent state with $T_3 = -1$, which means an equivalent carbon-14 state. To the extent that the nuclear force is truly charge-independent, and the Coulomb effect has been properly removed, these two states should have the same energy. And indeed, the lowest four energy states of carbon-14 and oxygen-14 have identical spin and parity and similar energies. Also, both even-even nuclei have a 0^+ ground state, as even-even nuclei should. These ground states have $t_T = 1$, the smallest possible.

Now each of these multiplets should also have a version with $T_3 = 0$, which means a nitrogen-14 state. So any state that carbon-14 and oxygen-14 have should also exist for nitrogen-14. For example, the ground states of carbon-14 and oxygen-14 should also appear as a 0^+ state with $t_T = 1$ in the nitrogen-14 spectrum. Indeed, if you inspect the energy levels for nitrogen-14 in figure 14.46, exactly halfway in between the carbon-14 and oxygen-14 ground state energies, there it is!

Ideally speaking, these three states should have the same height in the figure. But it would be difficult to remove the Coulomb effect completely. And charge independence is not exact either, even though it is quite accurate.

A similar $T_3 = 0$ state can readily be found for the first three excited levels of carbon-14 and oxygen-14. In each case there is a nitrogen-14 state with exactly the same spin and parity and $t_T = 1$ right in between the matching carbon-14 and oxygen-14 levels. (To be sure, ENSDF does not list the t_T values for carbon-14 above the ground state. But common sense says they must be the same as the corresponding states in nitrogen-14 and carbon-14. For the first excited state of carbon-14, this is confirmed in [50, p. 11].)

Figure 14.46 also shows that nitrogen-14 has a lot more low energy states

than carbon-14 or oxygen-14. Square nucleon type can explain that too: all the low-lying states of nitrogen-14 that are not shared with carbon-14 and oxygen-14 are $t_T = 0$ states. These states are not possible for the other two nuclei.

Nothing is perfect, of course. The first state with nonzero t_T in the nitrogen spectrum besides the mentioned four isobaric analog states is the 0^-, $t_T = 1$ state at 8.8 MeV, just below the 3^- analog state. Carbon-14 has a 0^- state immediately above the 3^- state, but oxygen-14 has no obvious candidate.

Despite such imperfections, consideration of nucleon type is quite helpful for understanding the energy levels of light nuclei. And a lot of it carries over to heavier nuclei, [50, p. 12] and [36, p. 57]. While heavier nuclei have significant Coulomb energy, this long-range force is apparently often not that important here.

Now all that is needed is a good name. "Nucleon type" or "nucleon class" are not acceptable; they would give those hated outsiders and pesky students a general idea of what physicists were talking about. However, physicists noted that there is a considerable potential for confusion between nucleon type and spin, since both are described by the same mathematics. To maximize that potential for confusion, physicists decided that nucleon type should be called "spin."

Of course, physicists themselves still have to know whether they are talking about nucleon type or spin. Therefore some physicists called nucleon type "isobaric spin," because what differentiates isobars is the value of the net T_3. Other physicists talked about "isotopic spin," because physicists like to think of isotopes, and hey, isotopes have nucleon type too. Some physicists took the isowhatever spin to be $\frac{1}{2}$ for the proton, others for the neutron. However, that confused physicists themselves, so eventually it was decided that the proton has $\frac{1}{2}$. Also, a great fear arose that the names might cause some outsiders to suspect that the "spin" being talked about was not really spin. If you think about it, "isobaric angular momentum" or "isotopic angular momentum" does not make much sense. So physicists shortened the name to "isospin." Isospin means "equal spin" plain and simple; there is no longer anything to give the secret away that it is something completely different from spin. However, the confusion of having two different names for the same quantity was missed. Therefore, the alternate term "i-spin" was coined besides isospin. It too has nothing to give the secret away, and it restores that additional touch of confusion.

Isospin is conserved when only the nuclear force is relevant. As an example, consider the reaction in which a deuteron kicks an alpha particle out of an oxygen-16 nucleus:

$$^{16}_{8}\mathrm{O} + ^{2}_{1}\mathrm{H} \quad \rightarrow \quad ^{14}_{7}\mathrm{N} + ^{4}_{2}\mathrm{He}$$

The oxygen is assumed to be in the ground state. That is a $t_T = 0$ state, in agreement with the fact that oxygen-16 is a light nucleus with $T_3 = 0$. The deuteron can only be in a $t_T = 0$ state; that is the only bound state. The

alpha particle will normally be in the ground state, since it takes over 20 MeV to excite it. That ground state is a $t_T = 0$ one, since it is a light $T_3 = 0$ nucleus. Conservation of isospin then implies that the nitrogen-14 must have $t_T = 0$ too. The nitrogen can come out excited, but it should not come out in its lowest excited state, the 0^+ $t_T = 1$ state shared with carbon-14 and oxygen-14 in figure 14.46. Indeed, experiments show that this lowest excited state is only produced in negligible amounts compared to the surrounding states.

Selection rules for which nuclear decays occur can also be formulated based on isospin. If the electromagnetic force plays a significant part, T_3 but not $\vec{T}$ is conserved. The weak force does not conserve T_3 either, as beta decay shows. For example, the ground states of oxygen-14 and carbon-14 in figure 14.46 will beta-decay to the ground state of nitrogen 14, changing both T_3 and t_T. (Oxygen-16 will also beta-decay to the corresponding isobaric analog state of nitrogen-14, a decay that is called "superallowed," because it is unusually fast. It is much faster than to the ground state, even though decay to the ground state releases more energy. Carbon-14 has too little energy to decay to the analog state.)

Despite the lack of isospin conservation, isospin turns out to be very useful for understanding beta and gamma decay. See for example the discussion of superallowed beta decays in chapter 14.19, and the isospin selection rules for gamma decay in section 14.20.2.

14.18.3 Draft: Additional points

There are other particles besides nucleons that are also pretty much the same except for electric charge, and that can also be described using isospin. For example, the positive, neutral, and negatively charged pions form an isospin triplet of states with $t_T = 1$. Isospin was quite helpful in recognizing the existence of the more basic particles called quarks that make up baryons like nucleons and mesons like pions. In final analysis, the usefulness of isospin is a consequence of the approximate properties of these quarks.

Some sources incorrectly credit the concept of isospin to Heisenberg. But he did not understand what he was doing. Heisenberg did correctly guess that protons and neutrons might be described as two variants of the same particle. He then applied the only quantum approach for a two-state particle to it that he knew, that of spin. However, the mathematical machinery of spin is designed to deal with two-state properties that are preserved under rotations of an axis system, compare {A.19}. That is an inappropriate mathematical approach to describe nucleon type in the absence of charge independence. And at the time Heisenberg himself believed that the nuclear force was far from charge-independent.

(Because the nuclear force is in fact approximately charge-independent, unlike Heisenberg assumed, isospin is preserved under rotations of the abstract 1,2,3 coordinate system as defined in the first subsection. Phrased more simply,

without charge independence, energy eigenfunctions would not have definite values of square isospin t_T. That would make isospin self-evidently "entirely useless," as Wigner pointed out. This point is not very clear from the example of two nucleons in empty space, as discussed above. That is because there the spatial wave function happens to be symmetric under particle exchange even without charge independence. But if you express the isospin states in the general wave function (14.41) in terms of the singlet and triplet states, you quickly see the problem.)

The recognition that isospin was meaningful only in the presence of charge independence, and the proposal that the nuclear force is indeed quite accurately charge-independent, was mostly due to Wigner, in part with Feenberg. Some initial steps had already been taken by other authors. In particular, Cassen & Condon had already proposed to write wave functions in a form to include isospin,

$$\psi = \psi(\vec{r}_1, S_{z1}, T_{31}, \vec{r}_2, S_{z2}, T_{32}, \ldots)$$

and proposed symmetry under particle exchange in that form. This is the form of wave functions as written down earlier for the two-nucleon system. Still Wigner is considered the founding father of the study of isospin. His identification of isospin for complex nuclei as we know it today, as a preserved quantum number due to charge independence, is the foundation charter of nuclear isospin. Wigner is also the infernal idiot who decided that "nucleon type" should be called "spin."

See Wilkinson, [50, p. vi, 1-13], for a more extensive discussion of these historical issues. A very different history is painted by Henley in the next chapter in the same book. In this history, Heisenberg receives all the credit. Wigner does not exist. However, the author of this history implicitly admits that Heisenberg did think that the nuclear force was far from charge-independent. Maybe the author understood isospin too poorly to recognize that that is a rather big problem. Certainly there is no discussion. Or the author had a personal issue with Wigner and was willing to sacrifice his scientific integrity for it. Either way, the credibility of the author of this particular history is zero.

14.18.4 Draft: Why does this work?

It may seem astonishing that all this works. Why would nucleon type resemble spin? Spin is a vector in three-dimensional space, not a simple number. Why would energy eigenstates be unchanged under rotations in some weird abstract space?

The simplest and maybe best answer is that nature likes this sort of mathematics. Nature just loves creation and annihilation operators. But still, why would that lead to preserved lengths of vectors in an abstract spaces?

An answer can be obtained by looking a bit closer at square spin. Consider first two spin ½ fermions. Compare the dot product of their spins to the operator

$\widehat{P}^s_{12}$ that exchanges their spins:

$$\widehat{\vec{S}}_1 \cdot \widehat{\vec{S}}_2 |0\,0\rangle = -\tfrac{3}{4}\hbar^2 |0\,0\rangle \qquad \widehat{P}^s_{12}|0\,0\rangle = -|0\,0\rangle$$
$$\widehat{\vec{S}}_1 \cdot \widehat{\vec{S}}_2 |1\,m_s\rangle = \tfrac{1}{4}\hbar^2 |1\,m_s\rangle \qquad \widehat{P}^s_{12}|1\,m_s\rangle = |1\,m_s\rangle$$

The first set of relations is derived in {A.10}. The second set can be verified by looking at the expressions of the spin states (5.26).

Comparing the two sets of relations, it is seen that the dot product of two spins is closely related to the operator that exchanges the two spins:

$$\widehat{\vec{S}}_1 \cdot \widehat{\vec{S}}_2 = \tfrac{1}{4}\hbar^2 (2\widehat{P}^s_{12} - 1)$$

Now consider the square spin of a system of I fermions. By definition

$$\widehat{\vec{S}}^2 \equiv \left(\sum_{i=1}^{I} \widehat{\vec{S}}_i \right) \cdot \left(\sum_{\underline{i}=1}^{I} \widehat{\vec{S}}_{\underline{i}} \right) = \sum_{i=1}^{I} \sum_{\underline{i}=1}^{I} \widehat{\vec{S}}_i \cdot \widehat{\vec{S}}_{\underline{i}}$$

Split up the sum into terms that have i and $\underline{i}$ equal, respectively not equal:

$$\widehat{\vec{S}}^2 = \sum_{i=1}^{I} \widehat{\vec{S}}_i^{\,2} + 2 \sum_{i=1}^{I} \sum_{\underline{i}=i+1}^{I} \widehat{\vec{S}}_i \cdot \widehat{\vec{S}}_{\underline{i}}$$

The first sum is just the square spin angular momentum of the individual fermions. The second sum can be written in terms of the exchange operators using the expression above. Doing so and cleaning up gives:

$$\widehat{\vec{S}}^2 = \hbar^2 (I - \tfrac{1}{4} I^2) + \hbar^2 \sum_{i=1}^{I} \sum_{\underline{i}=i+1}^{I} \widehat{P}^s_{i\underline{i}}$$

Similarly then for isospin as defined in the first subsection,

$$\widehat{\vec{T}}^2 = (I - \tfrac{1}{4} I^2) + \sum_{i=1}^{I} \sum_{\underline{i}=i+1}^{I} \widehat{P}^T_{i\underline{i}}$$

Square isospin by itself does not have direct physical meaning. However, the exchange operators do. In particular, charge independence means that exchanging nucleon types does not make a difference for the energy. That then means that $\widehat{\vec{T}}^2$ commutes with the Hamiltonian. That makes it a conserved quantity according to the rules of quantum mechanics, chapter 4.5.1 and/or {A.19}.

It may be noted that the exchange operators do not commute among themselves. That makes the symmetry requirements so messy. However, it is possible to restrict consideration to exchange operators of the form $\widehat{P}_{i\,i+1}$. See [14] for more.

Infinitesimal "rotations" of a state in 1,2,3 isospin state correspond to applying small multiples of the operators $\widehat{T}_1$, $\widehat{T}_2$ and $\widehat{T}_3$, compare {A.19}. According to the definitions of $\widehat{T}_1$ and $\widehat{T}_2$, this corresponds to applying small multiples of the charge creation and annihilation operators. So it amounts to gradually changing protons into neutrons and vice-versa. As a simple example, a 180° rotation around the 1 or 2 axis inverts the 3-component of every nucleon. That turns every proton into a neutron and vice-versa.

14.19 Draft: Beta decay

14.19.1 Draft: Introduction

Beta decay is the decay mechanism that affects the largest number of nuclei. It is important in a wide variety of applications, such as betavoltaics and PET imaging.

In standard beta decay, or more specifically, beta-minus decay, a nucleus converts a neutron into a proton. The number of neutrons N decreases by one unit, and the number of protons Z increases by one. So the neutron excess decreases by two. Beta decay moves nuclei with too many neutrons closer to the stable range.

Unlike the neutron, the proton has a positive charge, so by itself, converting a neutron into a proton would create charge out of nothing. However, that is not possible as net charge is preserved in nature. In beta decay, the nucleus also emits a negatively charged electron, making the net charge that is created zero as it should.

But there is another problem with that. Now a neutron with spin $\frac{1}{2}$ is converted into a proton and an electron, each with spin $\frac{1}{2}$. That violates angular momentum conservation. (Regardless of any orbital angular momentum, the net angular momentum would change from half-integer to integer.) In beta decay, the nucleus also emits a second particle of spin $\frac{1}{2}$, thus keeping the net angular momentum half-integer. Fermi called that second particle the neutrino, since it was electrically neutral and so small that it was initially impossible to observe. In fact, even at the time of writing, almost a century later, the mass of the neutrino, though known to be nonzero, is too small to measure.

Nowadays the neutrino emitted in beta decay is more accurately identified as the electron antineutrino. An antineutrino is the antiparticle of an ordinary neutrino, just like the positron is the antiparticle of the electron. (Particles and antiparticles are exact opposites in all properties except mass, but including charge, allowing a particle and the corresponding antiparticle to annihilate each other, leaving only photons.)

The reason that an antineutrino is emitted rather than a neutrino is known as "conservation of lepton number." Leptons are elementary particles that do not

respond to the "strong force," including electrons and neutrinos. The net lepton number is defined as the number of leptons, minus the number of antileptons. It is found that this number is conserved in nature. So when in beta decay the nucleus emits an electron, a lepton, and an antineutrino, an antilepton, the lepton number stays unchanged as it should (like net angular momentum and net charge stay unchanged, as already noted).

The antineutrino does not affect the basics of beta decay, as it has no charge and virtually zero mass. However, the antineutrino does affect the detailed analysis; for one, the antineutrino can come out with a lot of kinetic energy, thus reducing the otherwise expected kinetic energy of the electron.

In beta decay, the new nucleus must be lighter than the original one. Classical mass conservation would say that the reduction in nuclear mass must equal the mass of the emitted electron plus the (negligible) mass of the antineutrino. However, Einstein's mass-energy relation implies that that is not quite right. Mass is equivalent to energy, and the rest mass reduction of the nucleus must also provide the kinetic energies of the electron and neutrino, as well as the (much smaller) one that the nucleus itself picks up during the decay by "recoil".

Still, the bottom line is that the nuclear mass reduction must be *at least* the rest mass of the electron (plus antineutrino). In energy units, it must be at least 0.511 MeV, the rest mass energy of the electron. The first subsection below will graphically examine which nuclei have enough energy to beta decay.

Beta-plus decay is the opposite of beta decay. In beta-plus decay, the nucleus converts a neutron into a proton instead of the other way around. To conserve charge, the nucleus can emit a positron, and with it, an electron neutrino to conserve angular momentum and lepton number.

However, while converting a proton into a neutron, the nucleus has a much easier way to conserve charge. Instead of emitting a positively charged positron, it can absorb a negatively charged electron from the atom it is in. The electron's charge then cancels that of the proton. To preserve angular momentum and lepton number, an electron neutrino is again emitted. This process is called "electron capture" (or also "K-capture" or "L-Capture" depending on the electron shell name from which the electron is swiped). Now the nuclear mass reduction does not need to provide the 0.512 MeV rest mass energy of a positron. Instead the nuclear mass can increase up to the 0.512 MeV rest mass energy of the electron that disappears.

So electron capture can occur in circumstances where positron creation is not possible. However, if the nuclear mass reduction is plenty for both electron capture and positron emission, the latter tends to dominate. The reason is the large quantum mechanical uncertainty in position of the low-energy atomic electron. This uncertainty dwarfs the size of the nucleus. It makes it very unlikely for the electron to be found inside the nucleus. A high-energy positron created by the nucleus itself can be created in any state, including high energy ones with short wave lengths.

Also note electron capture is of course not possible if somehow the nucleus has been stripped of all its atomic electrons, like might occur in space.

Electron capture is also called "inverse beta decay," because an electron being absorbed by a nucleus is much like a movie of an electron being emitted played backwards in time. But there are some problems with this idea. For one, the time-reversed movie would also have an electron antineutrino going into the nucleus, not an electron neutrino coming out.

Still, absorption of a particle is much like the emission of the corresponding antiparticle, at least as far as conservation laws other than energy are concerned. For example, capture of an electron adds one unit of negative charge, while emission of a positron removes one unit of positive charge. Either way, the nuclear charge becomes one unit more negative. In those terms, the notion of "inverse beta decay" may not be that far out, especially since the neutrino is a minor actor in the first place.

14.19.2 Draft: Energetics Data

As the introduction explained, in beta decay a nucleus converts a neutron into a proton, thus changing into a different nucleus. It can only occur if the nuclear mass reduction exceeds the 0.511 MeV rest mass energy of the electron emitted in the process.

Figures 14.47 through 14.50 show the nuclear mass reduction for beta decay as the vertical coordinate. The reduction exceeds the rest mass energy of the electron only above the horizontal center bands. The left half of each square indicates the nucleus before beta decay, the right half the one after the decay. The horizontal coordinate indicates the atomic numbers, with the values and element symbols as indicated. Neutron numbers are listed at the square itself. Lines connect pairs of nuclei with equal neutron excess.

If the left-half square is colored blue, beta decay is observed. Blue left-half squares are only found above the center bands, so the mass reduction is indeed at least the mass of the electron. However, some blue left-half squares are right on top of the band. Their beta decay should be very slow.

Note that some left-half squares above the band may not be blue. The color indicates the *dominant* decay process, so if the left-hand nucleus also experiences another decay mode in addition to beta decay, like alpha decay or beta-plus decay with another nucleus, and at a higher rate, its half square will not be blue. However, there should be no left-half squares above the band that are stable green.

Note here that even-even nuclei ${}^{48}_{20}\mathrm{Ca}$ and ${}^{96}_{40}\mathrm{Zr}$ are *not* stable, However, their beta decay is so extremely slow that double-beta decay dominates. Normal beta-decay has never been observed for them. This is not just because the energy release is small, but more importantly because these transitions are strongly "forbidden" in the sense discussed in section 14.19.6.

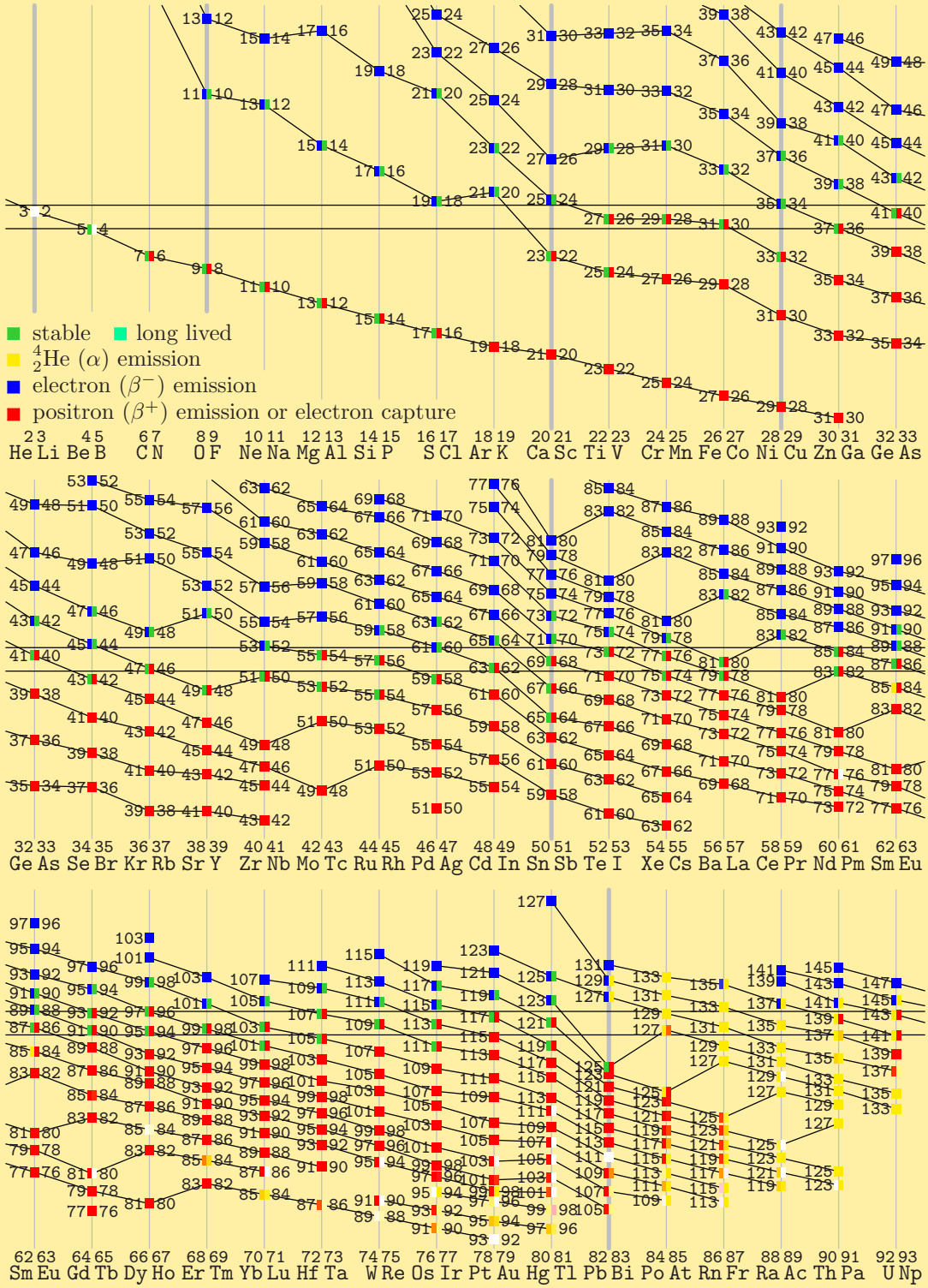


Figure 14.47: Energy release in beta decay of even-odd nuclei. [pdf]

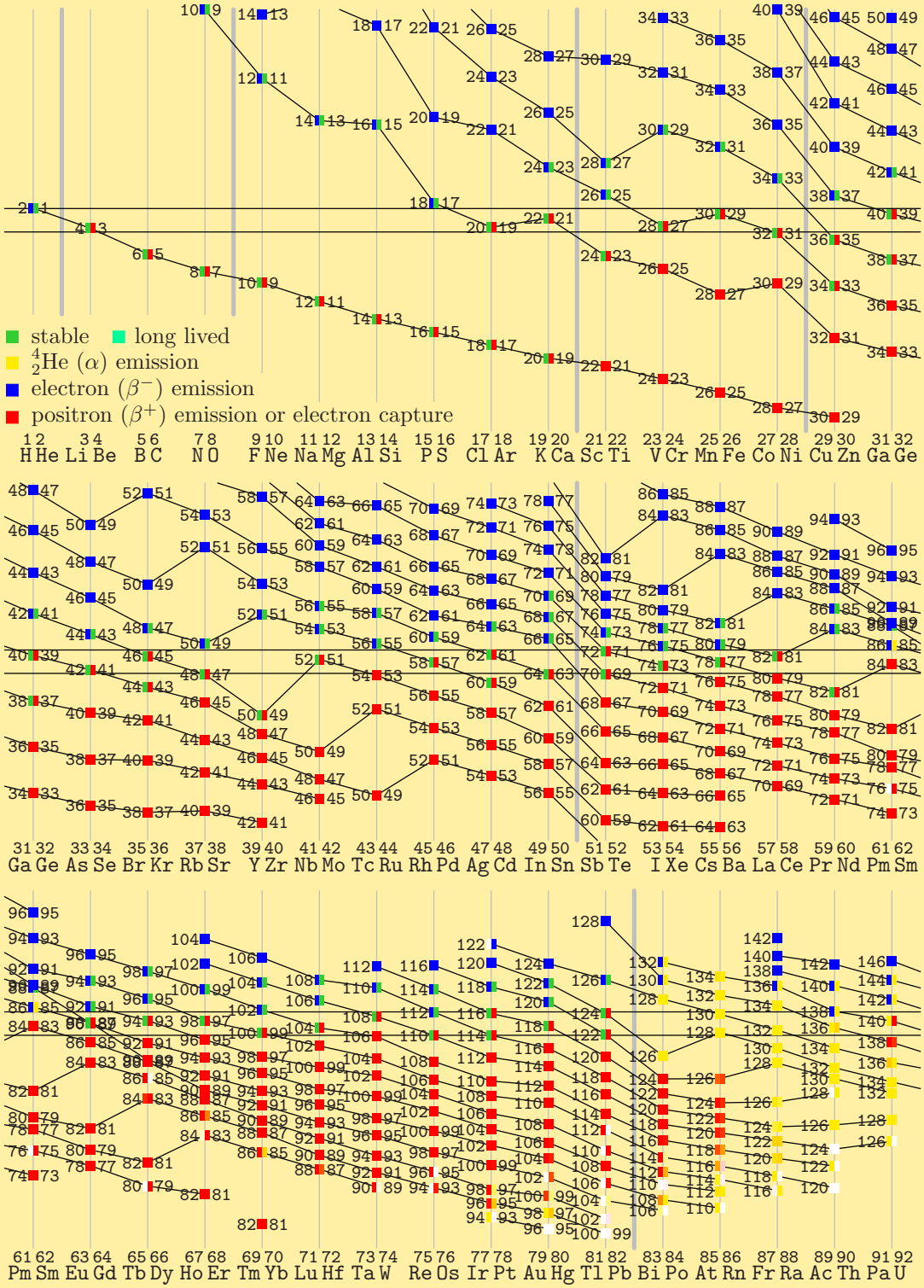


Figure 14.48: Energy release in beta decay of odd-even nuclei. [pdf]

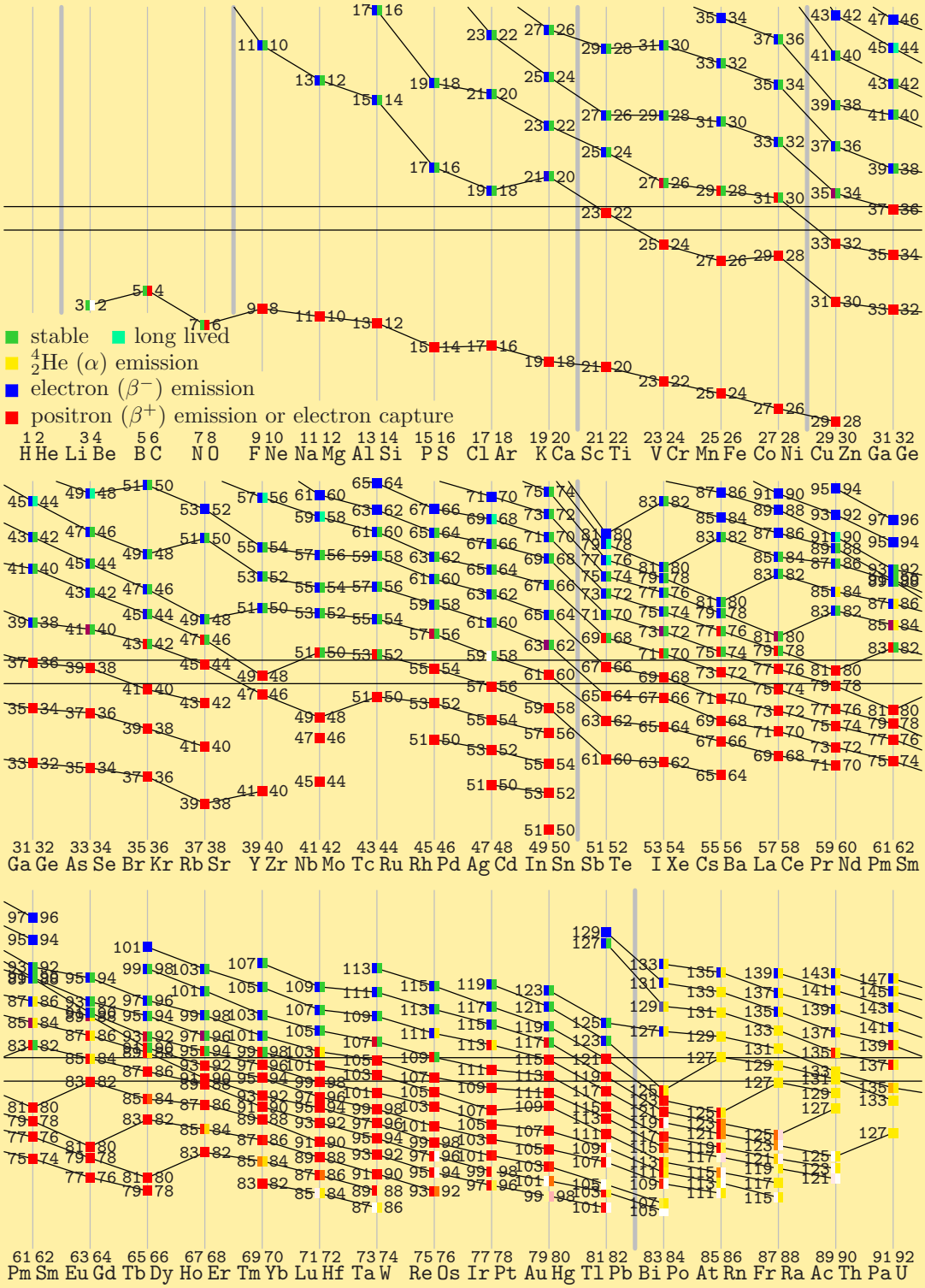


Figure 14.49: Energy release in beta decay of odd-odd nuclei. [pdf]

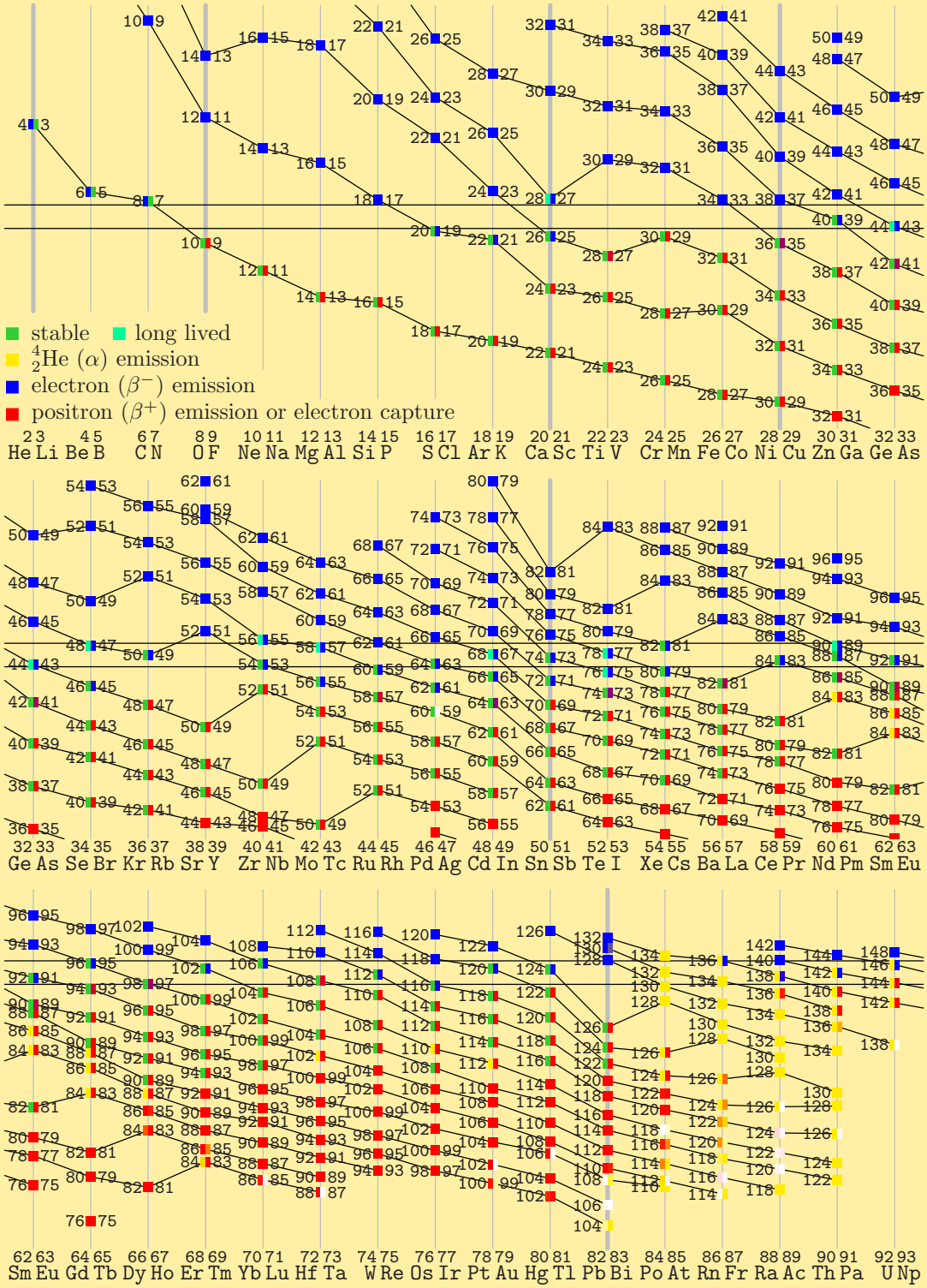


Figure 14.50: Energy release in beta decay of even-even nuclei. [pdf]

In beta-plus decay, the nucleus converts a proton into a neutron instead of the other way around. To find the energy release in that process, the figures may be read the other way around. The nucleus before the decay is now the right hand one, and the decay is observed when the right-half square is red.

The energy release is now positive downward, and it is now below the center bands that the nuclear mass reduction is sufficient to produce the rest mass of a positron that can carry the proton' positive charge away. The positron, the anti-particle of the electron, has the same mass as the electron but opposite charge.

But note that red right-half squares extend to *within* the center bands. The reason is that instead of emitting a positron, the nucleus can capture an electron from the atomic electron cloud surrounding the nucleus. In that case, rather than having to come up with an electron mass worth of energy, the nucleus receives an infusion of that amount of energy. So the required energy goes down by two electron masses.

It follows that the left-hand nucleus will suffer beta decay if the square is above the top of the band, while the right-hand nucleus will suffer electron capture if the square is below the top of the band. Therefore at most one nucleus of each pair can be stable.

Note, once more, that color indicates the dominant decay mode. So right-half squares below the top of the band do not have to be red; they just should not be stable green. This is especially relevant for the odd-odd nuclei in figures 14.49 and 14.50. Odd-odd nuclei are unusually unstable, and the even-even nuclei they decay into are unusually stable. So it is quite likely that an odd-odd nuclei in the region of relatively stable nuclei finds that it has enough energy to both beta decay to the neighboring even-even nucleus of higher Z and beta-plus decay / electron capture to the neighboring even-even nucleus of lower Z. Then, if one of the two processes is relatively slow, because the process is just above the top of the band in figure 14.49, or just below the top of the band in figure 14.50, then the other process is likely to dominate. So the half square does not have the expected color. Left-hand squares just above the top of the band in figure 14.49 will be red, and right-hand squares just below the top of the band in figure 14.49 will be blue.

One example is ${}^{40}_{19}\mathrm{K}$ potassium-40, with 21 neutrons. It appears above the band in figure 14.49, indicating that it suffers beta decay. But it also appears below the band in figure 14.50, so that it also suffers electron capture and positron emission. In this case, beta decay dominates beta-plus decay and electron capture 9 to 1. Recall that electron capture is relatively slow and the nucleus is just below the bottom of the band in figure 14.50, so beta-plus decay will be too.

14.19.3 Draft: Beta decay and magic numbers

The magic neutron numbers are quite visible in figures 14.47 through 14.50. For example, diagonal bands at neutron numbers 50, 82, and 126 are prominent in all four figures. Consider for example figure 14.48. For the 50/49 neutron nuclei, beta decay takes the tightly bound 50th neutron to turn into a proton. That requires relatively large energy, so the energy release is reduced. For the neighboring 52/51 nuclei, beta decay takes the much less tightly bound 52nd neutron, and the energy release is correspondingly higher.

The magic proton numbers tend to show up as step-downs in the curves. For example, consider the nuclei at the vertical $Z = 50$ line also in figure 14.48. In the In/Sn (indium/tin) beta decay, the beta decay neutron becomes the tightly bound 50th proton, and the energy release is correspondingly high. In the Sb/Te (antimony/tellurium) decay, the neutron becomes the less tightly bound 52nd proton, and the energy release is lower.

When the neutron and proton magic number lines intersect, combined effects can be seen. One pointed out by Mayer in her Noble prize acceptance lecture [[10]] is the decay of argon-39. It has 18 protons and 21 neutrons. If you interpolate between the neighboring pairs of nuclei on the same neutron excess line in figure 14.47, you would expect argon-39 to be below the top of the center band, hence to be stable against beta decay. But the actual energy release for argon-39 is unusually high, and beta decay it does. Why is it unusually high? For the previous pairs of nuclei, beta decay converts a neutron in the neutron shell that ends at magic number 20 into a proton in the corresponding proton shell. For the suvsequent pairs, beta decay converts a neutron in the neutron shell that ends at magic number 28 to a proton in the corresponding proton shell. Only for argon-39, beta decay converts a neutron in the neutron shell that end at magic number 28 into a proton in the lower energy proton shell that ends at magic number 20. The lowering of the major shell releases additional energy, and the decay has enough energy to proceed.

In figures 14.47 and 14.48, the lowest line for the lightest nuclei is unusually smooth. These lines correspond to a neutron excess of 1 or -1, depending on whether it is before or after the decay. The pairs of nuclei on these two lines are mirror nuclei. During beta decay the neutron that turns into a proton transfers from the neutron shells into the exact same position in the proton shells. Because of charge independence, the nuclear energy does not change. The Coulomb energy does change, but as a relatively small, long-range effect, it changes fairly gradually.

These lines also show that beta-plus decay and electron capture become energetically favored when the nuclei get heavier. That is to be expected since this are nuclei with almost no neutron excess. For the heavier ones, it is therefore energetically favorable to convert protons into neutrons, rather than the other way around.

14.19.4 Draft: Von Weizsäcker approximation

Since the von Weizsäcker formula of section 14.10.2 predicts nuclear mass, it can be used to predict whether beta-minus or beta-plus/electron capture will occur.

The mathematics is relatively simple, because the mass number A remains constant during beta decay. For a given mass number, the von Weizsäcker formula is just a quadratic in Z. Like in the previous subsection, consider again pairs of nuclei with the same A and one unit difference in Z. Set the mass difference equal to the electron mass and solve the resulting equation for Z using simple algebra.

It is then seen that beta-minus decay, respectively beta-plus decay / electron capture occurs for a pair of nuclei depending whether the average Z value is less, respectively greater, than

$$\boxed{Z_{\rm bd} = A\frac{4C_a + m_{\rm n} - m_{\rm p} - m_{\rm e} + C_c C_z A^{-1/3}}{8C_a + 2C_c A^{2/3}}} \tag{14.50}$$

where the constants $C_.$ are as given in section 14.10.2. The nucleon pairing energy must be ignored in the derivation, so the result may be off by a pair of nuclei for even-even and odd-odd nuclei.

The result is plotted as the black curve in the decay graph figure 14.51. It gives the location where the change in nuclear mass is just enough for either beta-minus decay or electron capture to occur, with nothing to spare. The curve locates the stable nuclei fairly well. For light nuclei, the curve is about vertical, indicating there are equal numbers of protons and neutrons in stable nuclei. For heavier nuclei, there are more neutrons than protons, causing the curve to deviate to the right, the direction of increasing neutron excess.

Because of the pairing energy, stable even-even nuclei can be found well away from the curve. Conversely, stable odd-odd nuclei are hard to find at all. In fact, there are only four: hydrogen-2 (deuterium), lithium-6, boron-10, and nitrogen-14. For comparison, there are 150 stable even-even ones. For nuclei of odd mass number, it does not make much difference whether the number of protons is odd or the number of neutrons: there are 49 stable odd-even nuclei and 53 stable even-odd ones.

(There is also the bizarre excited ${}^{180\rm m}_{73}$Ta nucleus that is stable, and is odd-odd to boot. But that is an excited state and another story, which is discussed under gamma decay. The ground state ${}^{180}_{73}$Ta has a half life of only 8 hours, as a relatively heavy odd-odd nucleus should.)

As an example of the instability of odd-odd nuclei, consider the curious case of potassium-40, ${}^{40}_{19}K$. It has both an odd number of protons, 19, and of neutrons, 21. Potassium-40 is pretty much on top of the stable line, as evident from the fact that both its neighbors, odd-even isotopes potassium-39 and potassium-41, are stable. But potassium-40 itself is unstable. It does have

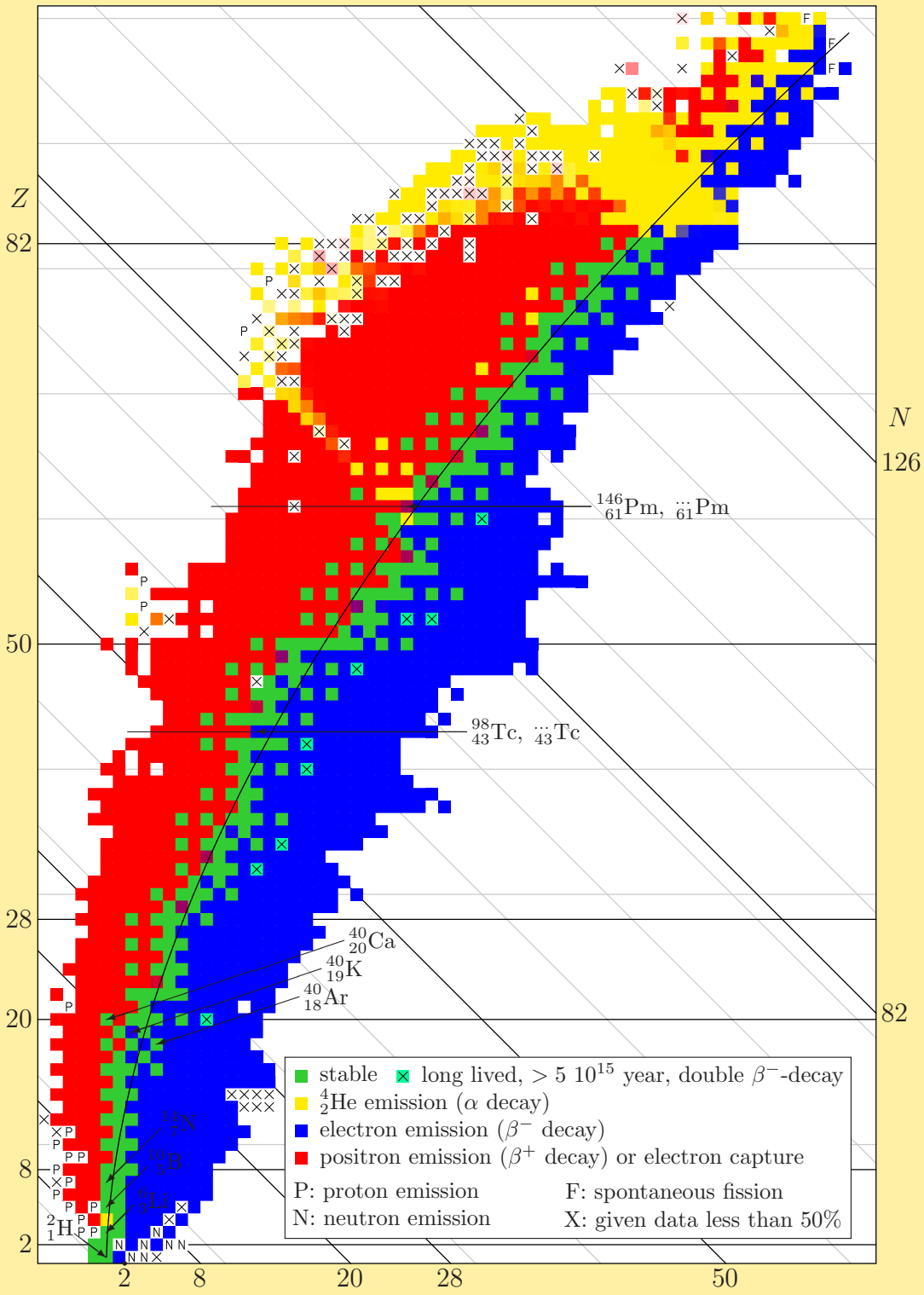


Figure 14.51: Examples of beta decay. [pdf]

a lifetime comparable to the age of the universe; long enough for significant quantities to accumulate. About 0.01% of natural potassium is potassium-40.

But decay it does. Despite the two billion year average lifetime, there are so many potassium-40 nuclei in a human body that almost 5 000 decay per second anyway. About 90% do so through beta decay and end up as the doubly-magic calcium-40. The other 10% decay by electron capture or positron emission and end up as even-even argon-40, with 18 protons and 22 neutrons. So potassium-40 suffers all three beta decay modes, the only relatively common nucleus in nature that does.

Admittedly only 0.001% decays through positron emission. The nuclear mass difference of 0.99 MeV with argon-40 is enough to create a positron, but not by much. Before a positron can be created, potassium is almost sure to have captured an electron already. For a nucleus like xenon-119 the mass difference with iodine-119 is substantially larger, 4.5 MeV, and about 4 in 5 xenon-119 nuclei decay by positron emission, and the fifth by electron capture.

It is energetically possible for the potassium-40 decay product calcium-40 to decay further into argon-40, by capturing two electrons from the atom. Energetically possible means that this does not require addition of energy, it liberates energy, so it can occur spontaneously. Note that calcium-40 would have to capture two electrons at the same time; capturing just one electron would turn it into potassium-40, and that requires external energy addition. In other words, calcium-40 would have to skip over the intermediate odd-odd potassium 40. While it is possible, it is believed that calcium-40 is stable; if it decays at all, its half-life must be more than 5.9 zettayear (5.9 10^{21} year).

But some even-even nuclei do decay through "double beta-minus" decay. For example, germanium-76 with 32 protons and 44 neutrons will in a couple of zettayear emit two electrons and so turn into even-even selenium-76, skipping over odd-odd arsenic-76 in the process. However, since the entire lifetime of the universe is much less than the blink of an eye compared to a zettayear, this does not get rid of much germanium-76. About 7.5% of natural germanium is germanium-76.

The reduced stability of odd-odd nuclei is the main reason that technetium (Tc) and promethium (Pm) can end up with no stable isotopes at all while their immediate neighbors have many. Both technetium and promethium have an odd-odd isotope sitting right on top of the separating line between beta-minus and beta-plus decay; technetium-98 respectively promethium-146. Because of the approximation errors in the von Weizsäcker formula, they are not quite on the theoretical curve in figure 14.51. However, examination of the experimental nuclear masses shows the excess mass reduction for beta-minus decay and electron capture to be virtually identical for these odd-odd nuclei. And in fact promethium-146 does indeed decay both ways. Technetium-98 could too, but does not; it finds it quicker to create an electron than to capture one from the surrounding atom.

Because the theoretical stable line slopes towards the right in figure 14.51, only one of the two odd-even isotopes next to technetium-98 should be unstable, and the same for the ones next to promethium-146. However, the energy liberated in the decay of these odd-even nuclei is only a few hundred keV in each case, far below the level for which the von Weizsäcker formula is anywhere meaningful. For technetium and promethium, neither neighboring isotope is stable. This is a qualitative failure of the von Weizsäcker model. But it is rare; it happens only for these two out of the lowest 82 elements. Few books even mention it is a fundamental failure of the formula.

14.19.5 Draft: Kinetic Energies

The kinetic energy of nuclear decay products is important to understand the correct nature of the decay.

Historically, one puzzling observation in beta decay was the kinetic energies with which the electrons came out. When the beta decay of a collection of nuclei of a given type is observed, the electrons come out with a range of kinetic energies. In contrast, in the alpha decay of a collection of nuclei of a given type, all alpha particles come out with pretty much the exact same kinetic energy.

Consider the reason. The total kinetic energy release in the decay of a given nucleus is called the "Q value." Following Einstein's famous relation $E = mc^2$, the Q value in alpha decay is given by the reduction in the net rest mass energy during the decay:

$$\boxed{Q = m_{\mathrm{N}1}c^2 - m_{\mathrm{N}2}c^2 - m_\alpha c^2} \tag{14.51}$$

where 1 indicates the nucleus before the decay and 2 the nucleus after the decay.

Since energy must be conserved, the reduction in rest mass energy given by the Q-value is converted into kinetic energy of the decay products. Classical analysis makes that:

$$Q = \tfrac{1}{2}m_{\mathrm{N}2}v_{\mathrm{N}2}^2 + \tfrac{1}{2}m_\alpha v_\alpha^2$$

This assumes that the initial nucleus is at rest, or more generally that the decay is observed in a coordinate system moving with the initial nucleus. Linear momentum must also be conserved:

$$m_{\mathrm{N}1}\vec{v}_{\mathrm{N}1} = m_{\mathrm{N}2}\vec{v}_{\mathrm{N}2} + m_\alpha\vec{v}_\alpha$$

but since the velocity of the initial nucleus is zero,

$$m_{\mathrm{N}2}\vec{v}_{\mathrm{N}2} = -m_\alpha\vec{v}_\alpha$$

Square both sides and divide by $2m_{\mathrm{N}2}$ to get:

$$\tfrac{1}{2}m_{\mathrm{N}2}v_{\mathrm{N}2}^2 = \frac{m_\alpha}{m_{\mathrm{N}2}}\tfrac{1}{2}m_\alpha v_\alpha^2$$

Now, excluding the special case of beryllium-8, the mass of the alpha particle is much smaller than that of the final nucleus. So the expression above shows that the kinetic energy of the final nucleus is much less than that of the alpha particle. The alpha particle runs away with almost all the kinetic energy. Its kinetic energy is almost equal to Q. Therefore it is always the same for a given initial nucleus, as claimed above. In the special case that the initial nucleus is beryllium-8, the final nucleus is also an alpha particle, and each alpha particle runs away with half the kinetic energy. But still, each alpha particle always comes out with a single value for its kinetic energy, in this case $\frac{1}{2}Q$.

In beta decay, things would be pretty much the same if just an electron was emitted. The electron too would come out with a single kinetic energy. The fact that it did not led Pauli to propose that another small particle also comes out. That particle could carry away the rest of the kinetic energy. It had to be electrically neutral like a neutron, because the nuclear charge change is already accounted for by the charge taken away by the electron. The small neutral particle was called the "neutrino" by Fermi. The neutrino was also required for angular momentum conservation: a proton and an electron each with spin $\frac{1}{2}$ have net spin 0 or 1, not $\frac{1}{2}$ like the original neutron.

The neutrino that comes out in beta-minus decay is more accurately called an electron antineutrino and usually indicated by $\bar{\nu}$. The bar indicates that it is counted as an antiparticle.

The analysis of the kinetic energy of the decay products changes because of the presence of an additional particle. The Q-value for beta decay is

$$\boxed{Q = m_{\mathrm{N}1}c^2 - m_{\mathrm{N}2}c^2 - m_{\mathrm{e}}c^2 - m_{\bar{\nu}}c^2} \tag{14.52}$$

However, the rest mass energy of the neutrino can safely be ignored. At the time of writing, numbers less than a single eV are bandied around. That is immeasurably small compared to the nuclear rest mass energies which are in terms of GeV. In fact, physicists would love the neutrino mass to be nonnegligible: then they could figure out what is was!

As an aside, it should be noted that the nuclear masses in the Q values are *nuclear* masses. Tabulated values are invariably *atomic* masses. They are different by the mass of the electrons and their binding energy. Other books therefore typically convert the Q-values to atomic masses, usually by ignoring the electronic binding energy. But using atomic masses in a description of nuclei, not atoms, is confusing. It is also a likely cause of mistakes. (For example, [31, fig. 11.5] seems to have mistakenly used atomic masses to relate isobaric nuclear energies.)

It should also be noted that if the initial and/or final nucleus is in an excited state, its mass can be computed from that of the ground state nucleus by adding the excitation energy, converted to mass units using $E = mc^2$. Actually, nuclear masses are usually given in energy units rather than mass units, so no conversion is needed.

Because the amount of kinetic energy that the neutrino takes away varies, so does the kinetic energy of the electron. One extreme case is that the neutrino comes out at rest. In that case, the given analysis for alpha decay applies pretty much the same way for beta decay if the alpha is replaced by the electron. This gives the maximum kinetic energy at which the electron can come out to be Q. (Unlike for the alpha particle, the mass of the electron is always small compared to the nucleus, and the nucleus always ends up with essentially none of the kinetic energy.) The other extreme is that the electron comes out at rest. In that case, it is the neutrino that pretty much takes all the kinetic energy. Normally, both electron and neutrino each take their fair share of kinetic energy. So usually the kinetic energy of the electron is somewhere in between zero and Q.

A further modification to the analysis for the alpha particle must be made. Because of the relatively small masses of the electron and neutrino, they come out moving at speeds close to the speed of light. Therefore the relativistic expressions for momentum and kinetic energy must be used, chapter 1.1.2.

Consider first the extreme case that the electron comes out at rest. The relativistic energy expression gives for the kinetic energy of the neutrino:

$$T_{\bar\nu} = \sqrt{(m_{\bar\nu}c^2)^2 + (pc)^2} - m_{\bar\nu}c^2 \tag{14.53}$$

where c is the speed of light and p the momentum. The nucleus takes only a very small fraction of the kinetic energy, so $T_{\bar\nu} \approx Q$. Also, whatever the neutrino rest mass energy $m_{\bar\nu}c^2$ may be exactly, it is certainly negligibly small. It follows that $T_{\bar\nu} \approx Q \approx pc$.

The small fraction of the kinetic energy that does end up with the nucleus may now be estimated, because the nucleus has the same magnitude of momentum p. For the nucleus, the nonrelativistic expression may be used:

$$T_{\mathrm{N2}} = \frac{p^2}{2m_{\mathrm{N2}}} = pc\frac{pc}{2m_{\mathrm{N2}}c^2} \tag{14.54}$$

The final fraction is very small because the energy release $pc \approx Q$ is in MeV while the nuclear mass is in GeV. Therefore the kinetic energy of the nucleus is indeed very small compared to that of the neutrino. If higher accuracy is desired, the entire computation may now be repeated, starting from the more accurate value $T_{\bar\nu} = Q - T_{\mathrm{N2}}$ for the kinetic energy of the neutrino.

The extreme case that the neutrino is at rest can be computed in much the same way, except that the rest mass energy of the electron is comparable to Q and must be included in the computation of pc. If iteration is not desired, an exact expression for pc can be derived using a bit of algebra:

$$pc = \sqrt{\frac{[E^2 - (E_{\mathrm{N2}} + E_{\mathrm{e}})^2][E^2 - (E_{\mathrm{N2}} - E_{\mathrm{e}})^2]}{4E^2}} \qquad E = E_{\mathrm{N2}} + E_{\mathrm{e}} + Q \tag{14.55}$$

where $E_{\rm N2} = m_{\rm N2}c^2$ and $E_{\rm e} = m_{\rm e}c^2$ are the rest mass energies of the final nucleus and electron. The same formula may be used in the extreme case that the electron is at rest and the neutrino is not, by replacing $E_{\rm e}$ by the neutrino rest mass, which is to all practical purposes zero.

In the case of beta-plus decay, the electron becomes a positron and the electron antineutrino becomes an electron neutrino. However, antiparticles have the same mass as the normal particle, so there is no change in the energetics. (There is a difference if it is written in terms of atomic instead of nuclear masses.) In case of electron capture, it must be taken into account that the nucleus receives an infusion of mass equal to that of the captured electron. The Q-value becomes

$$\boxed{Q = m_{\rm N1}c^2 + m_{\rm e}c^2 - m_{\rm N2}c^2 - m_{\bar\nu}c^2 - E_{\rm B,ce}} \tag{14.56}$$

where $E_{\rm B,ce}$ is the electronic binding energy of the captured electron. Because this is an inner electron, normally a K or L shell one, it has quite a lot of binding energy, too large to be ignored. After the electron capture, an electron farther out will drop into the created hole, producing an X-ray. If that electron leaves a hole behind too, there will be more X-rays. The energy in these X-rays subtracts from that available to the neutrino.

The binding energy may be ballparked from the hydrogen ground state energy, chapter 4.3, by simply replacing e^2 in it by e^2Z. That gives:

$$\boxed{E_{\rm B,ce} \sim 13.6\,Z^2 \text{ eV}} \tag{14.57}$$

The ballparks for electron capture in figure 14.54 use

$$E_{\rm B,ce} \sim \tfrac{1}{2}(\alpha Z)^2 m_{\rm e}c^2 \left(1 + \tfrac{1}{4}(\alpha Z)^2\right) \tag{14.58}$$

in an attempt to partially correct for relativistic effects, which are significant for heavier nuclei. Here $\alpha \approx 1/137$ is the so-called fine structure constant. The second term in the parentheses is the relativistic correction. Without that term, the result is the same as (14.57). See addendum {A.39} for a justification.

14.19.6 Draft: Forbidden decays

Energetics is not all there is to beta decay. Some decays are energetically fine but occur extremely slowly or not at all. Consider calcium-48 in figure fig:betdec2e. The square is well above the center band, so energy-wise there is no problem at all for the decay to scandium-48. But it just does not happen. The half life of calcium-48 is 64 10^{18} years, more than three billion times the entire lifetime of the universe. And when decay does happen, it is due to double beta decay; as of 2016, normal beta decay of calcium-48 has never been observed.

The big problem is angular momentum conservation. As an even-even nucleus, calcium-48 has zero spin, while scandium-48 has spin 6 in its ground

state. To conserve angular momentum during the decay, the electron and the antineutrino must therefore take six units of spin along. But to the extend that the nuclear size can be ignored, the electron and antineutrino come out of a mathematical point. That means that they come out with zero orbital angular momentum. They have half a unit of spin each, and there is no way to produce six units of net spin from that. The decay is forbidden by angular momentum conservation.

Of course, calcium-48 could decay to an excited state of scandium-48. Unfortunately, only the lowest two excited states are energetically possible, and these have spins 5 and 4. They too are forbidden.

14.19.6.1 Draft: Allowed decays

To understand what beta decays are forbidden, the first step is to examine what decays are allowed.

Consider the spins of the electron and antineutrino. They could combine into a net spin of zero. If they do, it is called a "Fermi decay." Since the electron and antineutrino take no spin away, in Fermi decays the nuclear spin cannot change.

The only other possibility allowed by quantum mechanics is that the spins of electron and antineutrino combine into a net spin of one; that is called a "Gamow-Teller decay." The rules of quantum mechanics for the addition of angular momentum vectors imply:

$$\boxed{|j_{\mathrm{N}1} - j_{\mathrm{e}\bar{\nu}}| \leqslant j_{\mathrm{N}2} \leqslant j_{\mathrm{N}1} + j_{\mathrm{e}\bar{\nu}}} \tag{14.59}$$

where $j_{\mathrm{N}1}$ indicates the spin of the nucleus before the decay, $j_{\mathrm{N}2}$ the one after it, and $j_{\mathrm{e}\bar{\nu}}$ is the combined angular momentum of electron and antineutrino. Since $j_{\mathrm{e}\bar{\nu}} = 1$ for allowed Gamow-Teller decays, spin can change one unit or stay the same. There is one exception; if the initial nuclear spin is zero, the final spin cannot be zero but must be one. Transitions from spin zero to zero are only allowed if they are Fermi ones. But they are allowed.

Putting it together, the angular momentum can change by up to one unit in an allowed beta decay. Also, if there is no orbital angular momentum, the parities of the electron and antineutrino are even, so the nuclear parity cannot change. In short

$$\boxed{\text{allowed:} \qquad |\Delta j_{\mathrm{N}}| \leqslant 1 \qquad \Delta\pi_{\mathrm{N}} = \text{no}} \tag{14.60}$$

where Δ indicates the nuclear change during the decay, j_{N} the spin of the nucleus, and π_{N} its parity.

One simple example of an allowed decay is that of a single neutron into a proton. Since this is a $\frac{1}{2}^+$ to $\frac{1}{2}^+$ decay, both Fermi and Gamow-Teller decays

occur. The neutron has a half-life of about ten minutes. It can be estimated that the decay is 18% Fermi and 82% Gamow-Teller, [31, p. 290].

Some disclaimers are in order. Both the discussion above and the following one for forbidden decays are nonrelativistic. But neutrinos are very light particles that travel at very close to the speed of light. For such relativistic particles, orbital angular momentum and spin get mixed-up. That is much like they get mixed-up for the photon. That was such a headache in describing electromagnetic transitions in chapter 7.4.3. Fortunately, neutrinos turn out to have some mass. So the given arguments apply at least under some conditions, even if such conditions are never observed.

A much bigger problem is that neutrinos and antineutrinos do not conserve parity. That is discussed in more detail a later subsection, Above, this book simply told you a blatant lie when it said that the electron-antineutrino system, (or the positron-neutrino system in beta-plus decay), comes off with zero parity. A system involving a single neutrino or antineutrino does not have definite parity. And parity is not conserved in the decay process anyway. But the initial and final nuclear states do have definite parity (to within very high accuracy). Fortunately, it turns out that you get the right answers for the change in *nuclear* parity if you simply assume that the electron and antineutrino come off with the parity given by their "orbital" angular momentum.

No you cannot have your money back. You did not pay any.

A relativistic description of neutrinos can be found in {A.44}.

14.19.6.2 Draft: Forbidden decays allowed

As noted at the start of this subsection, beta decay of calcium-48 requires a spin change of at least 4 and that is solidly forbidden. But forbidden is not quite the same as impossible. There is a small loophole. A nucleus is not really a mathematical point, it has a nonzero size.

Classically that would not make a difference, because the orbital angular momentum would be much too small to make up the deficit in spin. A rough ballpark of the angular momentum of, say, the electron would be pR, with p its linear momentum and R the nuclear radius. Compare this with the quantum unit of angular momentum, which is $\hbar$. The ratio is

$$\frac{pR}{\hbar} = \frac{pc\,R}{\hbar c} = \frac{pc\,R}{197 \text{ MeV fm}}$$

with c the speed of light. The product pc is comparable to the energy release in the beta decay and can be ballparked as on the order of 1 MeV. The nuclear radius ballparks to 5 fm. As a result, the classical orbital momentum is just a few percent of $\hbar$.

But quantum mechanics says that the orbital momentum *cannot* be a small fraction of $\hbar$. Angular momentum is quantized to values $\sqrt{l(l+1)}\hbar$ where l must

be an integer. For $l = 0$ the angular momentum is zero, for $l = 1$ the angular momentum is $\sqrt{2}\hbar$. There is nothing in between. An angular momentum that is a small fraction of $\hbar$ is not possible. Instead, what is small in quantum mechanics is the *probability* that the electron has angular momentum $l = 1$. If you try long enough, it may happen.

In particular, $pR/\hbar$ gives a rough ballpark for the quantum amplitude of the $l = 1$ state. (The so-called Fermi theory of beta decay, {A.45}, can be used to justify this and other assumptions in this section.) The probability is the square magnitude of the quantum amplitude, so the probability of getting $l = 1$ is roughly $(pR/\hbar)^2$ smaller than getting $l = 0$. That is about 3 or 4 orders of magnitude less. It makes decays that have $l = 1$ that many orders of magnitude slower than allowed decays, all else being the same. But if the decay is energetically possible, and allowed decays are not, it will eventually happen. (Assuming of course that some completely different decay like alpha decay does not happen first.)

Decays with $l = 1$ are called "first-forbidden decays." The electron and neutrino can then take up to 2 units of angular momentum away through their combined orbital angular momentum and spin. So the nuclear spin can change up to two units. Orbital angular momentum has negative parity if l is odd, so the parity of the nucleus must change during the decay. Therefore the possible changes in nuclear spin and parity are:

$$\boxed{\text{first-forbidden:} \qquad |\Delta j_{\text{N}}| \leqslant 2 \qquad \Delta\pi_{\text{N}} = \text{yes}} \tag{14.61}$$

That will not do for calcium-48, because at least 4 units of spin change is needed. In "second-forbidden decays," the electron and neutrino come out with a net orbital angular momentum $l = 2$. Second forbidden decays are another 3 or 4 order of magnitude slower still than first forbidden ones. The nuclear parity remains unchanged like in allowed decays. Where both allowed and second forbidden decays are possible, the allowed decay should be expected to have occurred long before the second forbidden one has a chance. Therefore, the interesting second-forbidden cases cases are those that are not allowed ones:

$$\boxed{\text{second-forbidden:} \qquad |\Delta j_{\text{N}}| = 2 \text{ or } 3 \qquad \Delta\pi_{\text{N}} = \text{no}} \tag{14.62}$$

In third forbidden decays, $l = 3$. The transitions that become possible that were not in first forbidden ones are:

$$\boxed{\text{third-forbidden:} \qquad |\Delta j_{\text{N}}| = 3 \text{ or } 4 \qquad \Delta\pi_{\text{N}} = \text{yes}} \tag{14.63}$$

These transitions are still another 3 or 4 orders of magnitude slower than second forbidden ones. And they do not work for calcium-48, as both the calcium-48 ground state and the three reachable scandium-48 states all have equal, positive, parity.

Beta decay of calcium-48 is possible through fourth-forbidden transitions:

$$\boxed{\text{fourth-forbidden:} \qquad |\Delta j_{\rm N}| = 4 \text{ or } 5 \qquad \Delta\pi_{\rm N} = \text{no}} \tag{14.64}$$

This allows decay to either the 5^+ and 4^+ excited states of scandium-48. However, fourth forbidden decays are generally impossibly slow.

14.19.6.3 Draft: The energy effect

There is an additional effect slowing down the beta decay of the 0^+ calcium-48 ground state to the 5^+ excited scandium-48 state. The energy release, or Q-value, of the decay is only about 0.15 MeV.

One reason that is bad news, (or good news, if you like calcium-48), is because it makes the momentum of the electron and neutrino correspondingly small. The ratio $pR/\hbar$ is therefore quite small at about 0.01. And because this is a fourth forbidden decay, the transition is slowed down by a ballpark $((pR/\hbar)^{-2})^4$; that means a humongous factor 10^{16} for $pR/\hbar = 0.01$. If a 1 MeV allowed beta decay may take on the order of a day, you can see why calcium-48 is effectively stable against beta decay.

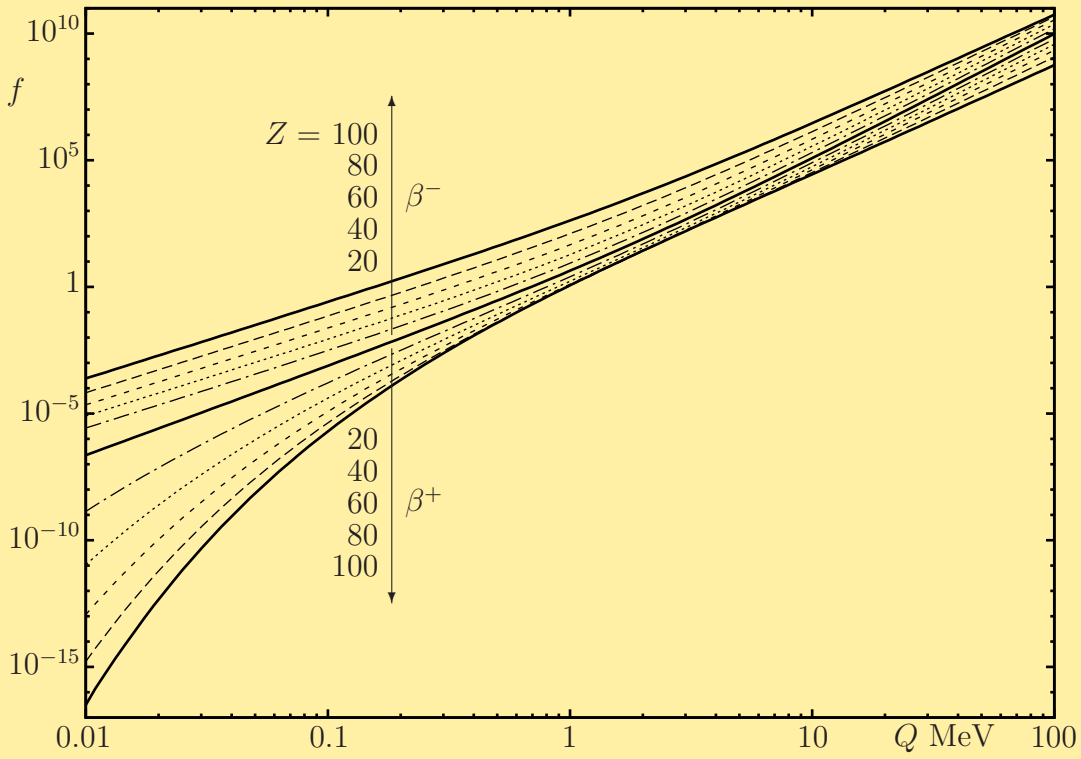


Figure 14.52: The Fermi integral. It shows the effects of energy release and nuclear charge on the beta decay rate of allowed transitions. Other effects exists. [pdf]

There is another, smaller, effect. Even if the final nucleus is the 5^+ excited scandium-48 state, with a single value for the magnetic quantum number, there is still more than one final state to decay to. The reason is that the relative amounts of energy taken by the electron and neutrino can vary. Additionally, their directions of motion can also vary. The actual net decay rate is an integral of the individual decay rates to all these different states. If the Q-value is low, there are relatively few states available, and this reduces the total decay rate too. The amount of reduction is given by the so-called "Fermi integral" shown in figure 14.52. A decay with a Q value of about 0.15 MeV is slowed down by roughly a factor thousand compared to one with a Q value of 1 MeV.

The Fermi integral shows beta plus decay is additionally slowed down, because it is more difficult to create a positron at a strongly repelling positively charged nucleus. The relativistic Fermi integral also depends on the nuclear radius, hence a bit on the mass number. Figure 14.52 used a ballpark value of the mass number for each Z value, {A.45}.

The Fermi integral applies to allowed decays, but the general idea is the same for forbidden decays. In fact, half-lives $\tau_{1/2}$ are commonly multiplied by the Fermi integral f to produce a "comparative half-life," or "ft-value" that is relatively insensitive to the details of the decay besides the degree to which it is forbidden. The ft-value of a given decay can therefore be used to ballpark to what extent the decay is forbidden.

You see how calcium-48 can resist beta-decay for 64 10^{18} years. (Zirconium-96 with a half-life of 24 10^{18} years has similar resistance to plain beta decay.)

14.19.7 Draft: Data and Fermi theory

Figure 14.53 shows nuclei that decay primarily through beta-minus decay in blue. Nuclei that decay primarily through electron capture and beta-plus decay are shown in red. The sizes of the squares indicate the decay rates. For reference, the stable and double-beta decay nuclei are shown as full-size green squares.

Note the tremendous range of decay rates. It corresponds to half-lives ranging from milliseconds to 10^{17} years. This is much like the tremendous range of half-lives in alpha decay. Decays lasting more than about twenty years are shown as a minimum-size dot in figure 14.53; many would be invisible shown on the true scale.

The decay rates in figure 14.53 are color coded according to a guesstimated value for how forbidden the decay is. Darker red or blue indicate more forbidden decays. Note that more forbidden decays tend to have much lower decay rates. (Lightly colored squares indicate nuclei for which the degree to which the decay is forbidden could not be guesstimated by the automated procedures used.)

Figure 14.54 shows the decay rates normalized with a theoretical guesstimate for them. Note the greatly reduced range of variation that the guesstimate achieves, crude as it may be. One major success story is for forbidden decays.

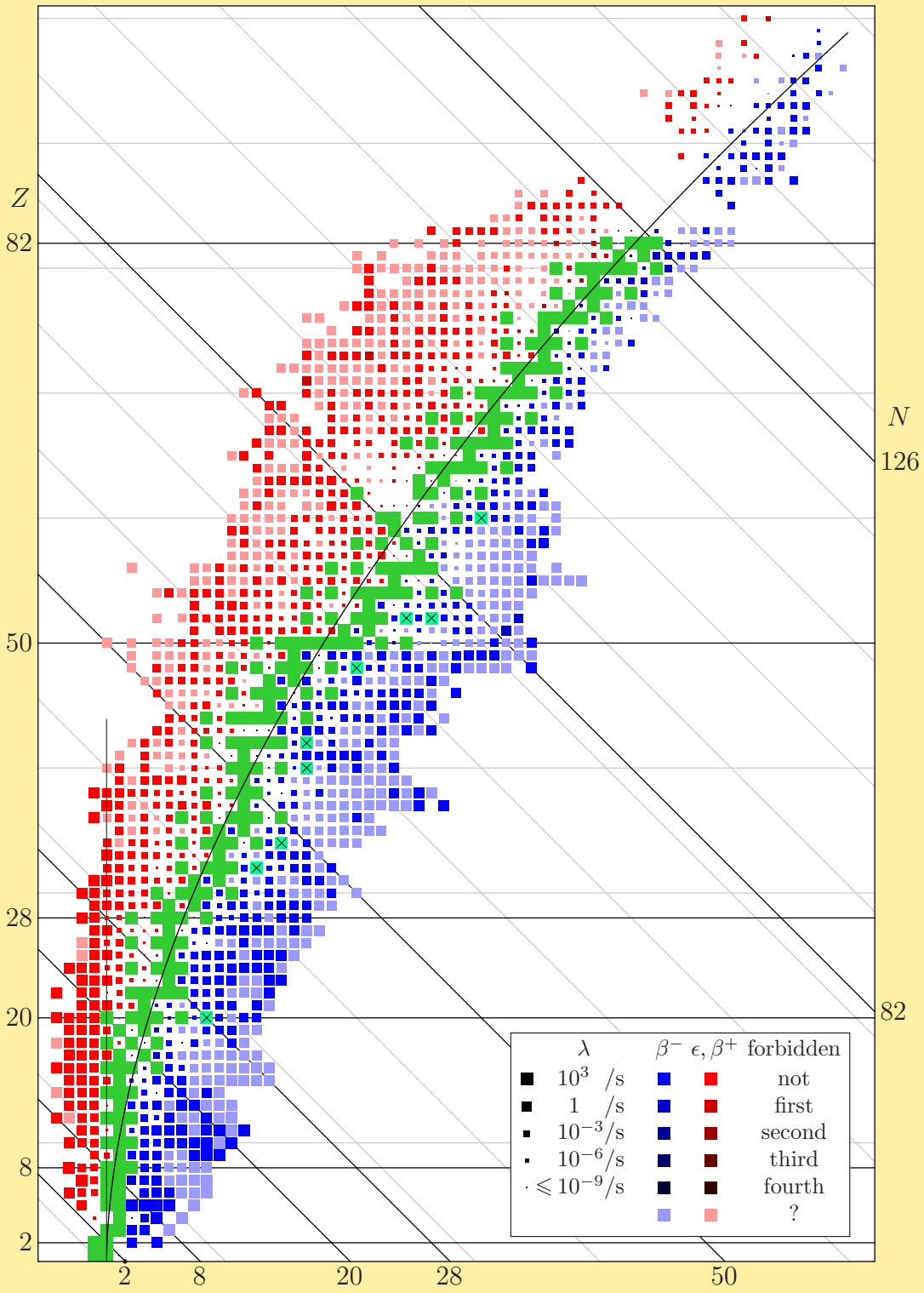


Figure 14.53: Beta decay rates. [pdf][con]

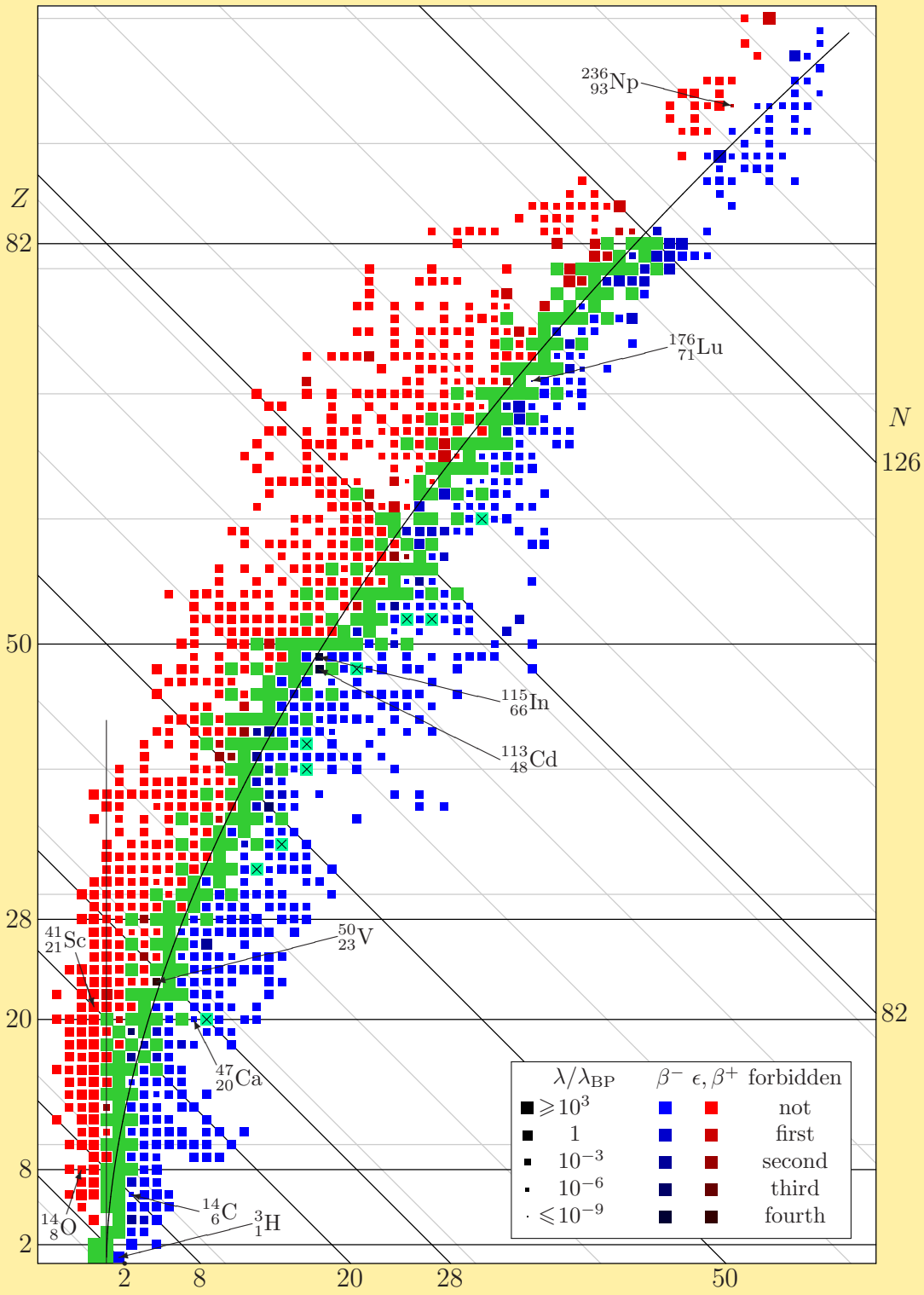


Figure 14.54: Beta decay rates as fraction of a ballparked value. [pdf][con]

These are often so slow that they must be shown as minimum-size dots in figure 14.53 to be visible. However, in figure 14.54 they join the allowed decays as full-size squares. Consider in particular the three slowest decays among the data set. The slowest of all is vanadium-50, with a half-life of 150 10^{15} year, followed by cadmium-113 with 8 10^{15} year, followed by indium-115 with 441 10^{12} year. (Tellurium-123 has only a lower bound on its half life listed and is not included.) These decay times are long enough that all three isotopes occur naturally. In fact, almost all naturally occurring indium is the "unstable" isotope indium-115. Their dots in figure 14.53 become full squares in figure 14.54.

Another eye-catching success story is ${}^{3}_{1}\mathrm{H}$, the triton, which suffers beta decay into ${}^{3}_{2}\mathrm{He}$, the stable helion. The decay is allowed, but because of its miniscule energy release, or Q-value, it takes 12 years anyway. Scaled with the ballpark, this slow decay too becomes a full size square.

The ballparks were obtained from the "Fermi theory" of beta decay, as discussed in detail in addendum {A.45}. Unlike the relatively simple theory of alpha decay, the Fermi theory is elaborate even in a crude form. Taking beta-minus decay as an example, the Fermi theory assumes a pointwise interaction between the wave functions of the neutron that turns into a proton and those of the electron/antineutrino pair produced by the decay. (Quantum mechanics allows the neutron before the decay to interact with the electron and neutrino that would exist if it had already decayed. That is a "twilight" effect, as discussed in chapter 5.3 and more specifically in addendum {A.24} for gamma decay.) The strength of the interaction is given by empirical constants.

Note that for many nuclei no ballparks were found. One major reason is that the primary decay mechanism is not necessarily to the ground state of the final nucleus. If decay to the ground state is forbidden, decay to a less-forbidden excited state may dominate. Therefore, to correctly estimate the decay rate for a given nucleus requires detailed knowledge about the excited energy states of the final nucleus. The energies of these excited states must be sufficiently accurately known, and they may not be. In particular, for a few nuclei, the energy release of the decay, or Q-value, was computed to be negative even for the ground state. This occurred for the electron capture of ${}^{163}_{67}\mathrm{Ho}$, ${}^{193}_{78}\mathrm{Pt}$, ${}^{194}_{80}\mathrm{Hg}$, ${}^{202}_{82}\mathrm{Pb}$, and ${}^{205}_{82}\mathrm{Pb}$, and for the beta decay of ${}^{187}_{75}\mathrm{Re}$ and ${}^{241}_{94}\mathrm{Pu}$. According to the Fermi theory, the decay cannot occur if the Q-value is negative. However, the Q-values in question are much smaller than the estimated electronic binding energy (14.8). In fact they are comparable to the difference in electronic binding energy between initial and final nucleus or less. Since the binding energy is just an estimate, the computed Q-values, and therefore the guesstimated decay rate, should not be trusted.

In addition to the energy of the excited states, their spins and parities must also be accurately known. The reason is that they determine to what level the decay is forbidden, hence slowed down. The computer program that produced figures 14.53 and 14.54 assumed conservatively that if no unique value for spin

and/or parity was given, it might be anything. Also, while there was obviously no way for the program to account for any excited states whose existence is not known, the program did allow for the possibility that there might be additional excited states above the highest energy level known. This is especially important well away from the stable line where the excited data are often sparse or missing altogether. All together, for about one third of the nuclei processed, the uncertainty in the ballparked decay rate was judged too large to be accepted. For the remaining nuclei, the level to which the decay was forbidden was taken from the excited state that gave the largest contribution to the decay rate.

The Fermi ballparks were constructed such that the true decay rate should not be significantly more than the ballparked one. In general they met that requirement, although for about 1% of the nuclei, the true decay rate was more ten times the ballparked ones, reaching up to 370 times for $^{253}_{100}$Fm. All these cases were for first-forbidden decays with relatively low Q-values. Since they included both beta minus and electron capture decays, a plausible explanation may be poor Q-values. However, for forbidden decays, the correction of the electron/positron wave function for the effect of the nuclear charge is also suspect.

Note that while the true decay rate should not be much more than the ballparked one, it is very possible for it to be much less. The ballpark does not consider the details of the nuclear wave function, because that is in general prohibitively difficult. The ballpark simply hopes that if a decay is not strictly forbidden by spin or parity at level l, the nuclear wave function change will not for some other reason make it almost forbidden. But in fact, even if the decay is theoretically possible, the part of the Hamiltonian that gives rise to the decay may produce a nuclear wave function that has little probability of being the right one. In that case the decay is slowed down proportional to that probability.

As an example, compare the decay processes of scandium-41 and calcium-47. Scandium-41, with 21 protons and 20 neutrons, decays into its mirror twin calcium-41, with 20 protons and 21 neutrons. The decay is almost all due to beta-plus decay to the ground state of calcium-41. According to the shell model, the lone proton in the $4f_{7/2}$ proton shell turns into a lone neutron in the $4f_{7/2}$ neutron shell. That means that the nucleon that changes type is already in the right state. The only thing that beta decay has to do is turn it from a proton into a neutron. And that is in fact all that the decay Hamiltonian does in the case of Fermi decay. Gamow-Teller decays also change the spin. The nucleon does not have to be moved around spatially. Decays of this type are called "superallowed." (More generally, superallowed decays are defined as decays between isobaric analog states, or isospin multiplets. Such states differ only in nucleon type. In other words, they differ only in the net isospin component T_3.) Superallowed decays proceed at the maximum rate possible. Indeed the decay of scandium-41 is at 1.6 times the ballparked value.

All the electron capture / beta-plus decays of the nuclei immediately to the

left of the vertical $Z = N$ line in figures 14.53 and 14.54 are between mirror nuclei, and all are superallowed. They are full-size squares in figure 14.54. Superallowed beta-minus decays occur for the triton mentioned earlier, as well as for a lone neutron.

But now consider the beta-minus decay process of calcium-47 to scandium-47. Calcium-47 has no protons in the $4f_{7/2}$ proton shell, but it has 7 neutrons in the $4f_{7/2}$ neutron shell. That means that it has a 1-neutron "hole" in the $4f_{7/2}$ neutron shell. Beta decay to scandium-47 will turn one of the 7 neutrons into a lone proton in the $4f_{7/2}$ proton shell.

At least one source claims that in the odd-particle shell model "all odd particles are treated equivalently," so that we might expect that the calcium-47 decay is superallowed just like the scandium-41 one. That is of course not true. The odd-particle shell model does emphatically not treat all odd particles equivalently. It only says that, effectively, an even number of nucleons in the shell pair up into a state of zero net spin, leaving the odd particle to provide the net spin and electromagnetic moments. That does not mean that the seventh $4f_{7/2}$ neutron can be in the same state as the lone proton after the decay. In fact, if the seventh neutron was in the same state as the lone proton, it would blatantly violate the antisymmetrization requirements, chapter 5.7. Whatever the state of the lone proton might be, 7 neutrons require 6 more independent states. And each of the 7 neutrons must occupy all these 7 states equally. It shows. The nuclear wave function of calcium-47 produced by the decay Hamiltonian matches up very poorly with the correct final wave function of scandium-47. The true decay rate of calcium-47 is therefore about 10 000 times smaller than the ballpark.

As another example, consider the beta-plus decay of oxygen-14 to nitrogen-14. Their isobaric analog states were identified in figure 14.46. Decay to the ground state is allowed by spin and parity, at a ballparked decay rate of 0.23/s. However, the true decay proceeds at a rate 0.01/s, which just happens to be 1.6 times the ballparked decay rate to the 0^+ *excited* isobaric analog state. One source notes additionally that over 99% of the decay is to the analog state. So decay to the ground state must be contributing less than a percent to the total decay. And that is despite the fact that decay to the ground state is allowed too and has the greater Q-value. The effect gets even clearer if you look at the carbon-14 to nitrogen-14 beta-minus decay. Here the decay to the isobaric analog state violates energy conservation. The decay to the ground state is allowed, but it is more than 10 000 times slower than ballpark.

Superallowed decays like the one of oxygen-14 to the corresponding isobaric analog state of nitrogen-14 are particularly interesting because they are 0^+ to 0^+ decays. Such decays cannot occur through the Gamow-Teller mechanism, because in Gamow-Teller decays the electron and neutrino take away one unit of angular momentum. That means that decays of this type can be used to study the Fermi mechanism in isolation.

The horror story of a poor match up between the nuclear wave function produced by the decay Hamiltonian and the final nuclear wave function is lutetium-176. Lutetium-176 has a 7^- ground state, and that solidly forbids decay to the 0^+ hafnium-176 ground state. However, hafnium has energetically allowed 6^+ and 8^+ excited states that are only first-forbidden. Therefore you would not really expect the decay of lutetium-176 to be particularly slow. But the spin of the excited states of hafnium is due to collective nuclear rotation, and these states match up extremely poorly with the ground state of lutetium-176 in which the spin is intrinsic. The decay rate is a stunning 12 orders of magnitude slower than ballpark. While technically the decay is only first-forbidden, lutetium is among the slowest decaying unstable nuclei, with a half-life of almost 40 10^{12} year. As a result, it occurs in significant quantities naturally. It is commonly used to determine the age of meteorites. No other ground state nucleus gets anywhere close to that much below ballpark. The runner up is neptunium-236, which is 8 orders of magnitude below ballpark. Its circumstances are similar to those of lutetium-176.

The discussed examples show that the Fermi theory does an excellent job of predicting decay rates if the differences in nuclear wave functions are taken into account. In fact, if the nuclear wave function can be accurately accounted for, like in 0^+ to 0^+ superallowed decays, the theory will produce decay rates to 3 digits accurate, [31, table 9.2]. The theory is also able to give accurate predictions for the distribution of velocities with which the electrons or positrons come out. Data on the velocity distributions can in fact be used to solidly determine the level to which the decay is forbidden by plotting them in so-called "Fermi-Kurie plots." These and many other details are outside the scope of this book.

14.19.8 Draft: Parity violation

For a long time, physicists believed that the fundamental laws of nature behaved the same when seen in the mirror. The strong nuclear force, electromagnetism, and gravity all do behave the same when seen in the mirror. However, in 1956 Lee and Yang pointed out that the claim had not been tested for the weak force. If it was untrue there, it could explain why what seemed to be a single type of K-meson could decay into end products of different parity. The symmetry of nature under mirroring leads to the law of conservation of parity, chapter 7.3. However, if the weak force is not the same under mirroring, parity can change in weak processes, and therefore, the decay products could have any net parity, not just that of the original K-meson.

Wu and her coworkers therefore tested parity conservation for the beta decay of cobalt-60 nuclei. These nuclei were cooled down to extremely low temperatures to cut down on their thermal motion. That allowed their spins to be aligned with a magnetic field, as in the left of figure 14.55. It was then observed

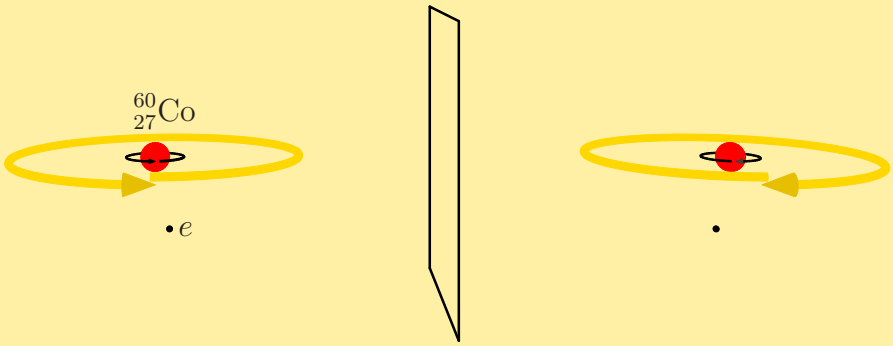


Figure 14.55: Parity violation. In the beta decay of cobalt-60, left, the electron preferentially comes out in the direction that a left-handed screw rotating with the nuclear spin would move. Seen in the mirror, right, that becomes the direction of a right-handed screw.

that the electrons preferentially came out in the direction of motion of a left-handed screw rotating with the nuclear spin. Since a left-handed screw turns into a right-handed one seen in the mirror, it followed that indeed the weak force is not the same seen in the mirror. The physics in the mirror is not the correct physics that is observed.

Since the weak force is weak, this does not affect parity conservation in other circumstances too much. Formally it means that eigenfunctions of the Hamiltonian are not eigenfunctions of the parity operator. However, nuclear wave functions still have a single parity to very good approximation; the amplitude of the state of opposite parity mixed in is of the order of 10^{-7}, [31, p. 313]. The probability of measuring the opposite parity is the square of that, much smaller still. Still, if a decay is absolutely forbidden when parity is strictly preserved, then it might barely be possible to observe the rare decays allowed by the component of the wave function of opposite parity.

An additional operation can be applied to the mirror image in 14.55 to turn it back into a physically correct decay. All particles can be replaced by their antiparticles. This operation is called "charge conjugation," because among other things it changes the sign of the charge for each charged particle. In physics, you are always lucky if a name gets some of it right. Some of the particles involved may actually be charged, and "conjugation" is a sophisticated-sounding term to some people. It is also a vague term that quite conceivably could be taken to mean "reversal of sign" by people naive enough to consider "conjugation" sophisticated. Charge conjugation turns the electrons going around in the loops of the electromagnet in figure 14.55 into positrons, so the current reverses direction. That must reverse the sign of the magnetic field if the physics is right. But so will the magnetic moment of anticobalt-60 nucleus change sign, so it stays aligned with the magnetic field. And physicist believe the positrons will preferentially come out of anticobalt-60 nuclei along the motion of a right-handed

screw.

Besides this combined charge conjugation plus parity (CP) symmetry of nature, time symmetry is also of interest here. Physical processes should remain physically correct when run backwards in time, the same way you can run a movie backwards. It turns out that time symmetry too is not completely absolute, and neither is CP symmetry for that matter. However, if all three operations, charge conjugation (C), mirroring (P), and time inversion (T), together are applied to a physical process, the resulting process is believed to always be physically correct. There is a theorem, the CPT theorem, that says so under relatively mild assumptions.

14.20 Draft: Gamma Decay

Nuclear reactions and decays often leave the final nucleus in an quantum state of elevated energy. Such an excited state may lower its energy by emitting a photon of electromagnetic radiation. That is called gamma decay. It is a common way to evolve to the ground state.

Gamma decay is in many respect similar to alpha and beta decay discussed in earlier sections. However, the type of nucleus does not change in gamma decay. Both the atomic and mass number of the nucleus stay the same. (Of course, an excited nuclear state can suffer alpha or beta decay instead of gamma decay. That however is not the subject of this section.)

Gamma decay of excited nuclei is the direct equivalent of the decay of excited electron states in atoms. The big difference between gamma decay and the radiation emitted by the electrons in atoms is energy. The energy of the photons emitted by nuclei is typically even higher than that of the X-ray photons emitted by inner electrons. Therefore the radiation emitted by nuclei is generally referred to as "gamma rays."

Both atomic radiation and nuclear gamma decay were analyzed in considerable detail in chapter 7.4 through 7.8 and addenda {A.20} through {A.25}. There is no point in repeating all that here. Instead this section will merely summarize the key points and discuss some actual observations.

However, the existing data on gamma decay is enormous. Consider NuDat 2, a standard data base. At the time of writing, it contains over 3 100 nuclei. Almost every nucleus but the deuteron has many excited energy levels; there are over 160 000 in NuDat 2. Gamma decays can proceed between different states, and NuDat 2 contains over 240 000 of them. There is no way that this book can cover all that data. The coverage given in this section will therefore be anecdotal or random rather than comprehensive.

However, based on a simple model, at least ballpark transition rates will be established. These are called the Weisskopf units. They are commonly used as reference values, to give some context to the measured transition rates.

One big limitation of gamma decay is for nuclear states of zero spin. A state of zero spin cannot transition to another state of zero spin by emitting a photon. As discussed in chapter 7.4, this violates conservation of angular momentum.

But there are other ways that a nucleus can get rid of excess energy besides emitting an electromagnetic photon. One way is by kicking an atomic electron out of the surrounding atom. This process is called "internal conversion" because the electron is outside the nucleus. It allows transitions between states of zero spin.

For atoms, two-photon emission is a common way to achieve decays between states of zero angular momentum. However, for nuclei this process is less important because internal conversion usually works so well.

Internal conversion is also important for other transitions. Gamma decay is slow between states that have little difference in energy and/or a big difference in spin. For such decays, internal conversion can provide a faster alternative.

If the excitation energy is high, it is also possible for the nucleus to create an electron and positron pair from scratch. Since the quantum uncertainty in position of the pair is far too large for them to be confined within the small nucleus, this is called "internal pair creation."

14.20.1 Draft: Energetics

The reduction in nuclear energy during gamma decay is called the Q-value. This energy comes out primarily as the energy of the photon, though the nucleus will also pick up a bit of kinetic energy, called the recoil energy.

Recoil energy will usually be ignored, so that Q gives the energy of the photon. The photon energy is related to its momentum and frequency through the relativistic mass-energy and Planck-Einstein relations:

$$Q = E_{\mathrm{N1}} - E_{\mathrm{N2}} = pc = \hbar\omega \tag{14.65}$$

Typical tabulations list nuclear excitation energies as energies, rather than as nuclear masses. Unfortunately, the energies are usually in eV instead of SI units.

In internal conversion, the nucleus does not emit a photon, but kicks an electron out of the surrounding atomic electron cloud. The nuclear energy reduction goes into kinetic energy of the electron, plus the binding energy required to remove the electron from its orbit:

$$Q = E_{\mathrm{N1}} - E_{\mathrm{N2}} = T_{\mathrm{e}} + E_{\mathrm{B,e}} \tag{14.66}$$

14.20.2 Draft: Forbidden decays

The decay rate in gamma decay is to a large extent dictated by what is allowed by conservation of angular momentum and parity. The nucleus is almost a mathematical point compared to the wave length of a typical photon emitted

in gamma decay. Therefore, it is difficult for the nucleus to give the photon additional orbital angular momentum. That is much like what happens in alpha and beta decay.

The photon has one unit of spin. If the nucleus does not give it additional orbital angular momentum, the total angular momentum that the photon carries off is one unit. That means that the nuclear spin cannot change by more than one unit.

(While this is true, the issue is actually somewhat more subtle than in the decay types discussed previously. For a photon, spin and orbital angular momentum are intrinsically linked. Because of that, a photon always has some orbital angular momentum. That was discussed in chapter 7.4.3 and in detail in various addenda such as {A.21}. However, the inherent orbital angular momentum does not really change the story. The bottom line remains that it is unlikely for the photon to be emitted with more than one unit of net angular momentum.)

The nuclear spin can also stay the same, instead of change by one unit, even if a photon with one unit of angular momentum is emitted, In classical terms the one unit of angular momentum can go into changing the direction of the nuclear spin instead of its magnitude, chapter 7.4.2. However, this only works if the nuclear spin is nonzero.

Parity must also be preserved, chapter 7.4. Parity is even, or 1, if the wave function stays the same when the positive direction of all three Cartesian axes is reversed. Parity is odd, or -1, if the wave function changes sign. Parities of separate sources are multiplied together to combine them. That is unlike for angular momentum, where separate angular momenta are added together.

In the normal, or "allowed," decays the photon is emitted with odd parity. Therefore, the nuclear parity must reverse during the transition, chapter 7.4.2.

(To be picky, the so-called weak force does not preserve parity. This creates a very small uncertainty in nuclear parities. That then allows a very small probability for transitions in which the apparent parity is not conserved. But the probability for this is so small that it can almost always be ignored.)

Allowed transitions are called electric transitions because the nucleus interacts mainly with the electric field of the photon. More specifically, they are called "electric dipole transitions" for reasons originating in classical electromagnetics, chapter 7.7.2. For practical purposes, a dipole transition is one in which the photon is emitted with one unit of net angular momentum.

Transitions in which the nuclear spin change is greater than one unit, or in which the nuclear parity does not change, or in which the spin stays zero, are called "forbidden." Despite the name, most such decays will usually occur given enough time. However they are generally much slower.

One way that forbidden transitions can occur is that the nucleus interacts with the magnetic field instead of the electric field. This produces what are called magnetic transitions. Magnetic transitions tend to be noticeably slower

than corresponding electric ones. In magnetic dipole transitions, the photon has one unit of net angular momentum just like in electric dipole transitions. However, the photon now has even parity. Therefore magnetic dipole transitions allow the nuclear parity to stay the same.

Transitions in which the nuclear spin changes by more than one unit are possible through emission of a photon with additional orbital angular momentum. That allows a net angular momentum of the photon greater than one. But at a price. Each unit of additional net angular momentum slows down the typical decay rate by roughly 5 orders of magnitude.

The horror story is tantalum-180m. There are at the time of writing 256 ground state nuclei that are classified as stable. And then there is the excited nucleus tantalum-180m. Stable nuclei should be in their ground state, because states of higher energy decay into lower energy ones. But tantalum-180m has never been observed to decay. If it decays at all, it has been established that its half life cannot be less than 10^{15} year. The universe has only existed for roughly 10^{10} years, and so tantalum-180m occurs naturally.

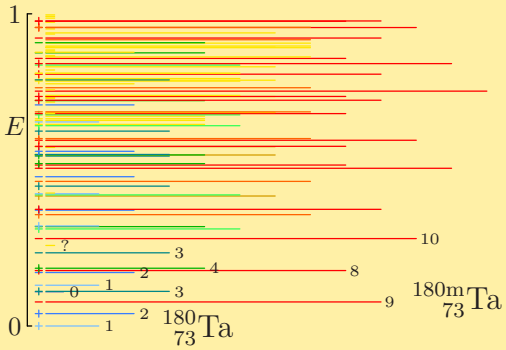


Figure 14.56: Energy levels of tantalum-180. [pdf]

The tantalum-180 ground state shows no such idiocy. It is unstable as any self-respecting heavy odd-odd nucleus should be. In fact it disintegrates within about 8 hours through both electron capture and beta-minus decay at comparable rates. But tantalum-180m is an excited state with a humongous spin of 9. Figure 14.56 shows the excited energy levels of tantalum-180; tantalum-180m is the second excited energy level. It can only decay to the 1^+ ground state and to an 2^+ excited state. It has very little energy available to do either. The decay would require the emission of a photon with at least seven units of orbital angular momentum, and that just does not happen in a thousand years. Nor in a petayear.

You might think that tantalum-180m could just disintegrate directly through electron capture or beta decay. But those processes have the same problem. There is just no way for tantalum-180m to get rid of all that spin without emitting particles with unlikely large orbital angular momentum. So tantalum-

180m will live forever, spinning too fast to reach the sweet oblivion of the quick death that waits below.

Electric transitions are often generically indicated as Eℓ and magnetic ones as Mℓ. Here ℓ indicates the net angular momentum of the photon. That is the maximum nuclear spin change that the transition can achieve. So electric dipole transitions are E1, and magnetic dipole transitions are M1. Names have also been given to the higher multipole orders. For example, $\ell = 2$ transitions are quadrupole ones, $\ell = 3$ octupole, $\ell = 4$ hexadecapole, $\ell = 5$ triakontadipole, $\ell = 6$ hexacontatetrapole, etcetera. (If you are wondering, the prefixes in these names are powers of two, expressed in a mixture of Latin and Greek.)

For electric transitions, the nuclear parity changes when ℓ is odd. For magnetic transitions, it changes when ℓ is even. The transition rules are summarized in table 14.4.

	maximum spin change	nuclear parity change
Eℓ:	ℓ	if ℓ is odd
Mℓ:	ℓ	if ℓ is even
No spin 0 to spin 0 transitions.		

Table 14.4: Nuclear spin and parity changes in electromagnetic multipole transitions.

That leaves transitions from nuclear spin 0 to nuclear spin 0. Such transitions cannot occur through emission of a photon, period. For such transitions, conservation of angular momentum would require that the photon is emitted without angular momentum. But a photon cannot have zero net angular momentum. You might think that the spin of the photon could be canceled through one unit of orbital angular momentum. However, because the spin and orbital angular momentum of a photon are linked, it turns out that this is not possible, {A.21}.

Decay from an excited state with spin zero to another state that also has spin zero is possible through internal conversion or internal pair production. In principle, it could also be achieved through two-photon emission, but that is a very slow process that has trouble competing with the other two.

One other approximate conservation law might be mentioned here, isospin. Isospin is conserved by nuclear forces, and its charge component is conserved by electromagnetic forces, section 14.18. It can be shown that to the extent that isospin is conserved, certain additional selection rules apply. These involve the quantum number of square isospin t_T, which is the isospin equivalent of

the azimuthal quantum number for the spin angular momentum of systems of fermions. Warburton & Weneser [48] give the following rules:

1. Electromagnetic transitions are forbidden unless $\Delta t_T = 0$, or ± 1. (Here "Δ" means the difference between the initial and final nuclear states).
2. Corresponding $\Delta t_T = \pm 1$ transitions in conjugate nuclei are identical in all properties. ("Conjugate" means that the two nuclei have the numbers of protons and neutrons swapped. "Corresponding" transitions means transitions between equivalent levels, as discussed in section 14.18.)
3. Corresponding E1 transitions in conjugate nuclei — whether $\Delta t_T = 0$ or ± 1 — have equal strengths.
4. $\Delta t_T = 0$ E1 transitions in self-conjugate nuclei are forbidden. (Self-conjugate nuclei have the same number of protons as neutrons.)
5. Corresponding $\Delta t_T = 0$ M1 transitions in conjugate nuclei are expected to be of approximately equal strength, within, say, a factor of two if the transitions are of average strength or stronger.
6. $\Delta t_T = 0$ M1 transitions in self-conjugate nuclei are expected to be weaker by a factor of 100 than the average M1 transition strength.
7. $\Delta t_T = 0$ Mℓ transitions in conjugate nuclei are expected to be of approximately equal strength if the transitions are of average strength or stronger.
8. $\Delta t_T = 0$ Mℓ transitions in self-conjugate nuclei are expected to be appreciably weaker than average.

The last four rules involve an additional approximation besides the assumption that isospin is conserved.

In a nutshell, expect that the transitions will be unexpectedly slow if the isospin changes by more than one unit. Expect the same for nuclei with equal numbers of protons and neutrons if the isospin does not change at all and it is an E1 or magnetic transition.

As an example, [31, p. 390], consider the decay of the 1^- isobaric analog state common to carbon-14, nitrogen-14, and oxygen-14 in figure 14.46. This state has $t_T = 1$. For oxygen-14, it is the lowest excited state. Its decay to the $t_T = 1$, 0^+, ground state is an E1 transition that is allowed by the spin, parity, and isospin selection rules. And indeed, the 1^- excited state decays rapidly to the ground state; the half-life is about 0.000 012 fs (femtoseconds). That is even faster than the Weisskopf ballpark for a fully allowed decay, subsection 14.20.4, which gives about 0.009 fs here. But for nitrogen-14, the equivalent transition is not allowed because of rule 4 above. Nitrogen-14 has 7 protons and 7 neutrons. And indeed, the partial half life of this transition is 2.7 fs. That is very much longer. Based on rule 3 above, it is expected that the decay rate of the 1^- state in carbon-14 is similar to the one in oxygen-14. Unfortunately, experimentally it has only been established that its half-life is less than 7 fs.

Some disclaimers are appropriate for this example. As far as the oxygen transition is concerned, the NuDat 2 [[12]] data *do not* say what the dominant decay process for the oxygen-14 state is. Nor what the final state is. So it might be another decay process that dominates. The next higher excited state, with 0.75 MeV more energy, decays 100% through proton emission. And two orders of magnitude faster than Weisskopf does seem a lot, figure 14.63.

As far as the nitrogen transition is concerned, the decay processes are listed in NuDat 2. The decay is almost totally due to proton emission, not gamma decay. The actual half-life of this state is 0.000 02 fs; the 2.7 fs mentioned above is computed using the given decay branching ratios. The 2.7 fs is way above the 0.006 fs Weisskopf estimate, but that is quite common for E1 transitions.

It may be more reasonable to compare the forbidden nitrogen 8 MeV to 2.3 MeV transition to the allowed 8 MeV to ground state, and 8 MeV to 4 MeV transitions. They are all three E1 transitions. Corrected for the differences in energy release, the forbidden transition is 20 times slower than the one to the ground state, and 25 times slower than the one to the 4 MeV state. So apparently, being forbidden seems to slow down this transition by a factor of roughly 20. It is significant, though it is not that big on the scale of figure 14.63.

As another example, the nitrogen transition from the 1^- 5.7 MeV $t_T = 0$ state to the $t_T = 0$ ground state is also forbidden, while the transition to the 0^+ $t_T = 1$ state is now permitted. And indeed, the decay to the ground state is about ten times slower, when corrected for energy release, [31, p. 391].

More comprehensive data may be found in [48].

14.20.3 Draft: Isomers

An "isomer" is a long lasting excited state of a nucleus. Usually, an excited nucleus that does not disintegrate through other means will drop down to lower energies through the emission of photons in the gamma ray range. It will then end up back in the ground state within a typical time in terms of fs, or about 10^{-15} second.

But sometimes a nucleus gets stuck in a metastable state that takes far longer to decay. Such a state is called an isomeric state. Krane [31, p. 174] ballparks the minimum lifetime to be considered a true isomeric state at roughly 10^{-9} s, Bertulani [5, p. 244] gives 10^{-15} s, and NuDat 2 [[12]] uses 10^{-1} s with qualification in their policies and 10^{-9} s in their glossary. Don't you love standardization? In any case, this book will not take isomers serious unless they have a lifetime comparable to 10^{-9} second. Why would an excited state that cannot survive for a millisecond be given the same respect as tantalum-180m, which shows no sign of kicking the bucket after 10^{15} years?

But then, why would any excited state be able to last very much more than the typical 10^{-15} s gamma decay time in the first place? The main reason is angular momentum conservation. It is very difficult for a tiny object like

a nucleus to give a photon much angular momentum. Therefore, transitions between states of very different angular momentum will be extremely slow, if they occur at all. Such transitions are highly "forbidden," or using a better term, "hindered."

If an excited state has a very different spin than the ground state, and there are no states in between the two that are more compatible, then that excited state is stuck. But why would low spin states be right next to high spin states? The main reason is found in the shell model, and in particular figure 14.15. According to the shell model, just below the magic numbers of 50, 82, and 126, high spin states are pushed into regions of low spin states by the so-called spin-orbit interaction. That is a recipe for isomerism if there ever was one.

Therefore, it should be expected that there will be many isomers below the magic numbers of 50, 82, and 126. And that these isomers will have the opposite parity of the ground state, because the high spin states are pushed into low spin states of opposite parity.

And so it is. Figure 14.57 shows the half-lifes of the longest-lasting exited states of even Z and odd N nuclei. The groups of isomers below the magic neutron numbers are called the "islands of isomerism." The difference in spin from the ground state is indicated by the color. A difference in parity is indicated by a minus sign. Half-lives over 10^{14} s are shown as full-size squares.

Figure 14.58 shows the islands for odd Z, even N nuclei.

For odd-odd nuclei, figure 14.59, the effects of proton and neutron magic numbers get mixed up. Proton and neutron excitations may combine into larger spin changes, providing one possible explanation for the isomers of light nuclei without parity change.

Besides tantalum-180m, which lives forever, also note bismuth-210m in figure 14.59. Bismuth-210m has the same spin 9^- as tantalum-180m, but it does manage to decay after about 3 million years. But it does so through alpha-decay, rather than gamma-decay,

For even-even nuclei, figure 14.60, there is very little isomeric activity.

14.20.4 Draft: Weisskopf estimates

Gamma decay rates can be ballparked using the so-called "Weisskopf estimates:"

$$\boxed{\lambda^{\mathrm{E}\ell} = C_{\mathrm{E}\ell} A^{2\ell/3} Q^{2\ell+1} \qquad \lambda^{\mathrm{M}\ell} = C_{\mathrm{M}\ell} A^{(2\ell-2)/3} Q^{2\ell+1}} \tag{14.67}$$

ℓ :	1	2	3	4	5
$C_{\mathrm{E}\ell}$:	$1.0\ 10^{14}$	$7.3\ 10^{7}$	34	$1.1\ 10^{-5}$	$2.4\ 10^{-12}$
$C_{\mathrm{M}\ell}$:	$3.1\ 10^{13}$	$2.2\ 10^{7}$	10	$3.3\ 10^{-6}$	$7.4\ 10^{-13}$

Here the decay rates are per second, A is the mass number, and Q is the energy release of the decay in MeV. Also ℓ is the maximum nuclear spin change possible for that transition. As discussed in subsection 14.20.2, electric transitions

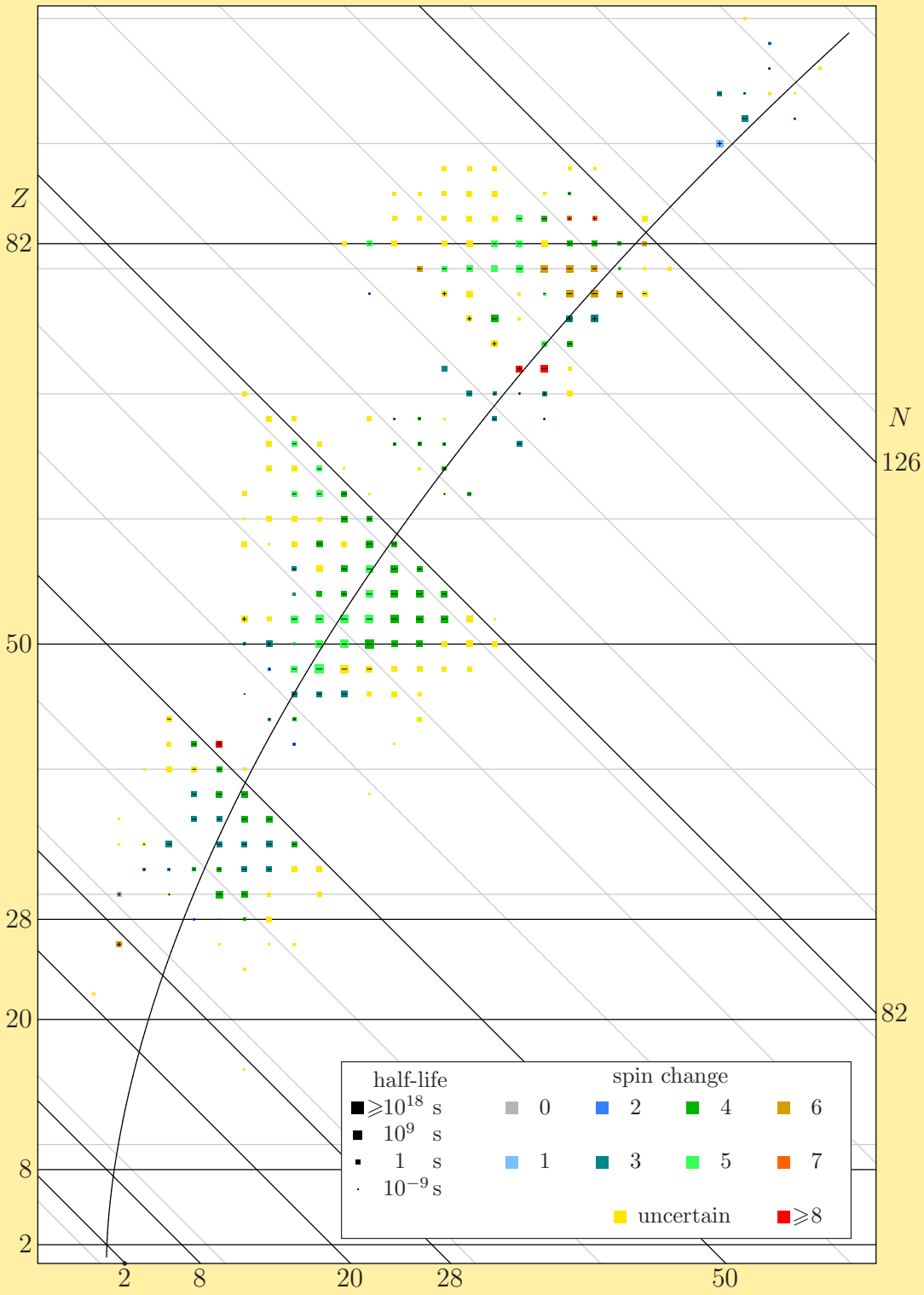


Figure 14.57: Half-life of the longest-lived even-odd isomers. [pdf][con]

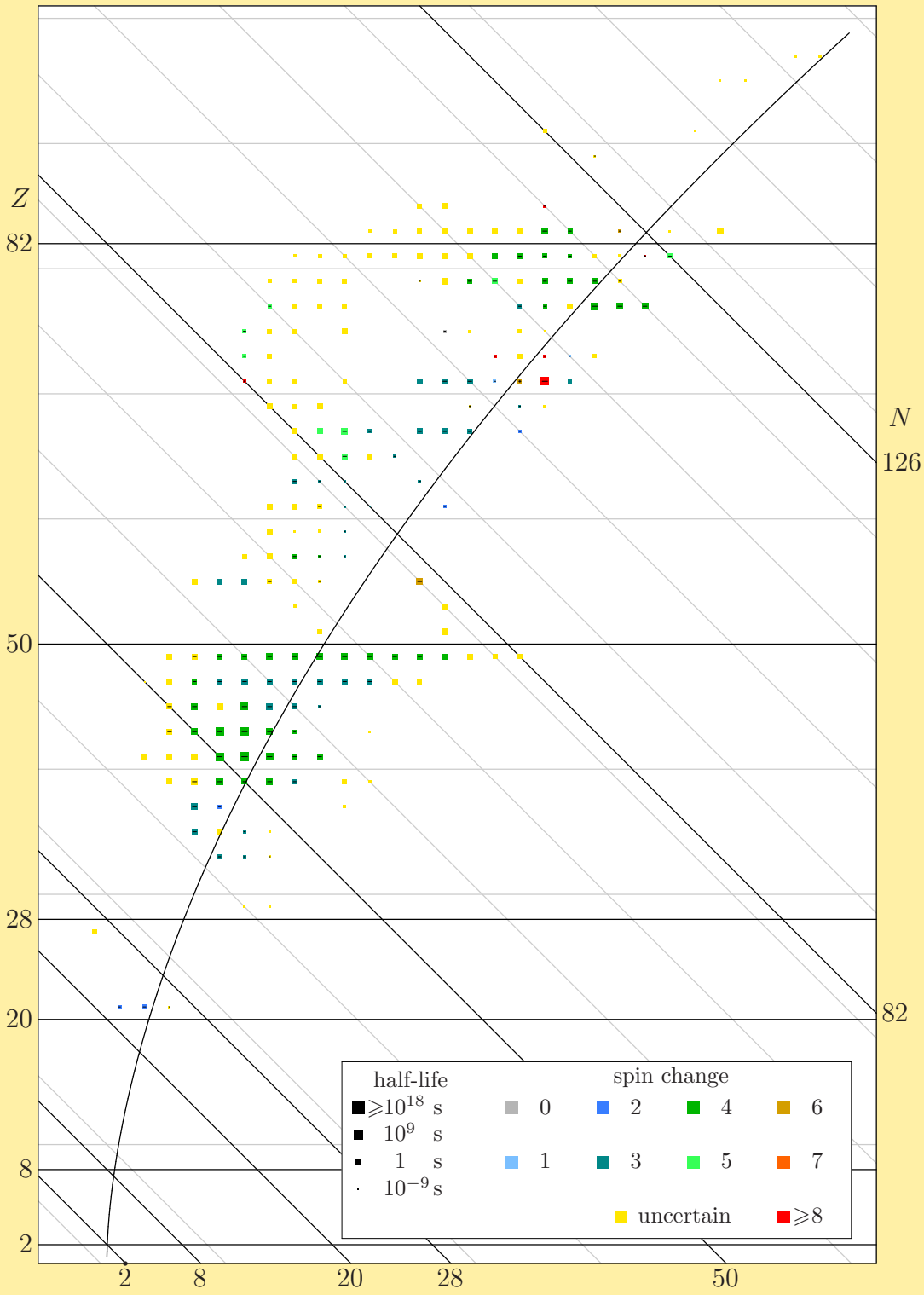


Figure 14.58: Half-life of the longest-lived odd-even isomers. [pdf][con]

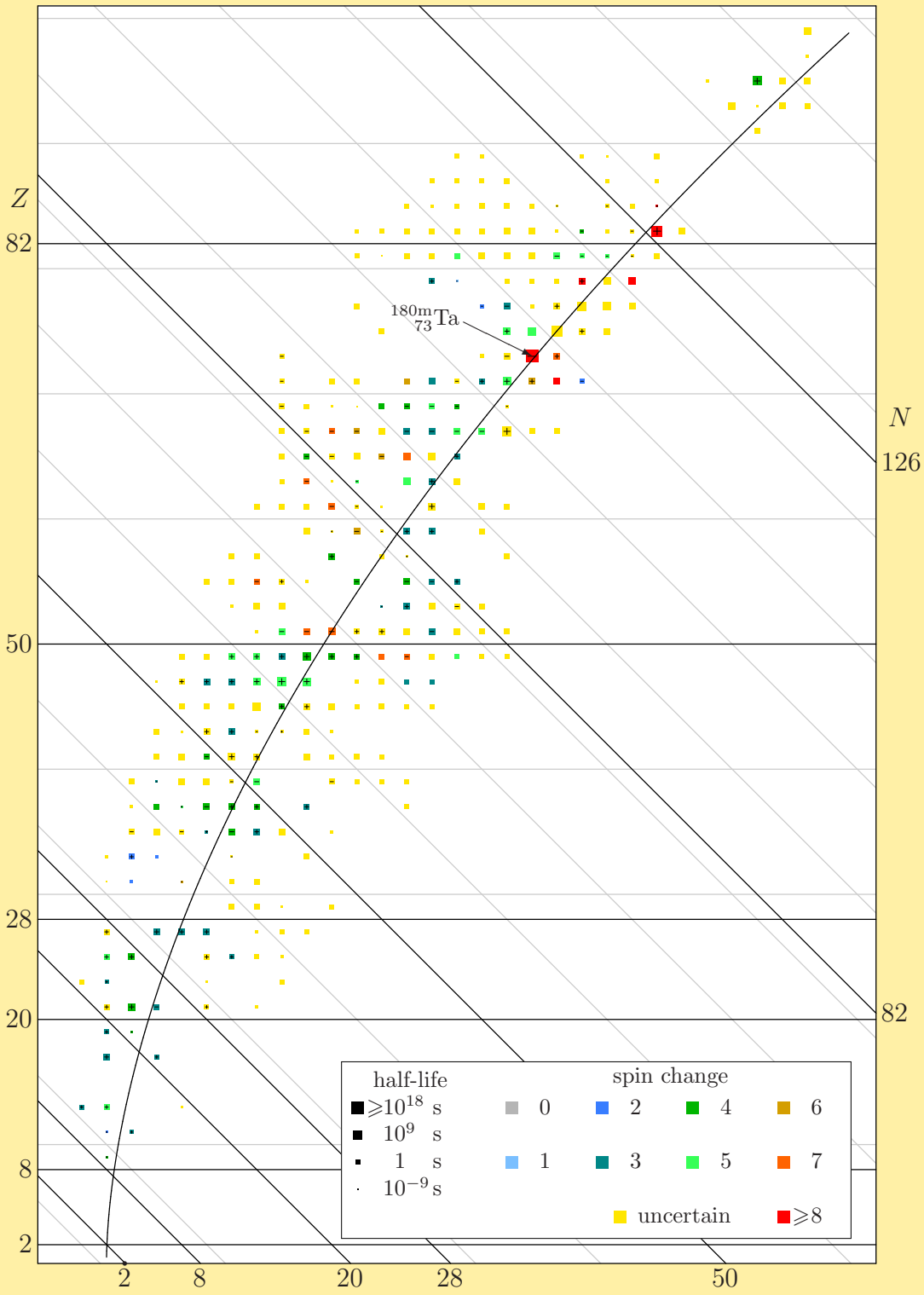


Figure 14.59: Half-life of the longest-lived odd-odd isomers. [pdf][con]

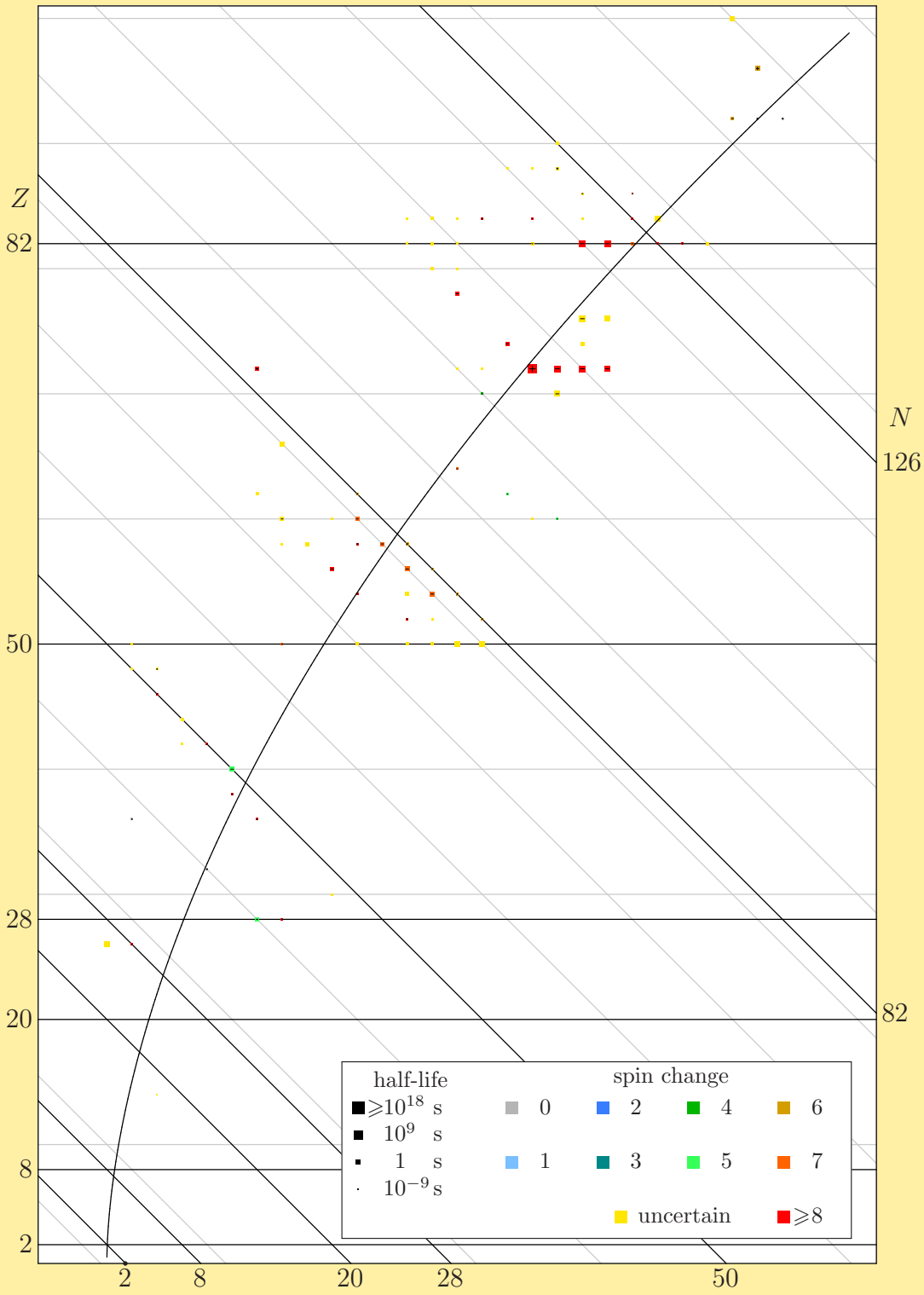


Figure 14.60: Half-life of the longest-lived even-even isomers. [pdf][con]

require that the nuclear parity flips over when ℓ is odd, and magnetic ones that it flips over when ℓ is even. In the opposite cases, the nuclear parity must stay the same. If there is more than one decay process involved, add the individual decay rates.

The estimates are plotted in figure 14.61. For magnetic transitions the better Moszkowski estimates are shown in figure 14.62. (Internal conversion is discussed in subsection 14.20.6, where the ballparks are given.)

A complete derivation and discussion of these estimates can be found {A.25}. Note that many sources have errors in their formulae and/or graphs or use non-SI units, {A.25.9}. The correct formulae in SI units are in {A.25.8}.

These estimates are derived under the assumption that only a single proton changes states in the transition. They also assume that the multipole order is the lowest possible, given by the change in nuclear spin. And that the final state of the proton has angular momentum $\frac{1}{2}$. Some correction factors are available to allow for different multipole orders and different final angular momenta, {A.25.8}. There are also correction factors to allow for the fact that really the proton and the rest of the nucleus move around their common center of gravity. Similar correction factors can allow for the case that a single neutron instead of a proton makes the transition. See {A.25.8} for more.

The initial and final proton states assumed in the estimates are further very simple, {A.25.8}. They are like a shell model state with a simplified radial dependence. Corrections exist for that radial dependence too. But the way the estimates are mostly used in practice is as a reference. The actual decay rate in a transition is compared to the Weisskopf or Moszkowski estimates. These estimates are therefore used as "units" to express decay rates in.

If there is a big difference, it gives hints about the nature of the transition process. For example, the actual decay rate is often orders of magnitude smaller than the estimate. That can indicate that the state produced by the decay Hamiltonian has only a small probability of being the correct final nuclear state. In other words, there may be a "poor match-up" or "little overlap" between the initial and final nuclear states. It is implicit in the simple proton states used in the estimates that the state produced by the decay Hamiltonian has a good chance of being right. But actual E1 transitions can easily be three or more orders of magnitude slower than estimate, as shown in the next subsection. That is similar to what was observed for the ballparks for beta decays given in section 14.19.7. One reason may be that some of these transitions are approximately forbidden by isospin conservation.

Conversely, the observed transition rate may be several orders of magnitude more rapid than the estimate. That may indicate that a lot of nucleons are involved in the transition. Their contributions can add up. This is frequently related to shape changes in nonspherical nuclei. For example, E2 transitions, which are particularly relevant to deformed nuclei, may easily be orders of magnitude faster than estimate.

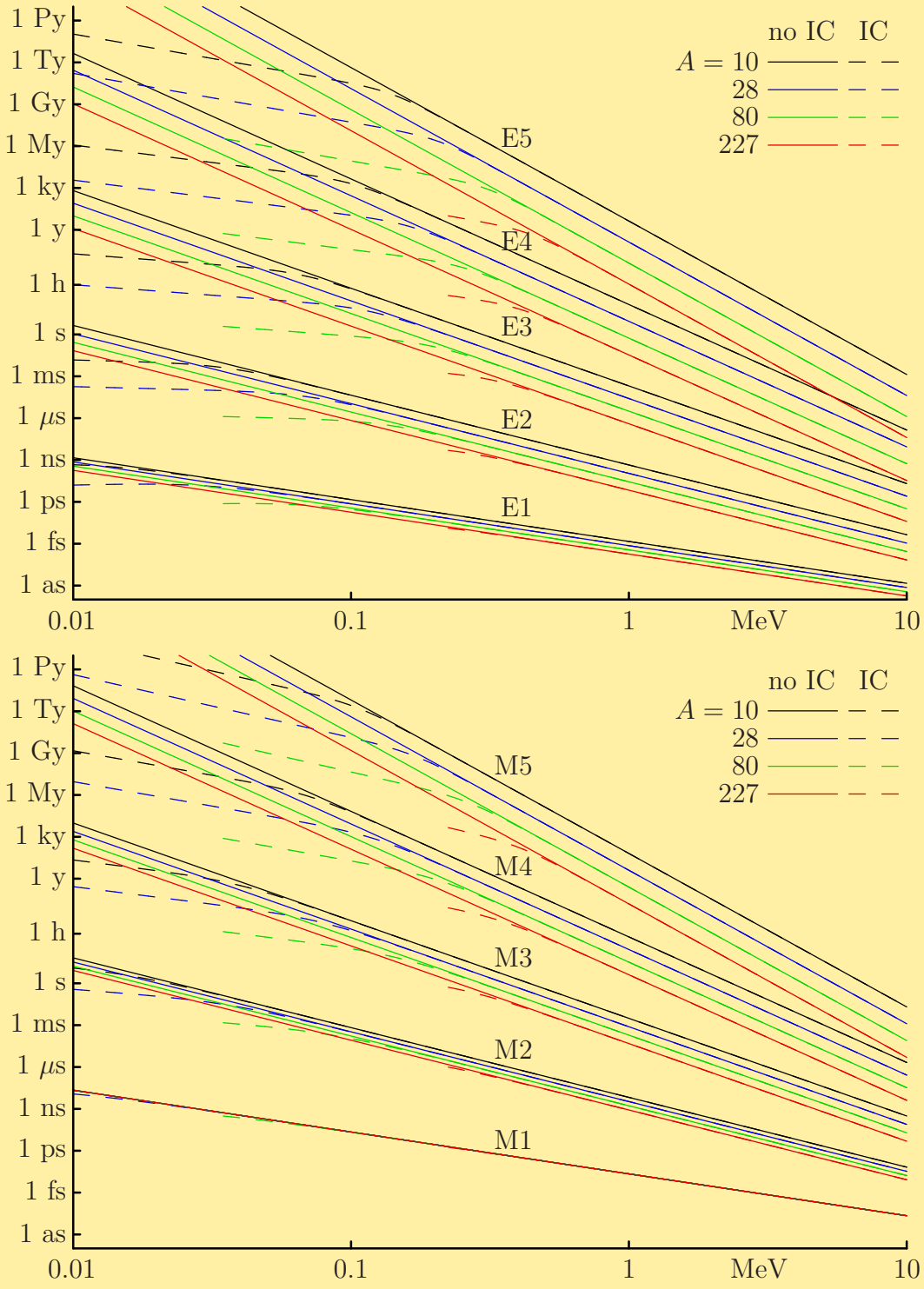


Figure 14.61: Weisskopf ballpark half-lifes for electromagnetic transitions versus energy release. Broken lines include ballparked internal conversion. [pdf]

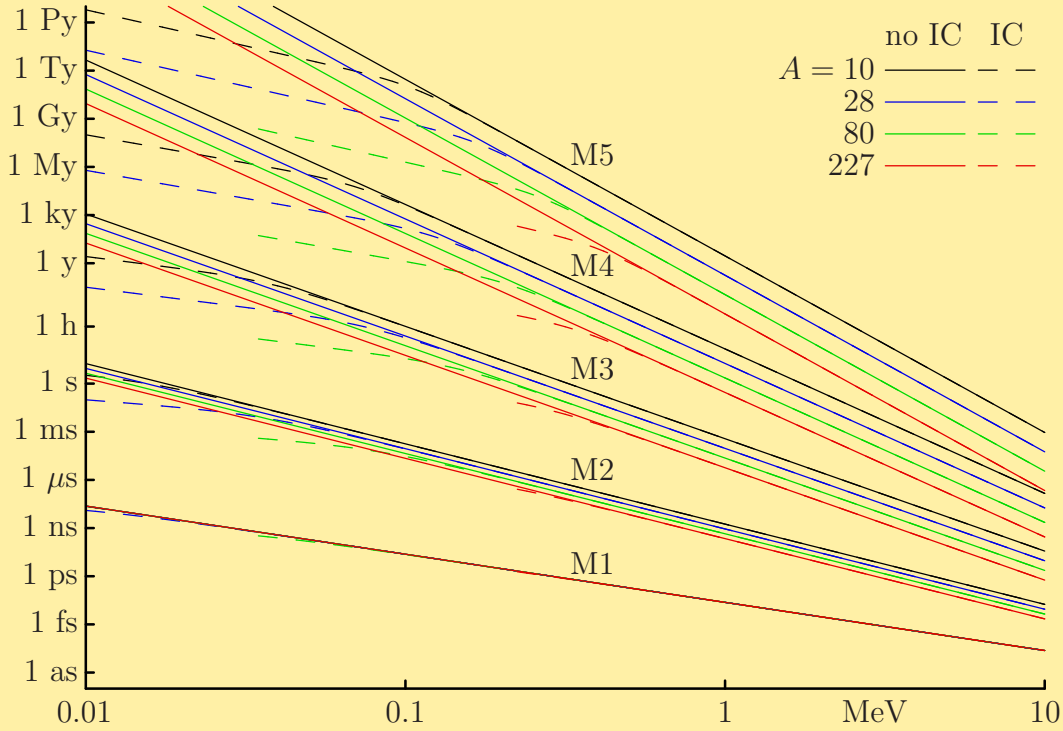


Figure 14.62: Moszkowski ballpark half-lifes for magnetic transitions versus energy release. Broken lines include ballparked internal conversion. [pdf]

Interestingly, M4 transitions tend to be quite close to the mark. Recall that the shell model puts the highest spin states of one harmonic oscillator shell right among the lowest spin states of the next lower shell, 14.15. Transitions between these states involve a parity change and a large change in spin, leading to E3 and M4 transitions. They resemble single-particle transitions as the Weisskopf and Moszkowski estimates assume. The estimates tend to work well for them. One possible reason that they do not end up that much below ballpark as E1 transitions may be that these are heavy nuclei. For heavy nuclei the restrictions put on by isospin may be less confining.

Finally, what other books do not point out is that there is a problem with electrical transitions in the islands of isomerism. There is serious concern about the correctness of the very Hamiltonian used in such transitions, {N.14}. This problem does not seem to affect magnetic multipole transitions in the nonrelativistic approximation.

Another problem not pointed out in various other books is for magnetic transitions. Consider the shell model states, figure 14.15. They allow many transitions inside the bands that by their unit change in angular momentum and unchanged parity are M1 transitions. However, these states have a change in orbital angular momentum equal to two units. The single-particle model on which the Weisskopf and Moszkowski estimates are based predicts zero transition rate

for such M1 transitions. It does not predict the Moszkowski or Weisskopf values given above and in the figures. In general, the predicted single-particle transition rates are zero unless the multipole order ℓ satisfies, {A.25.8}

$$|l_1 - l_2| \leqslant \ell \leqslant l_1 + l_2$$

where l_1 and l_2 are the initial and final orbital azimuthal quantum numbers. Fortunately this is only an issue for magnetic transitions, {A.25.8}.

Note that the single-particle model does give a nontrivial prediction for say an $2p_{1/2}$ to $2p_{3/2}$ M1 transition. That is despite the fact that the simplified Hamiltonian on which it is based would predict zero transition rate for the model system. For say a $4p_{3/2}$ to 2p transition, the Weisskopf and Moszkowski units also give a nontrivial prediction. That, however, is due to the incorrect radial estimate (A.187). The correct single-particle model on which they are based would give this transition rate as zero. Fortunately, transitions like $4p_{3/2}$ to 2p are not likely to be much of a concern.

14.20.5 Draft: Comparison with data

This subsection compares the theoretical Weisskopf and Moszkowski estimates of the previous section with actual data. The data are from NuDat 2, ⟦12⟧. The plotted values are a broad but further quite random selection of data of apparently good quality. A more precise description of the data selection procedure is in {N.34}. Internal conversion effects, as discussed in subsection 14.20.6, have been mathematically removed using the conversion constants given by NuDat 2. Computed decay rates were checked against the decay rates in W.u. as given by NuDat 2.

Figures 14.63 and 14.64 show the results. What is plotted is the half life, scaled to an (harmonic) average nucleus size. In particular,

$$\text{E}\ell\text{:}\quad \tau_{1/2,\text{red}} = \tau_{1/2}\left(\frac{A}{32}\right)^{2\ell/3} \qquad \text{M}\ell\text{:}\quad \tau_{1/2,\text{red}} = \tau_{1/2}\left(\frac{A}{32}\right)^{2(\ell-1)/3}$$

The horizontal coordinate in the figures indicates the energy release Q.

The multipole levels are color coded. The Weisskopf values are shown as broken lines. The solid lines are an attempt at a "best guess" based on the single-particle model. For the electric transitions, they simply use the empirical radial factor {A.25.8} (A.187). For the magnetic transitions, the best guess was based on the more accurate Moszkowski estimates. The empirical radial factor was again used. The momentum factor {A.25.8} (A.189) at minimum multipole order was averaged between the proton and neutron values. The g values in those factors were in turn averaged between their free space and theoretical values.

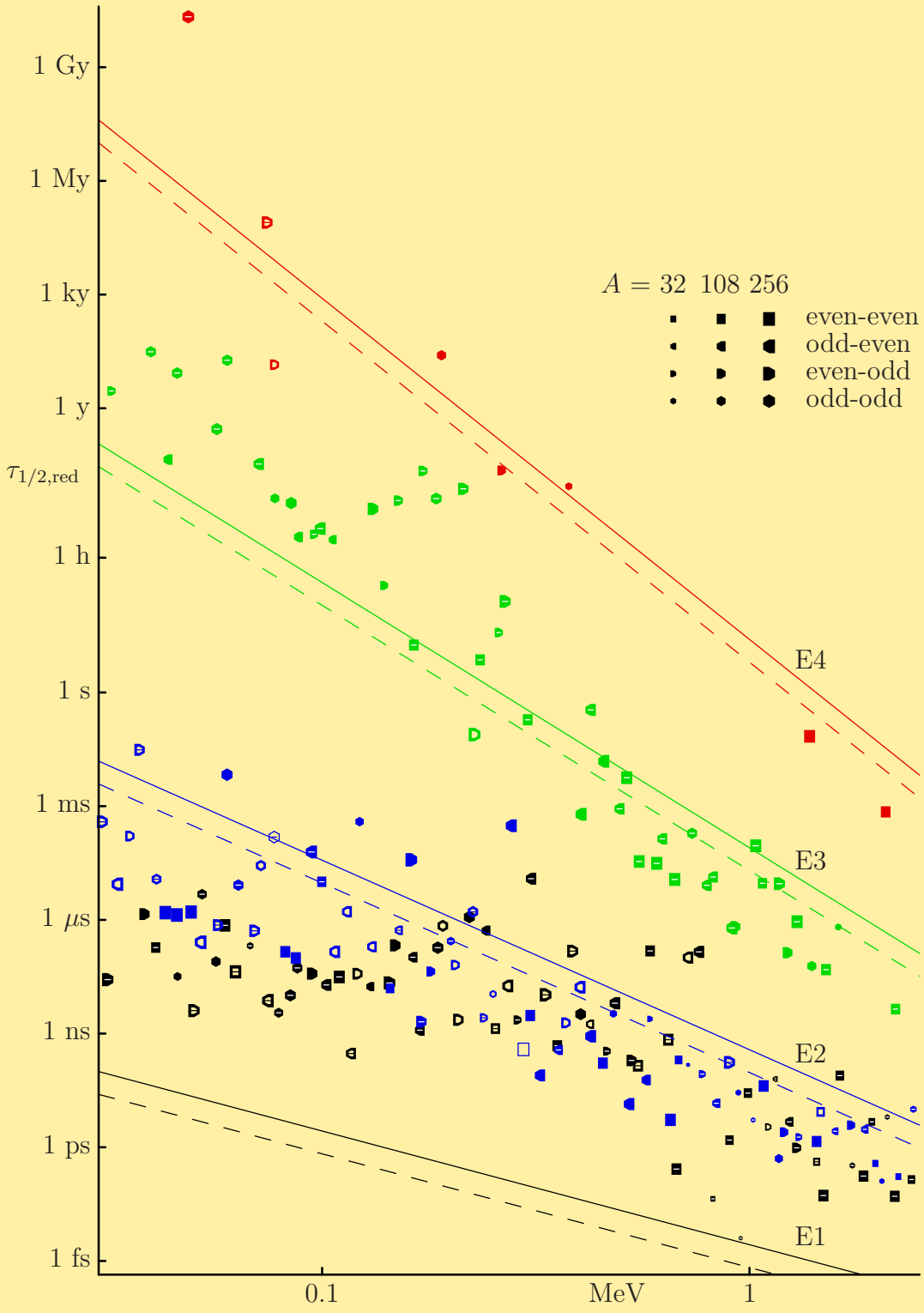


Figure 14.63: Comparison of electric gamma decay rates with theory. [pdf]

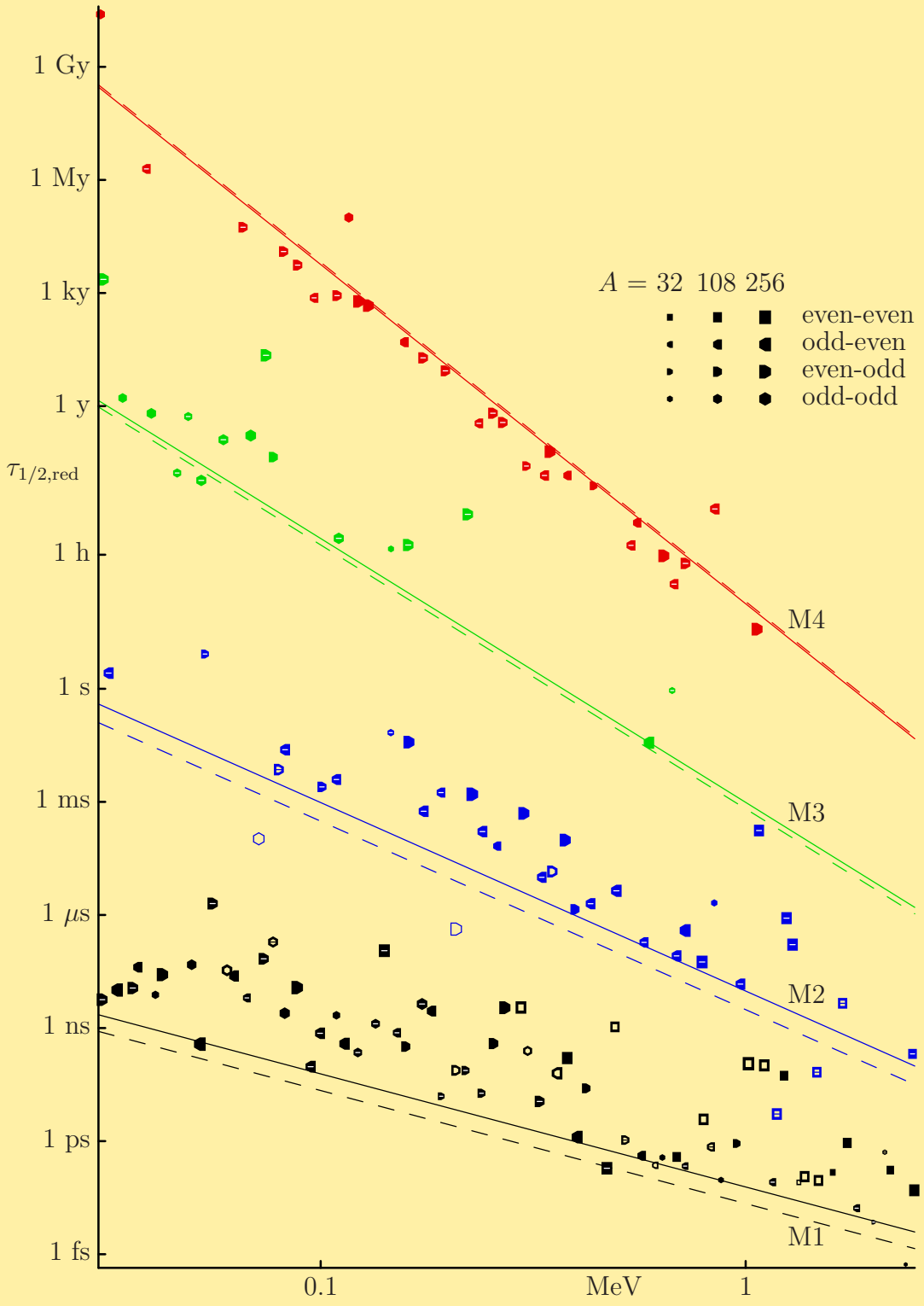


Figure 14.64: Comparison of magnetic gamma decay rates with theory. [pdf]

Symbol size indicates the nucleus size. Symbol shape indicates whether the numbers of protons and neutrons are even or odd. A minus sign indicates that the initial parity is odd; otherwise it is even. The final parity follows from the multipole order and type. An open symbol center indicates that the multipole level ℓ is higher than needed in the transition. More precisely, it indicates that it is higher than the change in nuclear spin.

Please, your mouth is hanging open. It makes you look very goofy. You can almost pretend that the magnetic data are not really as bad as they look, if you cover up those numbers along the vertical axis with your arm.

There is no doubt that if engineers got data like that, they would conclude that something is terribly and fundamentally wrong. Physicists however pooh-pooh the problems.

First of all, physics textbooks typically only present the M4 data graphically like this. Yes, the M4 transitions are typically "only" an order of magnitude or so off. According to the figures here, this "good" agreement happens *only* for the M4 data. Have a look at the E1 and E2 data. They end up pretty much in the same humongous cloud of scattered data. In physics textbooks you do not really see it, as these data are presented in separate histograms. And for some reason, in those histograms the E2 transitions are typically only half an order of magnitude above estimate, rather than 2.5 orders. (The E1 transitions in those histograms are similar to the data presented here.)

Consider now a typical basic nuclear textbook for physicist. According to the book, disagreements of several orders of magnitude from theory can happen. The difference between "can happen" and "are normal" is not defined. The book further explains: "In particular, experimental disintegration rates smaller than the ones predicted by [the Weisskopf estimates] can mean that [the Weisskopf radial factor {A.25.8} (A.187)] is not very reasonable and that the small overlap of the [initial and final nuclear wave functions] decreases the values of λ."

However, the "best guess" E1 line in figure 14.63 uses a better radial estimate. It is not exactly enough to get anywhere near the typical data.

And "poor overlap" is an easy cop-out since nuclear wave functions are not known. For example, it does not explain why some multipole orders like E1 have a very poor overlap of wave functions, and others do not.

Transition rates many orders of magnitude smaller than theory must have a good reason. Random deviations from theory are not a reasonable explanation. Having a transition rate *typically* four orders of magnitude smaller than a reasonable theoretical estimate is like routinely hitting the bull's eye of a 10 cm target to within a mm. There must be something causing this.

But what might that be? Conservation of angular momentum and parity are already fully accounted for. To be sure, conservation of isospin is not. However, isospin is an approximate symmetry. It is not accurate enough to explain reductions by 4 or 5 orders of magnitude. The two examples mentioned in subsection 14.20.2 managed just 1 order of magnitude slow down. And not

all E1 transitions are forbidden anyway. And light nuclei, for which isospin conservation is presumably more accurate, seem no worse than heavier ones in figure 14.63. Actually, the two best data are small nuclei.

To be sure, the above arguments implicitly assume that a bull's eye is hit by an incredibly accurate cancelation of opposite terms in the so-called matrix element that describes transitions. There is an alternate possibility. The final nuclear wave function could be zero where the initial one is nonzero and vice versa. In that case, the integrand in the matrix element is everywhere zero, and no accurate cancellations are needed.

But consider now the top half of figure 14.65. Here mixed E1 + M2 transitions are plotted. These transitions take sometimes place through the electric dipole mechanism and sometimes through the magnetic quadrupole one. Note that there are three very fast M2 transitions, the first, fourth, and tenth. These transitions occur at rates of 49, 58, respectively 21 times faster than the best guess based on the single-particle model. So the initial and final wave functions must unavoidably "overlap" very well. But how then to explain that the corresponding electric rates are 31 000, 7 800, and 200 times *slower* than best guess? The initial and final wave functions are the same. While there is a different operator sandwiched in between, {A.25}, the picture still becomes that one operator apparently achieves a bull's eye of perfect cancellation. There should be an explanation.

The example textbook also notes: "Experimental values higher than predicted by [the Weisskopf estimates] can mean, on the other hand, that the transition involves the participation of more than one nucleon or even a collective participation of the whole nucleus." The textbook notes in particular that the reason that most E2 transitions are faster than theory is due to the fact that these transitions are common among collective bands, especially rotational bands in deformed nuclei.

This is a well established argument. In principle 50 protons transitioning could indeed explain why many E2 transitions in figure 14.63 end up on the order of a rough factor 50^2 faster than theory.

But there are again some problems. For one, the mixed E1 + M2 transitions discussed earlier are not among states in the same rotational bands. The nuclear parity flips over in them. But the mentioned examples showed that several M2 transitions were also much faster than single-particle theory. While that might still be due to collective motion, it does not explain why the E1 transitions then were so slow.

Consider also the bottom of figure 14.65. Here mixed M1 + E2 transitions are plotted. Note that the E2 transitions are again much faster than theory, with few exceptions. But how then to explain that the M1 transitions between the same initial and final states are much slower than theory? More of these miraculously accurate cancellations? There are quite a few transitions at the higher energies where the M1 transition proceeds slower than the E2 one, despite

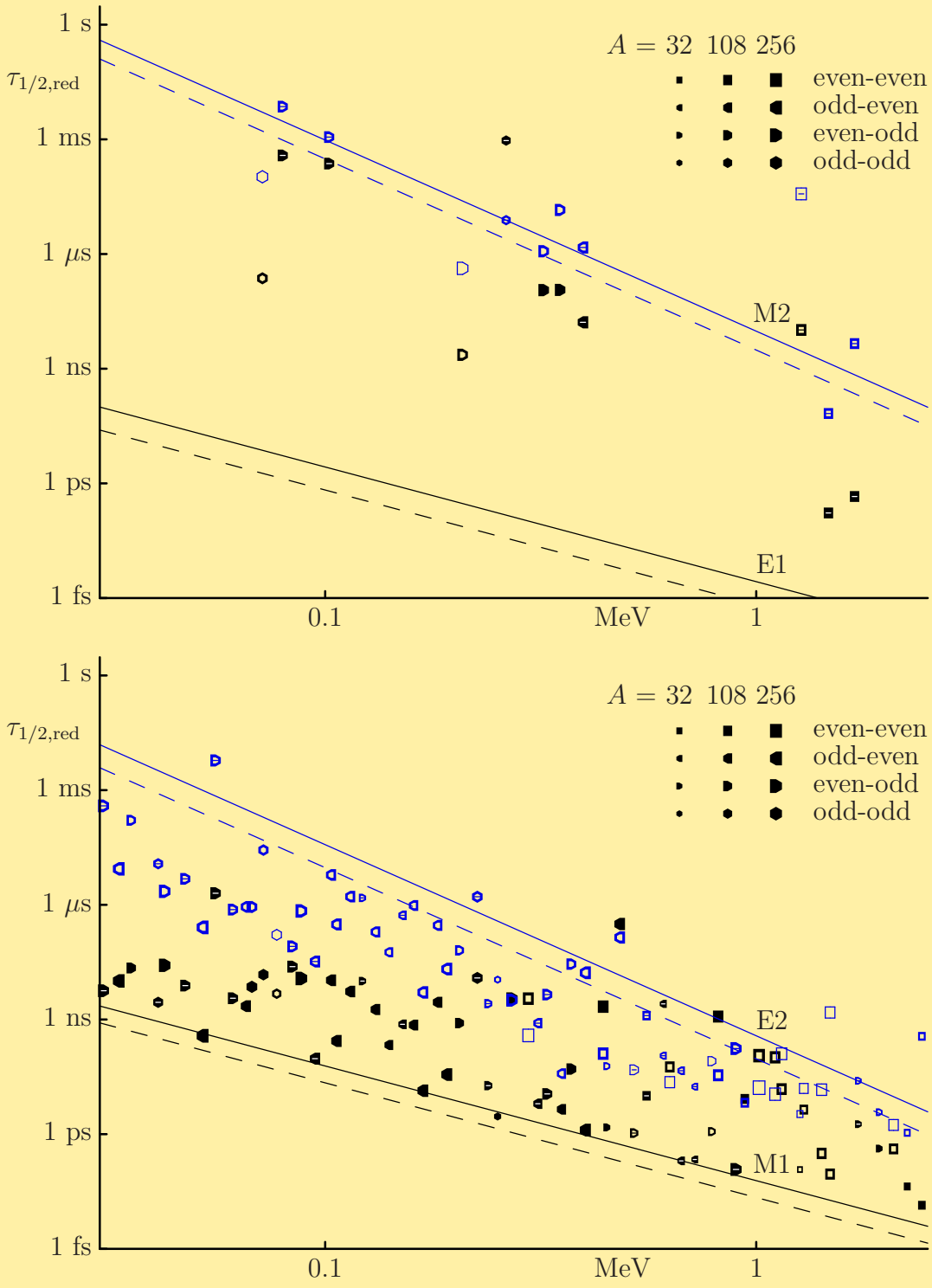


Figure 14.65: Comparisons of decay rates between the same initial and final states.

the difference in multipole order.

It is true that the orbital effect is relatively minor in magnetic transitions of minimal multipole order, {A.25.8}. But the transitions given by black symbols with open centers in the bottom of figure 14.65 are not of minimal multipole order. And in any case, "relatively minor" gets nowhere near to explaining differences of four or five orders of magnitude in relative decay rates.

The same problem exists for the idea that the special nature of transitions within rotational bands might somehow be responsible. Surely the idea of rotational bands is not be far accurate enough to explain the humongous differences? And some of the worst offenders in figure 14.65 are definitely not between states in the same rotational band. Those are again the ones where the black symbol has an open center; the nuclear spin does not change in those transitions.

There are 65 randomly chosen E1 transitions plotted in figure 14.63. Out of these 65, only one manages to achieve the "best guess" theoretical transition rate. That is a boron-10 transition. (You may have to strain your eyes to see it, it is such a small nucleus. It is right on top of the best-guess line, just before 1 MeV.) On the other hand, one transition is slower than best guess by more than 8 orders of magnitude, and another three are slower by more than 7 orders of magnitude.

Compare that with the 67 E2 transitions. Only one manages the three orders of magnitude slower than best guess that is so ho-hum for E1 transitions. That transition is a 2 340 keV $^{29}/_{2}{}^{+}$ to 2 063 keV $^{25}/_{2}{}^{+}$ $^{205}_{85}$At one. The amount of spin that is involved here is not exactly run-of-the mill. Note also that the three runners-up for being far above the E2 line are E1 transitions, rather than E2 ones...

Also, why do E3 transitions act much like E1 transitions in the first half of the energy range and like E2 ones over the second half? It seems weird.

The example textbook concludes: "In figure [...] one notes very good agreement between theoretical values and experimental ones for M4. This behavior is typical for transitions of high multipolarity." Based on figures 14.63 and 14.64, it seems very optimistic to call M4 transitions "typical" for high multipolarity.

Needless to say, then, the author of this book finds the discussion of measured gamma decay versus theory in standard nuclear text books grossly inadequate, and highly unconvincing where it is given at all. If you feel the same way, see note {N.36} for one alternative idea.

14.20.6 Draft: Internal conversion

In internal conversion, a nucleus gets rid of excitation energy by kicking an atomic electron out of the atom. This is most important for transitions between states of zero spin. For such transitions, the normal gamma decay process of emitting a photon is not possible since a photon cannot be emitted with zero angular momentum. However, the ejected electron, called the "conversion

electron," can keep whatever angular momentum it has. (For practical purposes, that is zero. Electrons that are not in s states have negligible probabilility of being found inside the nucleus.) Transitions in which no angular momentum is emitted by the nucleus are called E0 transitions.

A ballpark decay rate for E0 internal conversion can be found in Blatt & Weisskopf [7, 8, p. 621]. Where else. Converted to look similar to the gamma decay Weisskopf estimate (A.190), it reads

$$\boxed{\lambda^{\mathrm{E0}}_{\mathrm{Blatt\&Weisskopf}} \sim \alpha\omega(kR)^4 \frac{2}{25}\alpha(\alpha Z)^3 \left(\frac{2m_{\mathrm{e}}c^2}{Q}\right)^{9/2} \qquad \omega \equiv \frac{Q}{\hbar} \quad k \equiv \frac{Q}{\hbar c}} \tag{14.68}$$

Here $\alpha = e^2/4\pi\epsilon_0\hbar c \approx 1/137$ is the fine structure constant and Q is the energy release. Further $m_{\mathrm{e}}c^2$ is the rest mass energy of the electron, which is about half an MeV. The initial and final parities need to be the same.

Note that the first three factors in the expression above look much like an E2 electric transition. However, the next three factors are very small, though less so for heavy nuclei. On the other hand the final factor can be very large if the energy release Q is much less than an MeV. So E0 internal conversion is relatively speaking most effective for low-energy transitions in heavy nuclei.

Putting in the numbers gives the equivalent of (14.67) as

$$\boxed{\lambda^{\mathrm{E0}} = 3.8 Z^3 A^{4/3} Q^{1/2}} \tag{14.69}$$

Once again, the energy release should be in MeV and then the decay rate will be per second. Note that absolutely speaking the decay rate does in fact increase with the energy release, but very weakly.

Table 14.5 shows how the estimate stands up to scrutiny. The listed E0 transitions are all those for which NuDat 2, ⟦12⟧, gives useful and unambiguous data. All these turn out to be 0^+ to 0^+ transitions.

The second-last column in the table shows a scaled half life. It is scaled to some (harmonic) average nucleus size $Z = 16$, $A = 32$. In particular

$$\tau_{1/2,\mathrm{red}} \equiv \tau_{1/2} \frac{Z^3}{16^3} \frac{A^{4/3}}{32^{4/3}}$$

The final column shows what the scaled half life should be according to the theoretical estimate above. Note that, excluding the final three nuclei, the agreement is not too bad, as they come. What is an order of magnitude or so between friends? After the previous subsections everything would look accurate.

However, the final three nuclei decay much more rapidly than internal conversion predicts. A second decay process occurs here: electron-positron pair creation. This requires that the nuclear energy release Q is at least large enough to provide the rest mass energy of the electron and positron. That is a bit over a MeV. However, as the table suggests, to get a significant effect, more energy

Nucleus	Q	$\tau_{1/2}$ μs	$\tau_{1/2,\mathrm{red}}$ μs	I.C. Theory μs
$^{184}_{80}$Hg	0.375	0.00062	0.80	0.72
$^{72}_{36}$Kr	0.671	0.0263	0.88	0.54
$^{72}_{32}$Ge	0.691	0.4442	10.	0.53
$^{98}_{42}$Mo	0.735	0.0218	1.8	0.51
$^{192}_{82}$Pb	0.769	0.00075	1.1	0.50
$^{98}_{40}$Zr	0.854	0.064	4.4	0.47
$^{194}_{82}$Pb	0.931	0.0011	1.6	0.45
$^{96}_{40}$Zr	1.582	0.0380	2.6	0.35
$^{90}_{40}$Zr	1.761	0.0613	3.8	0.33
$^{68}_{28}$Ni	1.770	0.2760	4.0	0.33
$^{40}_{20}$Ca	3.353	0.00216	0.0057	0.24
$^{32}_{16}$S	3.778	0.00254	0.0025	0.23
$^{16}_{8}$O	6.048	0.00007	0.000003	0.18

Table 14.5: Half lifes for E0 transitions.

is needed. There should be enough additional energy to give the electron and positron relativistically nontrivial kinetic energies.

In transitions other than between states of zero spin, normal gamma decay is possible. But even in those decays, internal conversion and pair production may compete with gamma decay. They are especially important in highly forbidden gamma decays.

The so-called "internal conversion coefficient" α_ℓ gives the internal conversion rate of a transition as a fraction of its gamma decay rate:

$$\boxed{\alpha_\ell = \frac{\lambda_{\mathrm{IC}}}{\lambda_\gamma}} \tag{14.70}$$

The following ballpark values for the internal conversion coefficient in electric and magnetic transitions can be derived ignoring relativistic effects and electron binding energy:

$$\boxed{\alpha_{\mathrm{E}\ell} = \frac{1}{n^3}\frac{\ell}{\ell+1}\alpha(\alpha Z)^3\left(\frac{2m_{\mathrm{e}}c^2}{Q}\right)^{\ell+5/2} \qquad \alpha_{\mathrm{M}\ell} = \frac{1}{n^3}\alpha(\alpha Z)^3\left(\frac{2m_{\mathrm{e}}c^2}{Q}\right)^{\ell+3/2}} \tag{14.71}$$

Here, once again, ℓ is the multipole order of the decay, Q the nuclear energy release, and $\alpha \approx 1/137$ the fine structure constant. Further n is the principal quantum number of the atomic shell that the conversion electron comes from, Note the brilliance of using the same symbol for the internal conversion coefficients as for the fine structure constant. This book will use subscripts to keep them apart.

The above estimates are very rough. They are routinely off by a couple of orders of magnitude. However, they do predict a few correct trends. Internal conversion is relatively more important compared to gamma decay if the energy release Q of the decay is low, if the multipolarity ℓ is high, and if the nucleus is heavy. Ejection from the $n = 1$ K shell tends to dominate ejection from the other shells, but not to a dramatic amount.

(You might wonder why the earlier ballpark for E0 transitions looks mathematically like an $\ell = 2$ rate, instead of some $\ell = 0$ one. The reason is that E0 transitions do not create an electromagnetic field outside the nucleus, compare for example chapter 7.4.3. So the interaction with the electron is limited to the interior of the nucleus. That reduces the magnitude of the interaction greatly.)

Internal conversion is especially useful for investigating nuclei because the conversion coefficients are different for electric and magnetic transitions. Therefore, detailed decay measurements can shed light on the question whether a given transition is an electric or a magnetic one. Since they also depend strongly on the multipole order, they also help establish that. To be sure, the estimates above are not by far accurate enough to do these things. But much more accurate values have been computed using relativistic theories and tabulated.

Internal pair formation supplements internal conversion, [7, 8, p. 622]. The pair formation rate is largest where the internal conversion rate is smallest. That is in the region of low atomic number and high transition energies.

One very old reference incorrectly states that internal conversion happens when a gamma ray emitted by the nucleus knocks a surrounding electron out of the atom. Such a process, the photoelectric effect, is in principle possible, but its probability would be negligibly small. Note in particular that in many decays, almost no gamma rays are emitted but lots of conversion electrons. (While the interaction between the nucleus and the conversion electron is of course caused by photons, these are virtual photons. They would not come out of the nucleus even if you stripped away the atomic electrons.)

It may be noted that "internal conversion" is not unique to nuclei. Energetic atomic electron transitions can also get rid of their energy by ejection of another electron. The ejected electrons are called "Auger electrons." They are named after the physicist Auger, who was the first man to discover the process. (Some unscrupulous woman, Lise Meitner, had discovered and published it earlier, selfishly attempting to steal Auger's credit, {N.35}). In fact, internal conversion can give rise to additional Auger electrons as other electrons rush in to fill the internal converion electron hole. And so can electron capture.

Part IV

Supplementary Information

Appendix A

Addenda

This appendix describes a number of additional topics. They did not seem important enough to warrant including them in the main text. An addition is always a distraction; at the minimum you have to worry about whether you need to worry about it. However, many of the topics below are covered in well-known other texts. Obviously many other authors disagree about their importance. If they turn out to be right, you can find it here.

A.1 Classical Lagrangian mechanics

Lagrangian mechanics is a way to simplify complicated dynamical problems. This note gives a brief overview. For details and practical examples you will need to consult a good book on mechanics.

A.1.1 Introduction

As a trivial example of how Lagrangian mechanics works, consider a simple molecular dynamics simulation. Assume that the forces on the particles are given by a potential that only depends on the positions of the particles.

The difference between the net kinetic energy and the net potential energy is called the "Lagrangian." For a system of particles as considered here it takes the form

$$\mathcal{L} = \sum_j \tfrac{1}{2} m_j |\vec{v}_j|^2 - V(\vec{r}_1, \vec{r}_2, \ldots)$$

where j indicates the particle number and V the potential of the attractions between the particles and any external forces.

It is important to note that in Lagrangian dynamics, the Lagrangian must mathematically be treated as a function of the velocities and positions of the particles. While for a given motion, the positions and velocities are in turn a

function of time, time derivatives must be implemented through the chain rule, i.e. by means of total derivatives of the Lagrangian.

The "canonical momentum" $p^{\mathrm{c}}_{j,i}$ of particle j in the i direction, (with $i = 1$, 2, or 3 for the x, y, or z components respectively), is defined as

$$p^{\mathrm{c}}_{j,i} \equiv \frac{\partial \mathcal{L}}{\partial v_{j,i}}$$

For the Lagrangian above, this is simply the normal momentum $mv_{j,i}$ of the particle in the i-direction.

The Lagrangian equations of motion are

$$\frac{\mathrm{d}p^{\mathrm{c}}_{j,i}}{\mathrm{d}t} = \frac{\partial \mathcal{L}}{\partial r_{j,i}}$$

This is simply Newton's second law in disguise: the left hand side is the time derivative of the linear momentum of particle j in the i-direction, giving mass times acceleration in that direction; the right hand side is the minus the spatial derivative of the potential, which gives the force in the i direction on particle j. Obviously then, use of Lagrangian dynamics does not help here.

A.1.2 Generalized coordinates

One place where Lagrangian dynamics is very helpful is for macroscopic objects. Consider for example the dynamics of a Frisbee. Nobody is going to do a molecular dynamics computation of a Frisbee. What you do is approximate the thing as a "solid body," (or more accurately, a rigid body). The position of every part of a solid body can be fully determined using only six parameters, instead of the countless position coordinates of the individual atoms. For example, knowing the three position coordinates of the center of gravity of the Frisbee and three angles is enough to fully fix it. Or you could just choose three reference points on the Frisbee: giving three position coordinates for the first point, two for the second, and one for the third is another possible way to fix its position.

Such parameters that fix a system are called "generalized coordinates." The word generalized indicates that they do not need to be Cartesian coordinates; often they are angles or distances, or relative coordinates or angles. The number of generalized coordinates is called the number of degrees of freedom. It varies with the system. A bunch of solid bodies moving around freely will have six per solid body; but if there are linkages between them, like the bars in your car's suspension system, it reduces the number of degrees of freedom. A rigid wheel spinning around a fixed axis has only one degree of freedom, and so does a solid pendulum swinging around a fixed axis. Attach a second pendulum to its end, maybe not in the same plane, and the resulting compound pendulum has two degrees of freedom.

If you try to describe such systems using plain old Newtonian mechanics, it can get ugly. For each solid body you can apply that the sum of the forces must equal mass times acceleration of the center of gravity, and that the net moment around the center of gravity must equal the rate of change of angular momentum, which you then presumably deduce using the principal axis system.

Instead of messing with all that complex vector algebra, Lagrangian dynamics allows you to deal with just a single scalar, the Lagrangian. If you can merely figure out the net kinetic and potential energy of your system in terms of your generalized coordinates and their time derivatives, you are in business.

If there are linkages between the members of the system, the benefits magnify. A brute-force Newtonian solution of the three-dimensional compound pendulum would involve six linear momentum equations and six angular ones. Yet the thing has only two degrees of freedom; the angular orientations of the individual pendulums around their axes of rotation. The reason that there are twelve equations in the Newtonian approach is that the support forces and moments exerted by the two axes add another 10 unknowns. A Lagrangian approach allows you to just write two equations for your two degrees of freedom; the support forces do not appear in the story. That provides a great simplification.

A.1.3 Lagrangian equations of motion

This section describes the Lagrangian approach to dynamics in general. Assume that you have chosen suitable generalized coordinates that fully determine the state of your system. Call these generalized coordinates q_1, q_2, ... and their time derivatives $\dot{q}_1$, $\dot{q}_2$, ... The number of generalized coordinates K is the number of degrees of freedom in the system. A generic canonical coordinate will be indicated as q_k.

Now find the kinetic energy T and the potential energy V of your system in terms of these generalized coordinates and their time derivatives. The difference is the Lagrangian:

$$\begin{aligned}\mathcal{L}(q_1, q_2, \ldots, q_K, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_K, t) \\ \equiv \quad T(q_1, q_2, \ldots, q_K, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_K, t) - V(q_1, q_2, \ldots, q_K, t)\end{aligned}$$

Note that the potential energy depends only on the position coordinates of the system, but the kinetic energy also depends on how fast they change with time. Dynamics books give lots of helpful formulae for the kinetic energy of the solid members of your system, and the potential energy of gravity and within springs.

The canonical momenta are defined as

$$p_k^{\mathrm{c}} \equiv \frac{\partial \mathcal{L}}{\partial \dot{q}_k} \tag{A.1}$$

for each individual generalized coordinate q_k. The equations of motion are

$$\frac{\mathrm{d}p_k^{\mathrm{c}}}{\mathrm{d}t} = \frac{\partial \mathcal{L}}{\partial q_k} + Q_k \tag{A.2}$$

There is one such equation for each generalized coordinate q_k, so there are exactly as many equations as there are degrees of freedom. The equations are second order in time, because the canonical momenta involve first order time derivatives of the q_k.

The Q_k terms are called generalized forces, and are only needed if there are forces that cannot be modeled by the potential V. That includes any frictional forces that are not ignored. To find the generalized force Q_k at a given time, imagine that the system is displaced slightly at that time by changing the corresponding generalized coordinate q_k by an infinitesimal amount δq_k. Since this displacement is imaginary, it is called a "virtual displacement." During such a displacement, each force that is not modelled by V produces a small amount of "virtual work." The net virtual work divided by δq_k gives the generalized force Q_k. Note that frictionless supports normally do not perform work, because there is no displacement in the direction of the support force. Also, frictionless linkages between members do not perform net work, since the forces between the members are equal and opposite. Similarly, the internal forces that keep a solid body rigid do not perform work.

The bottom line is that normally the Q_k are zero if you ignore friction. However, any collisions against rigid constraints have to be modeled separately, just like in normal Newtonian mechanics. For an infinitely rigid constraint to absorb the kinetic energy of an impact requires infinite force, and Q_k would have to be an infinite spike if described normally. Of course, you could instead consider describing the constraint as somewhat flexible, with a very high potential energy penalty for violating it. Then make sure to use an adaptive time step in any numerical integration.

It may be noted that in relativistic mechanics, the Lagrangian is *not* the difference between potential and kinetic energy. However, the Lagrangian equations of motion (A.1) and (A.2) still apply.

The general concept that applies both nonrelativistically and relativistically is that of "action." The action $\mathcal{S}$ is defined as the time integral of the Lagrangian:

$$\boxed{\mathcal{S} \equiv \int_{t_1}^{t_2} \mathcal{L}\,\mathrm{d}t} \tag{A.3}$$

Here t_1 and t_2 are suitably chosen starting and ending times that enclose the time interval of interest. The action is unchanged by infinitesimal imaginary displacements of the system. It turns out that that is all that is needed for the Lagrangian equations of motion to apply.

See {D.3.1} for a derivation of the above claims.

A.1.4 Hamiltonian dynamics

For a system with K generalized coordinates the Lagrangian approach provides one equation for each generalized coordinate q_k. These K equations involve second order time derivatives of the K unknown generalized coordinates q_k. However, if you consider the time derivatives $\dot{q}_k$ as K additional unknowns, you get K *first* order equations for these $2K$ unknowns. An additional K equations are:

$$\frac{\mathrm{d}q_k}{\mathrm{d}t} = \dot{q}_k$$

These are no longer trivial because they now give the time derivatives of the first K unknowns in terms of the second K of them. This trick is often needed when using canned software to integrate the equations, because canned software typically only does systems of first order equations.

However, there is a much neater way to get $2K$ first order equations in $2K$ unknowns, and it is particularly close to concepts in quantum mechanics. Define the "Hamiltonian" as

$$H(q_1, q_2, \ldots, q_K, p_1^{\mathrm{c}}, p_2^{\mathrm{c}}, \ldots, p_K^{\mathrm{c}}, t) \equiv \sum_{k=1}^{K} \dot{q}_k p_k^{\mathrm{c}} - \mathcal{L}(q_1, q_2, \ldots, q_K, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_K, t) \tag{A.4}$$

In the right hand side expression, you must rewrite all the time derivatives $\dot{q}_k$ in terms of the canonical momenta

$$p_k^{\mathrm{c}} \equiv \frac{\partial \mathcal{L}}{\partial \dot{q}_k}$$

because the Hamiltonian must be a function of the generalized coordinates and the canonical momenta only. (In case you are not able to readily solve for the $\dot{q}_k$ in terms of the p_k^{c}, things could become messy. But in principle, the equations to solve are linear for given values of the q_k.)

In terms of the Hamiltonian, the equations of motion are

$$\frac{\mathrm{d}q_k}{\mathrm{d}t} = \frac{\partial H}{\partial p_k^{\mathrm{c}}} \qquad \frac{\mathrm{d}p_k^{\mathrm{c}}}{\mathrm{d}t} = -\frac{\partial H}{\partial q_k} + Q_k \tag{A.5}$$

where the Q_k, if any, are the generalized forces as before.

If the Hamiltonian does not explicitly depend on time and the generalized forces are zero, these evolution equations imply that the Hamiltonian does not change with time at all. For such systems, the Hamiltonian is the conserved total energy of the system. In particular for a nonrelativistic system, the Hamiltonian is the sum of the kinetic and potential energies, provided that the position of the system only depends on the generalized coordinates and not also explicitly on time.

See {D.3.2} for a derivation of the above claims.

A.1.5 Fields

The previous subsections discussed discrete mechanical objects like molecules, Frisbees, and pendulums. However, the Lagrangian and Hamiltonian formalisms can be generalized to fields like the electromagnetic field. That is mainly important for advanced physics like quantum field theories; these are not really covered in this book. But since it does appear in one advanced addendum, {A.22}, this subsection will summarize the main points.

The simplest classical field is the electrostatic potential φ. However, there may be more than one potential in a system. For example, in electrodynamics there are also vector potentials. So the generic potential will be indicated as φ_α, where the index α indicates what particular potential it is. A single potential φ_α is still characterized by infinitely many variables: there is a value of the potential at each position.

In addition there may be discrete variables. Electromagnetics would be pretty boring if you would not have some charged particles around. A generic coordinate of such a particle will be indicated as q_k. For example, if there is just one charged particle, q_1, q_2, and q_3 could represent the x, y, and z components of the position of the particle. If there are more particles, just keep increasing k.

Under the above conditions, the Lagrangian will involve an integral:

$$L = L_0 + \int \pounds \, \mathrm{d}^3\vec{r}$$

Here $\pounds$ is called the "Lagrangian density." It is essentially a Lagrangian per unit volume. The integral is over all space.

The first part L_0 is as before. It will depend on the discrete variables and their time derivatives:

$$L_0 = L_0(\ldots; q_k, \dot{q}_k; \ldots)$$

The dot indicates the time derivative of the variable.

The Lagrangian density $\pounds$ will depend on both the fields and the discrete coordinates:

$$\pounds = \pounds(\ldots; \varphi_\alpha, \varphi_{\alpha_t}, \varphi_{\alpha_x}, \varphi_{\alpha_y}, \varphi_{\alpha_z}; \ldots; q_k; \dot{q}_k; \ldots)$$

Here the subscripts on the field indicate partial derivatives:

$$\varphi_{\alpha_t} = \frac{\partial \varphi_\alpha}{\partial t} \quad \varphi_{\alpha_x} = \frac{\partial \varphi_\alpha}{\partial x} \quad \varphi_{\alpha_y} = \frac{\partial \varphi_\alpha}{\partial y} \quad \varphi_{\alpha_z} = \frac{\partial \varphi_\alpha}{\partial z}$$

In principle, there is no reason why the Lagrangian could not contain higher order derivatives, but fortunately you do not see such things in quantum field theories.

This brings up one practical point. Consider a contribution such as the potential energy of a particle called P with charge $\bar{q}_{\rm P}$ in an electrostatic field φ. Assuming that the particle is a point charge, that potential energy is $\bar{q}_{\rm P}\varphi_{\rm P}$ where $\varphi_{\rm P}$ is the potential evaluated at the position $\vec{r}_{\rm P}$ of the particle. But potentials evaluated at a point are problematic. You would really want the potentials to always appear inside integrals. To achieve that, you can assume that the particle is not really a point charge. That its charge is spread out just a little bit around the nominal position $\vec{r}_{\rm P}$. In that case, the potential energy takes the form:

$$\int \bar{q}_{\rm P}\delta^3_\varepsilon(\vec{r}-\vec{r}_{\rm P})\varphi(\vec{r};t)\,{\rm d}^3\vec{r} \approx \bar{q}_{\rm P}\varphi(\vec{r}_{\rm P};t)$$

Here $\delta^3_\varepsilon(\vec{r}-\vec{r}_{\rm P})$ is some chosen function that is zero except within some small distance ε of $\vec{r}_{\rm P}$, and that integrates to one. Because this function is zero except very close to $\vec{r}_p$, you can approximate $\varphi(\vec{r};t)$ by $\varphi(\vec{r}_{\rm P};t)$ and then take it out of the integral. That gives the original expression for the potential energy. But the integral is easier to use in the Lagrangian. Its integrand becomes part of the Lagrangian density. And you can always take the limit $\varepsilon \to 0$ at the end of the day to get point charges.

The Lagrangian equations for the discrete parameters are exactly the same as before, but of course now the Lagrangian includes the integral, {D.3.3}:

$$\frac{\rm d}{{\rm d}t}\left(\frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} + \int \frac{\partial \pounds}{\partial \dot{q}_k}\,{\rm d}^3\vec{r}\right) = \frac{\partial \mathcal{L}_0}{\partial q_k} + \int \frac{\partial \pounds}{\partial q_k}\,{\rm d}^3\vec{r} \tag{A.6}$$

There is one such equation for each discrete parameter q_k, valid at any time.

The Lagrangian equations for the field are based on the Lagrangian density instead of the Lagrangian itself. That is why you really want to have the terms involving the field as integrals. The equations are

$$\frac{\partial}{\partial t}\left(\frac{\partial \pounds}{\partial \varphi_{\alpha_t}}\right) + \frac{\partial}{\partial x}\left(\frac{\partial \pounds}{\partial \varphi_{\alpha_x}}\right) + \frac{\partial}{\partial y}\left(\frac{\partial \pounds}{\partial \varphi_{\alpha_y}}\right) + \frac{\partial}{\partial z}\left(\frac{\partial \pounds}{\partial \varphi_{\alpha_z}}\right) = \frac{\partial \pounds}{\partial \varphi_\alpha} \tag{A.7}$$

There is one such equation for each field φ_α, valid at any position and time.

The canonical momenta are now

$$p^{\rm c}_k \equiv \frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} + \int \frac{\partial \pounds}{\partial \dot{q}_k}\,{\rm d}^3\vec{r} \qquad \pi^{\rm c}_\alpha \equiv \frac{\partial \pounds}{\partial \varphi_{\alpha_t}} \tag{A.8}$$

Note that the field momentum $\pi^{\rm c}_\alpha$ is per unit volume.

The Hamiltonian is

$$H = \sum_k p^{\rm c}_k \dot{q}_k + \sum_\alpha \int \pi^{\rm c}_\alpha \varphi_{\alpha_t}\,{\rm d}^3\vec{r} - \mathcal{L} \tag{A.9}$$

The time derivatives $\dot{q}_k$ and φ_{α_t} must again be expressed in terms of the corresponding canonical momenta.

Hamilton's equations for discrete variables are as before:

$$\frac{\mathrm{d}q_k}{\mathrm{d}t} = \frac{\partial H}{\partial p_k^{\mathrm{c}}} \qquad \frac{\mathrm{d}p_k^{\mathrm{c}}}{\mathrm{d}t} = -\frac{\partial H}{\partial q_k} \tag{A.10}$$

The equations for the fields are a bit tricky. If there are no discrete variables, there is no problem. Then the Hamiltonian can be written in terms of a Hamiltonian density h as

$$H = \int h \, \mathrm{d}^3\vec{r}$$

In that case Hamilton's equations are

$$\frac{\partial \varphi_\alpha}{\partial t} = \frac{\partial h}{\partial \pi_\alpha^{\mathrm{c}}} \qquad \frac{\partial \pi_\alpha^{\mathrm{c}}}{\partial t} = -\frac{\partial h}{\partial \varphi_\alpha} + \frac{\partial}{\partial x}\left(\frac{\partial h}{\partial \varphi_{\alpha_x}}\right) + \frac{\partial}{\partial y}\left(\frac{\partial h}{\partial \varphi_{\alpha_y}}\right) + \frac{\partial}{\partial z}\left(\frac{\partial h}{\partial \varphi_{\alpha_z}}\right)$$

Unfortunately, if there are discrete parameters, products of integrals will appear. Then there is no Hamiltonian density. So the only thing you can do do is differentiate the full Hamiltonian H instead of a Hamiltonian density h. At the end of every differentiation, you will then need to drop an $\int$ and a $\mathrm{d}^3\vec{r}$. In particular, differentiate the Hamiltonian H until you have to start differentiating inside an integral, like, say,

$$\frac{\partial}{\partial \varphi_\alpha} \int \pounds \, \mathrm{d}^3\vec{r}$$

At that time, make the substitution

$$\frac{\partial}{\partial \varphi_\alpha} \int \pounds \, \mathrm{d}^3\vec{r} \quad \Longrightarrow \quad \frac{\partial \pounds}{\partial \varphi_\alpha}$$

This will produce the right answer, although the left hand side above is mathematically complete nonsense.

See {D.3.3} for a justification of this procedure and the other claims in this subsection.

A.2 An example of variational calculus

The problem to solve in addendum {A.22.1} provides a simple example of variational calculus.

The problem can be summarized as follows. Given is the following expression for the net energy of a system:

$$E = \frac{\epsilon_1}{2} \int (\nabla\varphi)^2 \, \mathrm{d}^3\vec{r} - \int \sigma_{\mathrm{p}} \varphi \, \mathrm{d}^3\vec{r} \tag{1}$$

Here the operator ∇ is defined as

$$\nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z} \quad \Longrightarrow \quad \nabla\varphi \equiv \hat{\imath}\frac{\partial\varphi}{\partial x} + \hat{\jmath}\frac{\partial\varphi}{\partial y} + \hat{k}\frac{\partial\varphi}{\partial z}$$

The integrals are over all space, or over some other given region. Further ϵ_1 is assumed to be a given positive constant and $\sigma_{\rm p} = \sigma_{\rm p}(\vec{r})$ is a given function of the position $\vec{r}$. The function $\varphi = \varphi(\vec{r})$ will be called the potential and is not given. Obviously the energy depends on what this potential is. Mathematicians would say that E is a "functional," a number that depends on what a function is.

The energy E will be minimal for some specific potential $\varphi_{\rm min}$. The objective is now to find an equation for this potential $\varphi_{\rm min}$ using variational calculus.

To do so, the basic idea is the following: imagine that you start at $\varphi_{\rm min}$ and then make an infinitesimally small change $\mathrm{d}\varphi$ to it. In that case there should be no change $\mathrm{d}E$ in energy. After all, if there was an negative change in E, then E would decrease. That would contradict that $\varphi_{\rm min}$ produces the lowest energy of all. If there was an positive infinitesimal change in E, then a change in potential of opposite sign would give a negative change in E. Again that produces a contradiction to what is given.

The typical physicist would now work out the details as follows. The slightly perturbed potential is written as

$$\varphi(\vec{r}) = \varphi_{\rm min}(\vec{r}) + \delta\varphi(\vec{r})$$

Note that the d in $\mathrm{d}\varphi$ has been renotated as δ. That is because everyone does so in variational calculus. The symbol does not make a difference, the idea remains the same. Note also that $\delta\varphi$ is a function of position; the change away from $\varphi_{\rm min}$ is normally different at different locations. You are in fact allowed to choose anything you like for the function $\delta\varphi$, as long as it is sufficiently small and it is zero at the limits of integration.

Now just take differentials like you typically do it in calculus or physics. If in calculus you had some expression like f^2, you would say $\mathrm{d}f^2 = 2f\mathrm{d}f$. (For example, if f is a function of a variable t, then $\mathrm{d}f^2/\mathrm{d}t = 2f\mathrm{d}f/\mathrm{d}t$. But physicists usually do not bother with the $\mathrm{d}t$; then they do not have to worry what exactly f is a function of.) Similarly

$$\delta(\nabla\varphi)^2 = 2(\nabla\varphi)\cdot\delta(\nabla\varphi)$$

where

$$\delta\nabla\varphi = \nabla(\varphi_{\rm min} + \delta\varphi) - \nabla(\varphi_{\rm min}) = \nabla\delta\varphi$$

so

$$\delta(\nabla\varphi)^2 = 2(\nabla\varphi)\cdot(\nabla\delta\varphi)$$

For a change starting from $\varphi_{\min}$:

$$\delta(\nabla\varphi)^2 = 2(\nabla\varphi_{\min})\cdot(\nabla\delta\varphi)$$

(Note that φ by itself gets approximated as $\varphi_{\min}$, but $\delta\varphi$ is the completely arbitrary change that can be anything.) Also,

$$\delta(\sigma_{\rm p}\varphi) = \sigma_{\rm p}\delta\varphi$$

because $\sigma_{\rm p}$ is a given constant at every position.

Total you get for the change in energy that must be zero

$$0 = \delta E = \frac{\epsilon_1}{2}\int 2(\nabla\varphi_{\min})\cdot(\nabla\delta\varphi)\,{\rm d}^3\vec{r} - \int \sigma_{\rm p}\delta\varphi\,{\rm d}^3\vec{r} \tag{2}$$

A conscientious mathematician would shudder at the above manipulations. And for good reason. Small changes are not good mathematical concepts. There is no such thing as "small" in mathematics. There are just limits where things go to zero. What a mathematician would do instead is write the change in potential as a some multiple λ of a chosen function $\varphi_{\rm c}$. So the changed potential is written as

$$\varphi(\vec{r}) = \varphi_{\min}(\vec{r}) + \lambda\varphi_{\rm c}(\vec{r})$$

The chosen function $\varphi_{\rm c}$ can still be anything that you want that vanishes at the limits of integration. But it is not assumed to be "small." So now no mathematical nonsense is written. The energy for this changed potential is

$$E = \frac{\epsilon_1}{2}\int[\nabla(\varphi_{\min}+\lambda\varphi_{\rm c})]^2\,{\rm d}^3\vec{r} - \int\sigma_{\rm p}(\varphi_{\min}+\lambda\varphi_{\rm c})\,{\rm d}^3\vec{r}$$

Now this energy is a function of the multiple λ. And that is a simple numerical variable. The energy must be smallest at $\lambda = 0$, because $\varphi_{\min}$ gives the minimum energy. So the above function of λ must have a minimum at $\lambda = 0$. That means that it must have a zero derivative at $\lambda = 0$. So just differentiate the expression with respect to λ. (You can differentiate as is, or simplify first and bring λ outside the integrals.) Set this derivative to zero at $\lambda = 0$. That gives the same result (2) as derived by physicists, except that $\varphi_{\rm c}$ takes the place of $\delta\varphi$. The result is the same, but the derivation is nowhere fishy.

This derivation will return to the notations of physicists. The next step is to get rid of the derivatives on $\delta\varphi$. Note that

$$\int(\nabla\varphi_{\min})\cdot(\nabla\delta\varphi)\,{\rm d}^3\vec{r} = \iiint\frac{\partial\varphi_{\min}}{\partial x}\frac{\partial\delta\varphi}{\partial x} + \frac{\partial\varphi_{\min}}{\partial y}\frac{\partial\delta\varphi}{\partial y} + \frac{\partial\varphi_{\min}}{\partial z}\frac{\partial\delta\varphi}{\partial z}\,{\rm d}x{\rm d}y{\rm d}z$$

The way to get rid of the derivatives on $\delta\varphi$ is by integration by parts. Integration by parts pushes a derivative from one factor on another. Here you see the real reason why the changes in potential must vanish at the limits of integration. If

they did not, integrations by parts would bring in contributions from the limits of integration. That would be a mess.

Integrations by parts of the three terms in the integral in the x, y, and z directions respectively produce

$$\int (\nabla\varphi_{\min}) \cdot (\nabla\delta\varphi)\, \mathrm{d}^3\vec{r} = \iiint -\frac{\partial^2\varphi_{\min}}{\partial x^2}\delta\varphi - \frac{\partial^2\varphi_{\min}}{\partial y^2}\delta\varphi - \frac{\partial^2\varphi_{\min}}{\partial z^2}\delta\varphi \,\mathrm{d}x\mathrm{d}y\mathrm{d}z$$

In vector notation, that becomes

$$\int (\nabla\varphi_{\min}) \cdot (\nabla\delta\varphi)\, \mathrm{d}^3\vec{r} = -\int (\nabla^2\varphi_{\min})\delta\varphi\, \mathrm{d}^3\vec{r}$$

Substituting that in the change of energy (2) gives

$$0 = \delta E = \int (-\epsilon_1\nabla^2\varphi_{\min} - \sigma_{\mathrm{p}})\delta\varphi\, \mathrm{d}^3\vec{r}$$

The final step is to say that this can only be true for whatever change $\delta\varphi$ you take if the parenthetical expression is zero. That gives the final looked-for equation for $\varphi_{\min}$:

$$-\epsilon_1\nabla^2\varphi_{\min} - \sigma_{\mathrm{p}} = 0 \tag{3}$$

To justify the above final step, call the parenthetical expression f for short. Then the variational statement above is of the form

$$\int f\delta\varphi\, \mathrm{d}^3\vec{r} = 0$$

where $\delta\varphi$ can be arbitrarily chosen as long as it is zero at the limits of integration. It is now to be shown that this implies that f is everywhere zero inside the region of integration.

(Note here that whatever function f is, it should not contain $\delta\varphi$. And there should not be any derivatives of $\delta\varphi$ anywhere at all. Otherwise the above statement is not valid.)

The best way to see that f must be zero everywhere is first assume the opposite. Assume that f is nonzero at some point P. In that case select a function $\delta\varphi$ that is zero everywhere except in a small vicinity of P, where it is positive. (Make sure the vicinity is small enough that f does not change sign in it.) Then the integral above is nonzero; in particular, it will have the same sign as f at P. But that is a contradiction, since the integral must be zero. So the function f cannot be nonzero at a point P; it must be zero everywhere.

(There are more sophisticated ways to do this. You could take $\delta\varphi$ as a positive multiple of f that fades away to zero away from point P. In that case the integral will be positive unless f is everywhere zero. And sign changes in f are no longer a problem.)

A.3 Galilean transformation

The Galilean transformation describes coordinate system transformations in nonrelativistic Newtonian physics. This note explains these transformation rules. Essentially the same analysis also applies to Lorentz transformations between observers using arbitrarily chosen coordinate systems. The small difference will be indicated.

Consider two observers A' and B' that are in inertial motion. In other words, they do not experience accelerating forces. The two observers move with a relative velocity of magnitude V relative to each other. Observer A' determines the time of events using a suitable clock. This clock displays the time $t_{A'}$ as a single number, say as the number of seconds since a suitably chosen reference event. To specify the position of events, observer A' uses a Cartesian coordinate system $(x_{A'}, y_{A'}, z_{A'})$ that is at rest compared to him. The origin of the coordinate system is chosen at a suitable location, maybe the location of the reference event that is used as the zero of time.

Observer B' determines time using a clock that indicates a time $t_{B'}$. This time might be zero at a different reference event than the time $t_{A'}$. To specify the position of events, observer B' uses a Cartesian coordinate system $(x_{B'}, y_{B'}, z_{B'})$ that is at rest compared to her. The origin of this coordinate system is different from the one used by observer A'. For one, the two origins are in motion compared to each other with a relative speed V.

The question is now, what is the relationship between the times and positions that these two observers attach to arbitrary events.

To answer this, it is convenient to introduce two additional observers A and B. Observer A is at rest compared to observer A'. However, she takes her zero of time and the origin of her coordinate system from observer B'. In particular, the location and time that A associates with her origin at time zero is also the origin at time zero for observer B':

$$(x_A, y_A, z_A, t_A) = (0,0,0,0) \qquad \Longleftrightarrow \qquad (x_{B'}, y_{B'}, z_{B'}, t_{B'}) = (0,0,0,0)$$

The other additional observer, B, is at rest compared to B'. Like observer A, observer B uses the same origin and zero of time as observer B':

$$(x_B, y_B, z_B, t_B) = (0,0,0,0) \qquad \Longleftrightarrow \qquad (x_{B'}, y_{B'}, z_{B'}, t_{B'}) = (0,0,0,0)$$

Observer B orients her coordinate system like A does.

That makes the relationship between A and B just like A and B as discussed for the Lorentz transform, figure 1.2. However, the classical Galilean transformation is much simpler than the Lorentz transformation. It is

$$\boxed{t_B = t_A \qquad x_B = x_A - Vt_A \qquad y_B = y_A \qquad z_B = z_A} \tag{A.11}$$

Note however that these classical formulae are only an approximation. They can only be used if the relative velocity V between the observers is much smaller

than the speed of light. In fact, if you take the limit $c \to \infty$ of the Lorentz transformation (1.6), you get the Galilean transformation above.

The question still is how to relate the times and locations that observer A' attaches to events to those that observer B' does. To answer that, it is convenient to do it in stages. First relate the times and locations that A' attaches to events to the ones that A does. Then use the formulae above to relate the times and locations that A attaches to events to the ones that B does. Or, if you want the relativistic transformation, at this stage use the Lorentz transformation (1.6). Finally, relate the times and locations that B attaches to events to the ones that B' does.

Consider then now the relationship between the times and locations that A' attaches to events and the ones that A does. Since observer A and A' are at rest relative to each other, they agree about differences in time between events. However, A' uses a different zero for time. Therefore, the relation between the times used by the two observers is

$$\boxed{t_{\rm A} = t_{\rm A'} - \tau_{\rm AA'}}$$

Here $\tau_{\rm AA'}$ is the time that observer A' associates with the reference event that observer A uses as time zero. It is a constant, and equal to $-\tau_{\rm A'A}$. The latter can be seen by simply setting $t_{\rm A'}$ zero in the formula above.

To specify the location of events, both observers A' and A use Cartesian coordinate systems. Since the two observers are at rest compared to each other, they agree on distances between locations. However, their coordinate systems have different origins. And they are also oriented under different angles. That makes the unit vectors $\hat\imath$, $\hat\jmath$, and $\hat k$ along the coordinate axes different. In vector form the relation between the coordinates is then:

$$\boxed{x_{\rm A}\hat\imath_{\rm A} + y_{\rm A}\hat\jmath_{\rm A} + z_{\rm A}\hat k_{\rm A} = (x_{\rm A'} - \xi_{\rm AA'})\hat\imath_{\rm A'} + (y_{\rm A'} - \eta_{\rm AA'})\hat\jmath_{\rm A'} + (z_{\rm A'} - \zeta_{\rm AA'})\hat k_{\rm A'}} \tag{A.12}$$

Here $\xi_{\rm AA'}$, $\eta_{\rm AA'}$, and $\theta_{\rm AA'}$ are the position coordinates that observer A' associates with the origin of the coordinate system of A. By putting $(x_{\rm A'}, y_{\rm A'}, z_{\rm A'})$ to zero in the expression above, you can relate this to the coordinates that A attaches to the origin of A'.

The above equations can be used to find the coordinates of A in terms of those of A. To do so, you will need to know the components of the unit vectors used by A' in terms of those used by A. In other words, you need to know the dot products in

$$\begin{aligned}
\hat\imath_{\rm A'} &= (\hat\imath_{\rm A'}\cdot\hat\imath_{\rm A})\hat\imath_{\rm A} + (\hat\imath_{\rm A'}\cdot\hat\jmath_{\rm A})\hat\jmath_{\rm A} + (\hat\imath_{\rm A'}\cdot\hat k_{\rm A})\hat k_{\rm A} \\
\hat\jmath_{\rm A'} &= (\hat\jmath_{\rm A'}\cdot\hat\imath_{\rm A})\hat\imath_{\rm A} + (\hat\jmath_{\rm A'}\cdot\hat\jmath_{\rm A})\hat\jmath_{\rm A} + (\hat\jmath_{\rm A'}\cdot\hat k_{\rm A})\hat k_{\rm A} \\
\hat k_{\rm A'} &= (\hat k_{\rm A'}\cdot\hat\imath_{\rm A})\hat\imath_{\rm A} + (\hat k_{\rm A'}\cdot\hat\jmath_{\rm A})\hat\jmath_{\rm A} + (\hat k_{\rm A'}\cdot\hat k_{\rm A})\hat k_{\rm A}
\end{aligned}$$

Then these relations allow you to sum the $\hat{\imath}_{\rm A}$ components in the right hand side of (A.12) to give $x_{\rm A}$. Similarly the $\hat{\jmath}_{\rm A}$ components sum to $y_{\rm A}$ and the $\hat{k}_{\rm A}$ components to $z_{\rm A}$.

Note also that if you know these dot products, you also know the ones for the inverse transformation, from A to A'. For example,

$$(\hat{\imath}_{\rm A} \cdot \hat{\imath}_{\rm A'}) = (\hat{\imath}_{\rm A'} \cdot \hat{\imath}_{\rm A}) \qquad (\hat{\imath}_{\rm A} \cdot \hat{\jmath}_{\rm A'}) = (\hat{\jmath}_{\rm A'} \cdot \hat{\imath}_{\rm A}) \qquad \text{etcetera}$$

(In terms of linear algebra, the dot products form a 3×3 matrix. This matrix is called "unitary," or as a real matrix also more specifically "orthonormal," since it preserves distances between locations. The matrix for the reverse transform is found by taking a transpose.)

The relationship between observers B and B' is a simplified version of the one between observers A and A'. It is simpler because B and B' use the same zero of time and the same origin. Therefore the formulae can be obtained from the ones given above by replacing A' and A by B and B' and dropping the terms related to time and origin shifts.

A.4 More on index notation

Engineering students are often much more familiar with linear algebra than with tensor algebra. So it may be worthwhile to look at the Lorentz transformation from a linear algebra point of view. The relation to tensor algebra will be indicated. If you do not know linear algebra, there is little point in reading this addendum.

A contravariant four-vector like position can be pictured as a column vector that transforms with the Lorentz matrix Λ. A covariant four-vector like the gradient of a scalar function can be pictured as a row vector that transforms with the inverse Lorentz matrix Λ^{-1}:

$$\overset{\hookrightarrow}{r}_{\rm B} = \Lambda \overset{\hookrightarrow}{r}_{\rm A} \qquad \left(\overset{\hookrightarrow}{\nabla} f\right)^{\rm T}_{\rm B} = \left(\overset{\hookrightarrow}{\nabla} f\right)^{\rm T}_{\rm A} \Lambda^{-1}$$

In linear algebra, a superscript T transforms columns into rows and vice-versa. Since you think of the gradient by itself as a column vector, the T turns it into a row vector. Note also that putting the factors in a product in the correct order is essential in linear algebra. In the second equation above, the gradient, written as a row, premultiplies the inverse Lorentz matrix.

In tensor notation, the above expressions are written as

$$x^\mu_{\rm B} = \lambda^\mu{}_\nu x^\nu_{\rm A} \qquad \partial_{\mu,{\rm B}} f = \partial_{\nu,{\rm A}} f \left(\lambda^{-1}\right)^\nu{}_\mu$$

The order of the factors is now no longer a concern; the correct way of multiplying follows from the names of the indices.

The key property of the Lorentz transformation is that it preserves dot products. Pretty much everything else follows from that. Therefore the dot product must now be formulated in terms of linear algebra. That can be done as follows:

$$\vec{r}_1 \cdot \vec{r}_2 \equiv \vec{r}_1^{\,\mathrm{T}} G \vec{r}_2 \qquad \text{where} \quad G = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The matrix G is called the "Minkowski metric." The effect of G on $\vec{r}_2$ is to flip over the sign of the zeroth, time, entry. Looking at it another way, the effect of G on the preceding $\vec{r}_1^{\,\mathrm{T}}$ is to flip over the sign of its zeroth entry. Either way, G provides the minus sign for the product of the time coordinates in the dot product.

In tensor notation, the above expression must be written as

$$\vec{r}_1 \cdot \vec{r}_2 \equiv x_1^\mu g_{\mu\nu} x_2^\nu$$

In particular, since space-time positions have superscripts, the metric matrix G needs to be assigned subscripts. That maintains the convention that a summation index appears once as a subscript and once as a superscript.

Since dot products are invariant,

$$\vec{r}_{1\mathrm{A}}^{\,\mathrm{T}} G \vec{r}_{2\mathrm{A}} = \vec{r}_{1\mathrm{B}}^{\,\mathrm{T}} G \vec{r}_{2\mathrm{B}} = \vec{r}_{1\mathrm{A}}^{\,\mathrm{T}} \Lambda^T G \Lambda \vec{r}_{2\mathrm{A}}$$

Here the final equality substituted the Lorentz transformation from A to B. Recall that if you take a transpose of a product, the order of the factors gets inverted. If the expression to the far left is always equal to the one to the far right, it follows that

$$\boxed{\Lambda^{\mathrm{T}} G \Lambda = G} \tag{A.13}$$

This must be true for any Lorentz transform. In fact, many sources *define* Lorentz transforms as transforms that satisfy the above relationship. Therefore, this relationship will be called the "defining relation." It is very convenient for doing the various mathematics. However, this sort of abstract definition does not really promote easy physical understanding.

And there are a couple of other problems with the defining relation. For one, it allows Lorentz transforms in which one observer uses a left-handed coordinate system instead of a right-handed one. Such an observer observes a mirror image of the universe. Mathematically at least. A Lorentz transform that switches from a normal right-handed coordinate system to a left handed one, (or vice-versa), is called "improper." The simplest example of such an improper transformation is $\Lambda = -G$. That is called the "parity transformation." Its effect is to flip over all spatial position vectors. (If you make a picture of

it, you can see that inverting the directions of the x, y, and z axes of a right-handed coordinate system produces a left-handed system.) To see that $\Lambda = -G$ satisfies the defining relation above, note that G is symmetric, $G^{\mathrm{T}} = G$, and its own inverse, $GG = I$.

Another problem with the defining relation is that it allows one observer to use an inverted direction of time. Such an observer observes the universe evolving to smaller values of her time coordinate. A Lorentz transform that switches the direction of time from one observer to the next is called "nonorthochronous." (Ortho indicates correct, and chronous time.) The simplest example of a non-orthochronous transformation is $\Lambda = G$. That transformation is called "time-reversal." Its effect is to simply replace the time t by $-t$. It satisfies the defining relation for the same reasons as the parity transformation.

As a result, there are four types of Lorentz transformations that satisfy the defining relation. First of all there are the normal proper orthochronous ones. The simplest example is the unit matrix I, corresponding to the case that the observers A and B are identical. Second, there are the improper ones like $-G$ that switch the handedness of the coordinate system. Third there are the nonorthochronous ones like G that switch the correct direction of time. And fourth, there are improper nonorthochronous transforms, like $-GG = -I$, that switch both the handedness and the direction of time.

These four types of Lorentz transforms form four distinct groups. You cannot gradually change from a right-handed coordinate system to a left-handed one. Either a coordinate system is right-handed or it is left-handed. There is nothing in between. By the same token, either a coordinate system has the proper direction of time or the exactly opposite direction.

These four groups are reflected in mathematical properties of the Lorentz transforms. Lorentz transform matrices have determinants that are either 1 or -1. That is easily seen from taking determinants of both sides of the defining equation (A.13), splitting the left determinant in its three separate factors. Also, Lorentz transforms have values of the entry $\lambda^0{}_0$ that are either greater or equal to 1 or less or equal to -1. That is readily seen from writing out the ${}^0{}_0$ entry of (A.13).

Proper orthochronous Lorentz transforms have a determinant 1 and an entry $\lambda^0{}_0$ greater or equal to 1. That can readily be checked for the simplest example $\Lambda = I$. More generally, it can easily be checked that $\lambda^0{}_0$ is the time dilatation factor for events that happen right in the hands of observer A. That is the physical reason that $\lambda^0{}_0$ must always be greater or equal to 1. Transforms that have $\lambda^0{}_0$ less or equal to -1 flip over the correct direction of time. So they are nonorthochronous. Transforms that switch over the handedness of the coordinate system produce a negative determinant. But so do nonorthochronous transforms. If a transform flips over both handedness and the direction of time, it has a time dilatation less or equal to -1 but a positive determinant.

For reasons given above, if you start with some proper orthochronous Lorentz

transform like $\Lambda = I$ and gradually change it, it stays proper and orthochronous. But in addition its determinant stays 1 and its time-dilatation entry stays greater or equal to 1. The reasons are essentially the same as before. You cannot gradually change from a value of 1 or above to a value of -1 or below if there is nothing in between.

One consequence of the defining relation (A.13) merits mentioning. If you premultiply both sides of the relation by G^{-1}, you immediately see that

$$\boxed{\Lambda^{-1} = G^{-1}\Lambda^T G} \tag{A.14}$$

This is the easy way to find inverses of Lorentz transforms. Also, since $G^2 = I$, $G^{-1} = G$. However, it cannot hurt to leave the expression as written. There are other applications in tensor algebra in which G^{-1} is not equal to G.

As already illustrated above, what multiplications by G do is flip over the sign of some entries. So to find an inverse of a Lorentz transform, just flip over the right entries. To be precise, flip over the entries in which one index is 0 and the other is not.

The above observations can be readily converted to tensor notation. First an equivalent is needed to some definitions used in tensor algebra but not normally in linear algebra. The " lowered covector" to a contravariant vector like position will be defined as

$$\tilde{\vec{r}}^{\,\mathrm{L}} \equiv \tilde{\vec{r}}^{\,\mathrm{T}} G$$

In words, take a transpose and postmultiply with the metric G. The result is a row vector while the original is a column vector.

Note that the dot product can now be written as

$$\vec{r}_1 \cdot \vec{r}_2 = \tilde{\vec{r}}_1^{\,\mathrm{L}} \vec{r}_2$$

Note also that lowered covectors are covariant vectors; they are row vectors that transform with the inverse Lorentz transform. To check that, simply plug in the Lorentz transformation of the original vector and use the expression for the inverse Lorentz transform above.

Similarly, the "raised contravector" to a covariant vector like a gradient will be defined as

$$\left(\tilde{\vec{\nabla}} f\right)^{\mathrm{TR}} \equiv G^{-1} \tilde{\vec{\nabla}} f$$

In words, take a transpose and premultiply by the inverse metric. The raised contravector is a contravariant vector. Forming a raised contravector of a lowered covector gives back the original vector. And vice-versa. (Note that metrics are symmetric matrices in checking that.)

In tensor notation, the lowered covector is written as

$$x_\mu = x^\nu g_{\nu\mu}$$

Note that the graphical effect of multiplying by the metric tensor is to "lower" the vector index.

Similarly, the raised contravector to a covector is

$$\partial^\mu f = \left(g^{-1}\right)^{\mu\nu} \partial_\nu f$$

It shows that the inverse metric can be used to "raise" indices. But do not forget the golden rule: raising or lowering an index is much more than cosmetic: you produce a fundamentally different vector.

(That is not true for so-called "Cartesian tensors" like purely spatial position vectors. For these the metric G is the unit matrix. Then raising or lowering an index has no real effect. By the way, the unit matrix is in tensor notation $\delta^\mu{}_\nu$. That is called the Kronecker delta. Its entries are 1 if the two indices are equal and 0 otherwise.)

Using the above notations, the dot product becomes as stated in chapter 1.2.5,

$$x_{1,\mu} x_2^\mu$$

More interestingly, consider the inverse Lorentz transform. According to the expression given above $\Lambda^{-1} = G^{-1}\Lambda^{\mathrm{T}} G$, so:

$$\left(\lambda^{-1}\right)^\mu{}_\nu = \left(g^{-1}\right)^{\mu\alpha} \lambda^\beta{}_\alpha g_{\beta\nu}$$

(A transpose of a matrix, in this case Λ, swaps the indices.) According the index raising/lowering conventions above, in the right hand side the heights of the indices of Λ are inverted. So you can *define* a new matrix with entries

$$\boxed{\lambda_\nu{}^\mu \equiv \left(\lambda^{-1}\right)^\mu{}_\nu}$$

But note that the so-defined matrix is *not* the Lorentz transform matrix:

$$\lambda_\nu{}^\mu \neq \lambda^\mu{}_\nu$$

It is a different matrix. In particular, the signs on some entries are swapped.

(Needless to say, various supposedly authoritative sources list both matrices as λ^μ_ν for that exquisite final bit of confusion. It is apparently not easy to get subscripts and superscripts straight if you use some horrible product like MS Word. Of course, the simple answer would be to use a place holder in the empty position that indicates whether or not the index has been raised or lowered. For example:

$$\lambda^{L\mu}_{\nu R} \neq \lambda^{\mu N}_{N\nu}$$

However, this is not possible because it would add clarity.)

Now consider another very confusing result. Start with

$$G^{-1}GG^{-1} = G^{-1} \qquad \Longrightarrow \qquad \left(g^{-1}\right)^{\mu\alpha} g_{\alpha\beta} \left(g^{-1}\right)^{\beta\nu} = \left(g^{-1}\right)^{\mu\nu}$$

According to the raising conventions, that can be written as

$$\boxed{g^{\mu\nu} \equiv \left(g^{-1}\right)^{\mu\nu}} \tag{A.15}$$

Does this not look exactly as if $G = G^{-1}$? That may be true in the case of Lorentz transforms and the associated Minkowski metric. But for more general applications of tensor algebra it is most definitely *not* true. Always remember the golden rule: names of tensors are only meaningful if the indices are at the right height. The right height for the indices of G is subscripts. So $g^{\mu\nu}$ does *not* indicate an entry of G. Instead it turns out to represent an entry of G^{-1}.

So physicists now have two options. They can write the entries of G^{-1} in the understandable form $(g^{-1})^{\mu\nu}$. Or they can use the confusing, error-prone form $g^{\mu\nu}$. So what do you think they all do? If you guessed option (b), you are making real progress in your study of modern physics.

Often the best way to verify some arcane tensor expression is to convert it to linear algebra. (Remember to check the heights of the indices when doing so. If they are on the wrong height, restore the omitted factor $g_{\cdot\cdot}$ or $(g^{-1})^{\cdot\cdot}$.) Some additional results that are useful in this context are

$$\Lambda^{-\mathrm{T}} G \Lambda^{-1} = G \qquad \Lambda G^{-1} \Lambda^{\mathrm{T}} = G^{-1} \qquad \Lambda G \Lambda^{\mathrm{T}} = G$$

The first of these implies that the inverse of a Lorentz transform is a Lorentz transform too. That is readily verified from the defining relation (A.13) by premultiplying by $\Lambda^{-\mathrm{T}}$ and postmultiplying by Λ^{-1}. The second expression is simply the matrix inverse of the first. Both of these expressions generalize to any symmetric metric G. The final expression implies that the transpose of a Lorentz transform is a Lorentz transform too. That is only true for Lorentz transforms and the associated Minkowski metric. Or actually, it is also true for any other metric in which $G^{-1} = G$, including Cartesian tensors. For these metrics, the final expression above is the same as the second expression.

A.5 The reduced mass

Two-body systems, like the earth-moon system of celestial mechanics or the proton-electron hydrogen atom of quantum mechanics, can be analyzed more simply using reduced mass. In this note both a classical and a quantum derivation will be given. The quantum derivation will need to anticipate some results on multi-particle systems from chapter 5.1.

In two-body systems the two bodies move around their combined center of gravity. However, in examples such as the ones mentioned, one body is much more massive than the other. In that case the center of gravity almost coincides with the heavy body, (earth or proton). Therefore, in a naive first approximation it may be assumed that the heavy body is at rest and that the

lighter one moves around it. It turns out that this naive approximation can be made exact by replacing the mass of the lighter body by an reduced mass. That simplifies the mathematics greatly by reducing the two-body problem to that of a single one. Also, it now produces the exact answer regardless of the ratio of masses involved.

The classical derivation is first. Let m_1 and $\vec{r}_1$ be the mass and position of the massive body (earth or proton), and m_2 and $\vec{r}_2$ those of the lighter one (moon or electron). Classically the force $\vec{F}$ between the masses will be a function of the difference $\vec{r}_{21} = \vec{r}_2 - \vec{r}_1$ in their positions. In the naive approach the heavy mass is assumed to be at rest at the origin. Then $\vec{r}_{21} = \vec{r}_2$, and so the naive equation of motion for the lighter mass is, according to Newton's second law,

$$m_2 \ddot{\vec{r}}_{21} = \vec{F}(\vec{r}_{21})$$

Now consider the true motion. The center of gravity is defined as a mass-weighted average of the positions of the two masses:

$$\vec{r}_{\rm cg} = w_1 \vec{r}_1 + w_2 \vec{r}_2 \qquad w_1 = \frac{m_1}{m_1 + m_2} \quad w_2 = \frac{m_2}{m_1 + m_2}$$

It is shown in basic physics that the net external force on the system equals the total mass times the acceleration of the center of gravity. Since in this case it will be assumed that there are no external forces, the center of gravity moves at a constant velocity. Therefore, the center of gravity can be taken as the origin of an inertial coordinate system. In that coordinate system, the positions of the two masses are given by

$$\vec{r}_1 = -w_2 \vec{r}_{21} \qquad \vec{r}_2 = w_1 \vec{r}_{21}$$

because the position $w_1\vec{r}_1 + w_2\vec{r}_2$ of the center of gravity must be zero in this system, and the difference $\vec{r}_2 - \vec{r}_1$ must be $\vec{r}_{21}$. (Note that the sum of the two weight factors is one.) Solve these two equations for $\vec{r}_1$ and $\vec{r}_2$ and you get the result above.

The true equation of motion for the lighter body is $m_2 \ddot{\vec{r}}_2 = \vec{F}(\vec{r}_{21})$, or plugging in the above expression for $\vec{r}_2$ in the center of gravity system,

$$m_2 w_1 \ddot{\vec{r}}_{21} = \vec{F}(\vec{r}_{21})$$

That is exactly the naive equation of motion if you replace m_2 in it by the reduced mass $m_2 w_1$, i.e. by

$$\boxed{m_{\rm red} = \frac{m_1 m_2}{m_1 + m_2}} \tag{A.16}$$

The reduced mass is almost the same as the lighter mass if the difference between the masses is large, like it is in the cited examples, because then m_2 can be ignored compared to m_1 in the denominator.

The bottom line is that the motion of the two-body system consists of the motion of its center of gravity plus motion around its center of gravity. The motion around the center of gravity can be described in terms of a single reduced mass moving around a fixed center.

The next question is if this reduced mass idea is still valid in quantum mechanics. Quantum mechanics is in terms of a wave function ψ that for a two-particle system is a function of both $\vec{r}_1$ and $\vec{r}_2$. Also, quantum mechanics uses the potential $V(\vec{r}_{21})$ instead of the force. The Hamiltonian eigenvalue problem for the two particles is:

$$H\psi = E\psi \qquad H = -\frac{\hbar^2}{2m_1}\nabla_1^2 - \frac{\hbar^2}{2m_2}\nabla_2^2 + V(\vec{r}_{21})$$

where the two kinetic energy Laplacians in the Hamiltonian H are with respect to the position coordinates of the two particles:

$$\nabla_1^2\psi \equiv \sum_{j=1}^{3} \frac{\partial^2\psi}{\partial r_{1,j}^2} \qquad \nabla_2^2\psi \equiv \sum_{j=1}^{3} \frac{\partial^2\psi}{\partial r_{2,j}^2}$$

Now make a change of variables from $\vec{r}_1$ and $\vec{r}_2$ to $\vec{r}_{\rm cg}$ and $\vec{r}_{21}$ where

$$\vec{r}_{\rm cg} = w_1\vec{r}_1 + w_2\vec{r}_2 \qquad \vec{r}_{21} = \vec{r}_2 - \vec{r}_1$$

The derivatives of ψ can be converted using the chain rule of differentiation:

$$\frac{\partial\psi}{\partial r_{1,j}} = \frac{\partial\psi}{\partial r_{{\rm cg},j}}\frac{\partial r_{{\rm cg},j}}{\partial r_{1,j}} + \frac{\partial\psi}{\partial r_{21,j}}\frac{\partial r_{21,j}}{\partial r_{1,j}} = \frac{\partial\psi}{\partial r_{{\rm cg},j}}w_1 - \frac{\partial\psi}{\partial r_{21,j}}$$

or differentiating once more and summing

$$\nabla_1^2\psi = \sum_{j=1}^{3} \frac{\partial^2\psi}{\partial r_{1,j}^2} = w_1^2 \sum_{j=1}^{3} \frac{\partial^2\psi}{\partial r_{{\rm cg},j}^2} - 2w_1 \sum_{j=1}^{3} \frac{\partial^2\psi}{\partial r_{{\rm cg},j}\partial r_{21,j}} + \sum_{j=1}^{3} \frac{\partial^2\psi}{\partial r_{21,j}^2}$$

and a similar expression for $\nabla_2^2\psi$, but with w_2 instead of w_1 and a plus sign instead of the minus sign. Combining them together in the Hamiltonian, and substituting for w_1 and w_2, the mixed derivatives drop out against each other and what is left is

$$H = -\frac{\hbar^2}{2(m_1+m_2)}\nabla_{\rm cg}^2 - \frac{\hbar^2}{2m_{\rm red}}\nabla_{21}^2 + V(\vec{r}_{21})$$

The first term is the kinetic energy that the total mass would have if it was at the center of gravity; the next two terms are kinetic and potential energy around the center of gravity, in terms of the distance between the masses and the reduced mass.

The Hamiltonian eigenvalue problem $H\psi = E\psi$ has separation of variables solutions of the form

$$\psi = \psi_{\rm cg}(\vec{r}_{\rm cg})\psi_{21}(\vec{r}_{21})$$

Substituting this and the Hamiltonian above into $H\psi = E\psi$ and dividing by $\psi_{\rm cg}\psi_{21}$ produces

$$-\frac{\hbar^2}{2(m_1+m_2)}\frac{1}{\psi_{\rm cg}}\nabla^2_{\rm cg}\psi_{\rm cg} + \frac{1}{\psi_{21}}\left[-\frac{\hbar^2}{2m_{\rm red}}\nabla^2_{21} + V\right]\psi_{21} = E$$

Call the first term in the left hand side $E_{\rm cg}$ and the second E_{21}. By that definition, $E_{\rm cg}$ would normally be a function of $\vec{r}_{\rm cg}$, because $\psi_{\rm cg}$ is, but since it is equal to $E - E_{21}$ and those do not depend on $\vec{r}_{\rm cg}$, $E_{\rm cg}$ cannot either, and must be a constant. By similar reasoning, E_{21} cannot depend on $\vec{r}_{21}$ and must be a constant too. Therefore, rewriting the definitions of $E_{\rm cg}$ and E_{21}, two separate eigenvalue problems are obtained:

$$-\frac{\hbar^2}{2(m_1+m_2)}\nabla^2_{\rm cg}\psi_{\rm cg} = E_{\rm cg}\psi_{\rm cg} \qquad \left[-\frac{\hbar^2}{2m_{\rm red}}\nabla^2_{21} + V\right]\psi_{21} = E_{21}\psi_{21}$$

The first describes the quantum mechanics of an imaginary total mass m_1+m_2 located at the center of gravity. The second describes an imaginary reduced mass $m_{\rm red}$ at a location $\vec{r}_{21}$ away from a fixed center that experiences a potential $V(\vec{r}_{21})$.

For the hydrogen atom, it means that if the problem with a stationary proton is solved using an reduced electron mass $m_{\rm p}m_{\rm e}/(m_{\rm p}+m_{\rm e})$, it solves the true problem in which the proton moves a bit too. Like in the classical analysis, the quantum analysis shows that in addition the atom can move as a unit, with a motion described in terms of its center of gravity.

It can also be concluded, from a slight generalization of the quantum analysis, that a constant external gravity field, like that of the sun on the earth-moon system, or of the earth on a hydrogen atom, causes the center of gravity to accelerate correspondingly, but does not affect the motion around the center of gravity at all. That reflects a basic tenet of general relativity.

A.6 Constant spherical potentials

This addendum describes the solutions of the Hamiltonian eigenvalue problem in spherical coordinates if the potential is constant.

These solutions are important for many reasons. For example, you might want to create a simplified model for the hydrogen atom that way. To do so, you could, for example, assume that the potential energy has a constant negative value up to say the Bohr radius and is zero beyond it. That is not really a very good model for the hydrogen atom. However, it works much better for nucleons in atomic nuclei, chapter 14.12.

The solutions in this note are also important for describing experiments in which particles are scattered from some target, {A.30}. And more fundamentally, they give the energy states of definite angular momentum for particles in empty space.

A.6.1 The eigenvalue problem

The Hamiltonian eigenvalue problem is

$$-\frac{\hbar^2}{2m}\nabla^2\psi + V\psi = E\psi$$

In this note it is assumed that the potential V is a constant in the radial region of interest.

To clean the problem up a bit, take the potential energy term to the other side, and also the coefficient of the Laplacian. That produces

$$-\nabla^2\psi = \frac{p_{\rm c}^2}{\hbar^2}\psi$$

where

$$p_{\rm c} \equiv \sqrt{2m(E-V)}$$

The constant $p_{\rm c}$ is what classical physics would take to be the linear momentum of a particle with total energy E and potential energy V.

A.6.2 The eigenfunctions

Because the potential is spherically symmetric like for the hydrogen atom, the eigenfunctions are of similar form:

$$\boxed{\psi_{Elm}(r,\theta,\phi) = R_{El}(r)Y_l^m(\theta,\phi)} \tag{A.17}$$

Here the Y_l^m are again the spherical harmonics. These eigenfunctions have definite square angular momentum $l(l+1)\hbar^2$ where l is an nonnegative integer. They also have definite angular momentum in the chosen z-direction equal to $m\hbar$, where m is an integer that satisfies $|m| \leqslant l$.

The radial functions R_{El} in the eigenfunctions ψ_{Elm} are different from those of the hydrogen atom. Depending on whatever is easiest in a given application, they can be written in two ways, {D.16}.

The first way is as

$$\boxed{R_{El} = c_{\rm s} j_l(p_{\rm c}r/\hbar) + c_{\rm c} n_l(p_{\rm c}r/\hbar) \qquad p_{\rm c} \equiv \sqrt{2m(E-V)}} \tag{A.18}$$

Here $c_{\rm s}$ and $c_{\rm c}$ are arbitrary constants. The functions j_l and n_l are called the "spherical Bessel functions" of the first and second kinds. The n_l are also called the "Neumann functions" and might instead be indicated by y_l or η_l.

Expressions for these Bessel functions can be found in advanced mathematical handbooks, [1]:

$$\boxed{j_l(x) = (-x)^l \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l \frac{\sin x}{x}} \tag{A.19}$$

$$\boxed{n_l(x) = -(-x)^l \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l \frac{\cos x}{x}} \tag{A.20}$$

The spherical Bessel functions are often convenient in a region of constant potential that includes the origin, because the Bessel functions of the first kind j_l give the solutions that are finite at the origin. (To see that, note that the Taylor series of $\sin x$ divided by x is a power series in x^2, and that $x\mathrm{d}x = \frac{1}{2}\mathrm{d}x^2$.) In particular for small x:

$$\boxed{j_l(x) = \frac{2^l l!}{(2l+1)!}x^l + O(x^{l+2}) \equiv \frac{x^l}{(2l+1)!!} + O(x^{l+2})} \tag{A.21}$$

Here !! is one of these unfortunate notations. The second ! means that all even factors are dropped from the factorial.

The Bessel functions of the second kind are singular at the origin and normally do not appear if the origin is part of the considered region.

Also, the spherical Bessel functions are real for real x. However, in a region where the potential V is larger than the energy E of the particles, the argument of the Bessel functions in (A.18) will be imaginary.

The other way to write the radial functions is as

$$\boxed{R_{El} = c_\mathrm{f} h_l^{(1)}(p_\mathrm{c} r/\hbar) + c_\mathrm{b} h_l^{(2)}(p_\mathrm{c} r/\hbar) \qquad p_\mathrm{c} \equiv \sqrt{2m(E-V)}}$$

where $h_l^{(1)}$ and $h_l^{(2)}$ are called the "spherical Hankel functions."

The spherical Hankel functions can again be found in advanced mathematical handbooks, [1]:

$$\boxed{h_l^{(1)}(x) = -\mathrm{i}(-x)^l \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l \frac{e^{\mathrm{i}x}}{x} = j_l(x) + \mathrm{i}n_l(x)} \tag{A.22}$$

$$\boxed{h_l^{(2)}(x) = \mathrm{i}(-x)^l \left(\frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}x}\right)^l \frac{e^{-\mathrm{i}x}}{x} = j_l(x) - \mathrm{i}n_l(x)} \tag{A.23}$$

The given expressions in terms of the spherical Bessel functions are readily inverted to give the Bessel functions in terms of the Hankel functions,

$$j_l(x) = \frac{h_l^{(1)}(x) + h_l^{(2)}(x)}{2} \qquad n_l(x) = \frac{h_l^{(1)}(x) - h_l^{(2)}(x)}{2\mathrm{i}} \tag{A.24}$$

For large x the spherical Hankel functions can be approximated as

$$\boxed{h_l^{(1)}(x) \sim (-i)^{l+1}\frac{e^{\mathrm{i}x}}{x} \qquad h_l^{(2)}(x) \sim i^{l+1}\frac{e^{-\mathrm{i}x}}{x}} \tag{A.25}$$

This asymptotic behavior tends to make the Hankel functions more convenient far from the origin. Exponentials are mathematically simpler and more fundamental than the sines and cosines in the asymptotic behavior of the Bessel functions.

A.6.3 About free space solutions

The most important case for which the energy eigenfunctions of the previous subsection apply is for particles in empty space. They describe energy states with definite square and z angular momentum. However, sometimes particles in empty space are better described by states of definite linear momentum. And in some cases, like in scattering problems, you need both types of solution. Then you also need to convert between them.

The energy states in empty space with definite square and z angular momentum are

$$\boxed{\psi_{Elm} = j_l(p_{\mathrm{c}}r/\hbar)Y_l^m(\theta,\phi) \qquad p_{\mathrm{c}} \equiv \sqrt{2mE}} \tag{A.26}$$

These states have square angular momentum $l(l+1)\hbar^2$ and angular momentum in the chosen z-direction $m\hbar$. They are nonsingular at the origin.

A state that has definite linear momentum p_{c} purely in the z-direction has an energy eigenfunction

$$\boxed{\psi_{\vec{k}} = e^{\mathrm{i}p_{\mathrm{c}}z/\hbar} \qquad p_{\mathrm{c}} \equiv \sqrt{2mE}} \tag{A.27}$$

This eigenfunction is not normalized, and cannot be normalized. However, neither can the eigenfunction ψ_{Elm} above be. It is the curse of eigenfunctions in infinite empty space. An introduction to the adjustments that must be made to deal with this is given in chapter 7.9.

It is sometimes necessary to write a linear momentum eigenfunction of the form (A.27) in terms of angular momentum ones of the form (A.26). Rayleigh worked out the correct combination a very long time ago, {D.16}:

$$\boxed{e^{\mathrm{i}p_{\mathrm{c}}z/\hbar} = \sum_{l=0}^{\infty} c_{\mathrm{w},l}\, j_l(p_{\mathrm{c}}r/\hbar)Y_l^0(\theta) \qquad c_{\mathrm{w},l} = \mathrm{i}^l\sqrt{4\pi(2l+1)}} \tag{A.28}$$

Note that only eigenfunctions with $m = 0$ are needed.

A.7 Accuracy of the variational method

This note has a closer look at the accuracy of the variational method.

Any approximate ground state wave function ψ may always be written as a combination of all the energy eigenfunctions ψ_1, ψ_2, ...:

$$\psi = c_1\psi_1 + \delta_2\psi_2 + \delta_3\psi_3 + \ldots$$

where c_1 and the δ_i for $i = 2, 3, \ldots$ are numerical coefficients. But if the approximation is any good at all, the coefficient c_1 of the correct ground state ψ_1 must be close to one, while the coefficients δ_i of the higher energy states must be small.

The wave function pollution with higher energy states can be related to the error in energy, call it ε, using a few simple manipulations. First the condition that ψ is normalized, $\langle\psi|\psi\rangle = 1$, works out to be

$$1 = \langle c_1\psi_1 + \delta_2\psi_2 + \ldots | c_1\psi_1 + \delta_2\psi_2 + \ldots\rangle = c_1^2 + \delta_2^2 + \delta_3^2 + \ldots$$

since the eigenfunctions $\psi_1, \psi_2, \ldots$ are orthonormal. Similarly, the expectation energy $\langle E\rangle = \langle\psi|H\psi\rangle$ of the approximate solution works out to be

$$\langle E\rangle = \langle c_1\psi_1 + \delta_2\psi_2 + \ldots | E_1c_1\psi_1 + E_2\delta_2\psi_2 + \ldots\rangle = c_1^2E_1 + \delta_2^2E_2 + \delta_3^2E_3 + \ldots$$

Multiplying the normalization condition by E_1 and subtracting it from the expression for the expectation energy above gives the error in energy as:

$$\varepsilon \equiv \langle E\rangle - E_1 = \delta_2^2(E_2 - E_1) + \delta_3^2(E_3 - E_1) + \ldots$$

Note first that since all the terms in the right hand side are positive, any approximate wave function has more expectation energy than the ground state E_1. It does not have to be a single energy eigenfunction of higher energy. But that should not be a surprise.

Nor is it surprising that the expression above shows that the error in energy ε will be small if the coefficients δ_i of the incorrect energy eigenfunctions are small and decrease suitably in magnitude when i increases.

However, note that while the errors in wave function are directly proportional to the coefficients δ_i, the error in energy is proportional to the *squares* of these coefficients. That makes the error in energy unexpectedly small, because the square of any small quantity is much smaller still. (This assumes that the term "small" is defined in a meaningful nondimensional way.)

That small error in energy is great because the computed energy is important for a number of things, like determining whether a stable ground state of the supposed form exists in the first place, and if it does, how fast it interacts with other energy eigenfunctions if there is uncertainty in energy, chapter 7.

While it may seem obvious that if the approximate wave function is close to the correct one, then the approximate energy will be close to the correct one,

the reverse is less trivial. If the approximate energy is close to the exact energy, does that necessarily mean that *entire wave function* is close to the exact one? Fortunately, the answer to that question is usually yes.

In particular, note from the expression for the error in energy above that for any coefficient δ_i

$$\delta_i \leq \sqrt{\frac{\varepsilon}{E_i - E_1}}$$

even in the worst-case scenario that all the error is in the i-th term. From the above, the amount δ_i of each polluting higher-energy eigenfunction function is small if ε is small.

But do also note the effect of the denominator. If it too is small, it may increase the possible error. The worst case occurs for the second lowest energy state. If the second-lowest energy E_2 is very close to the ground-state energy E_1, unusual good accuracy in energy may be required to ensure that the approximate wave function is accurate. (However, if E_2 equals the ground state energy, the second state is a ground state too; the ground state is then no longer unique. In that case the error from *some* valid ground state is described by the third energy state, not the second.)

Consider also the magnitude of the error in the approximate wave function. It is defined as

$$||\delta_2\psi_2 + \delta_3\psi_3 + \ldots|| = \sqrt{\delta_2^2 + \delta_3^2 + \ldots}$$

This can be related to the error in energy by noting that from its given expression

$$\varepsilon \geqslant \delta_2^2(E_2 - E_1) + \delta_3^2(E_2 - E_1) + \ldots$$

since $E_i - E_1$ is at least as big as $E_2 - E_1$. Comparing the expressions above shows that

$$||\delta_2\psi_2 + \delta_3\psi_3 + \ldots|| \leqslant \sqrt{\frac{\varepsilon}{E_2 - E_1}}$$

So if the error in energy ε is small, the magnitude of the error in the wave function is too.

The bottom line is that the lower you can get your expectation energy, the closer you will get to the true ground state energy. In addition the small error in energy will reflect in a small error in wave function too.

A.8 Positive ground state wave function

This addendum discusses why in at least the simplest cases a ground state wave function can be assumed to be real, positive, and unique (i.e. nondegenerate). It is assumed that the potential is a real function of position. That is true for the hydrogen molecular ion. It is also true for a single hydrogen atom and most

other simple systems, at least in the nonrelativistic approximations normally used in this book.

It should first be noted that if potentials are allowed that are positive infinity in a finite region, nonunique ground states that cannot be taken positive may in fact possible. Such a potential can provide an impenetrable boundary, completely separating one region of space from another. In that case the ground state wave functions at the two sides of the boundary become decoupled, allowing for indeterminacy in the combined ground state. Such artificial cases are not covered here. But you can readily find examples in lots of books on quantum mechanics, especially in one dimension. Here it will be assumed that the potentials stay away from positive infinity. For practical purposes, it may also be noted that if the potential becomes positive infinity at just a few points, it is usually not a problem unless the approach to singularity is very steep.

There is however a much more important restriction to the conclusions in this note: ground states may not be positive if you go to many-particle systems. That is discussed further in the final paragraphs of this addendum.

First consider why the ground state, and any other energy eigenstate, can be assumed to be real without loss of generality. Suppose that you had a complex eigenfunction ψ for the eigenvalue problem $H\psi = E\psi$. Write the eigenfunction as

$$\psi = \psi_{\rm r} + {\rm i}\psi_{\rm i}$$

where the real part $\psi_{\rm r}$ and the imaginary part $\psi_{\rm i}$ are real functions. Plugging this into the eigenvalue problem, you see that the real and imaginary parts each separately must satisfy the eigenvalue problem. (For the complex number $H\psi - E\psi$ to be zero, both its real and imaginary parts have to be zero.) So each of $\psi_{\rm r}$ and $\psi_{\rm i}$ is just as good an eigenfunction as ψ and each is real. Since the original complex ψ is a linear combination of the two, you do not need it separately. (If either $\psi_{\rm r}$ or $\psi_{\rm i}$ is zero, it is not an eigenfunction, but then it is not needed to describe ψ either. And if it is nonzero, it can be normalized.)

Next consider why the ground state can be taken to be positive, assuming, for now, that it is unique. What characterizes the ground state $\psi_{\rm gs}$ is that it has the lowest possible expectation value of the energy. The expectation energy can be written for arbitrary, but normalized, wave functions ψ as

$$\langle E \rangle = \frac{\hbar^2}{2m} \int_{\rm all} (\nabla\psi)^2 \, {\rm d}^3\vec{r} + \int_{\rm all} V\psi^2 \, {\rm d}^3\vec{r}$$

In the first term, the kinetic energy, integrations by part have been used to get rid of the second order derivatives. Note that $(\nabla\psi)^2$ stands for the sum of the square partial derivatives of ψ. Now by definition $\psi_{\rm gs}$ has the lowest possible value of the above expectation energy among all possible normalized functions ψ. But only terms square in ψ appear in the expectation energy. So $|\psi_{\rm gs}|$ has the same expectation energy. That means that $|\psi_{\rm gs}|$ is a ground state wave function

too. Under the given assumption that the ground state is unique, $|\psi_{\rm gs}|$ can be taken to be *the* ground state. That makes the ground state positive. (Note that a constant of magnitude one does not make a difference in an eigenfunction. So the original $\psi_{\rm gs}$ might well have been equal to $-|\psi_{\rm gs}|$. But $|\psi_{\rm gs}|$ is equivalent to that.)

Finally, it needs to be shown that the ground state is indeed unique as assumed above. That is really messy, so it has been banned to derivation {D.22}. It is based on the same idea that the absolute value of a ground state is a ground state too.

Regrettably the arguments above stop working for more than two electrons. To really understand the reason, you will first need to read chapter 5.6 on multiple-particle systems. But in a nutshell, the wave function for systems with multiple electrons must satisfy additional requirements, called the "anti-symmetrization" requirements. These requirements normally turn $|\psi_{\rm gs}|$ into an unacceptable wave function. Then obviously the above arguments fall apart. Fortunately, for just two electrons, there is a loophole in the requirement called "spin." That allows the hydrogen molecule, with two electrons, still to be covered.

The same problem occurs for atomic nuclei that contain multiple protons and/or neutrons. (For atomic nuclei, the potentials also tend to be far more complicated than assumed here. But that is another matter.) In general, particles for which antisymmetrization requirements apply are called fermions.

There is however a different class of particles called "bosons." For those, the wave function has to satisfy "symmetrization" requirements. Symmetrization requirements are still OK if you replace ψ by $|\psi|$. So the ideas above are helpful for understanding the ground state of large numbers of bosons. For example, they are helpful in understanding the superfluidity of liquid helium near its ground state, [18, pp. 321-323]. Complete helium atoms are bosons.

A.9 Wave function symmetries

Symmetries are very important in physics. For example, symmetries in wave functions are often quite helpful to understand the physics qualitatively.

As an example, the hydrogen molecular ion is mirror symmetric around its midplane. This midplane is the plane halfway in between the two nuclei, orthogonal to the line connecting them. To roughly understand what the mirror symmetry around this plane means, think of the midplane as an infinitely thin mirror. Take this mirror to be two-sided, so that you can look in it from either side. That allows you to see the mirror image of each side of the molecule. Simply put, the mirror symmetry of the ion means that the mirror image looks exactly the same as the original ion.

(If you would place the entire molecule at one side of the mirror, its entire mirror image would be at the other side of it. But except for this additional shift in location, everything would remain the same as in the case assumed here.)

Under the same terms, human beings are roughly mirror symmetric around the plane separating their left and right halves. But that symmetry is far from perfect. For example, if you part your hair at one side, your mirror image parts it at the other side. And your heart changes sides too.

To describe mirror symmetry more precisely, take the line through the nuclei to be the z-axis. And take z to be zero at the mirror. Then all that the mirror does mathematically is replace z by $-z$. For example, the mirror image of the nucleus at positive z is located at the corresponding negative z value. And vice-versa.

The effect of mirroring on any molecular wave function Ψ can be represented by a "mirror operator" $\mathcal{M}$. According to the above, all this operator does is replace z by $-z$:

$$\mathcal{M}\Psi(x,y,z) = \Psi(x,y,-z)$$

By definition a wave function is mirror symmetric if the mirror operator has no effect on it. Mathematically, if the mirror operator does not do anything, then $\mathcal{M}\Psi$ must be the same as Ψ. So mirror symmetry requires

$$\mathcal{M}\Psi(x,y,z) \equiv \Psi(x,y,-z) = \Psi(x,y,z)$$

The final equality above shows that a mirror-symmetric wave function is the same at positive values of z as at the corresponding negative values. Mathematicians might simply say that the wave function is symmetric around the xy-plane, (i.e. the mirror). The ground state $\psi_{\rm gs}$ of the molecular ion is mirror symmetric in this sense. The big question to be addressed in this addendum is, why?

The fundamental reason why the ion is mirror symmetric is a mathematical one. The mirror operator $\mathcal{M}$ commutes with the Hamiltonian H. Recall from chapter 4.5.1 what this means:

$$\mathcal{M}H = H\mathcal{M}$$

In words, it does not make a difference in which order you apply the two operators.

That can be seen from the physics. The Hamiltonian consists of potential energy V and kinetic energy T. Now it does not make a difference whether you multiply a wave function value by the potential before or after you flip the value over to the opposite z-position. The potential is the same at opposite z values, because the nuclei at the two sides of the mirror are the same. As far as the kinetic energy is concerned, if it involved a first-order z-derivative, there would be a change of sign when you flip over the sign of z. But the kinetic energy has

only a second order z-derivative. A second order derivative does not change. So all together it makes no difference whether you first mirror and then apply the Hamiltonian or vice-versa. The two operators commute.

Also according to chapter 4.5.1, that has a consequence. It implies that you can take energy eigenfunctions to be mirror eigenfunctions too. And the ground state is an energy eigenfunction. So it can be taken to be an eigenfunction of $\mathcal{M}$ too:

$$\mathcal{M}\psi_{\rm gs}(x,y,z) = \lambda\psi_{\rm gs}(x,y,z)$$

Here λ is a constant called the eigenvalue. But what would this eigenvalue be?

To answer that, apply $\mathcal{M}$ *twice*. That multiplies the wave function by the square eigenvalue. But if you apply $\mathcal{M}$ twice, you always get the original wave function back, because $-(-z) = z$. So the square eigenvalue λ^2 must be 1, in order that the wave function does not change when multiplied by it. That means that the eigenvalue λ itself can be either 1 or -1. So for the ground state wave function $\psi_{\rm gs}$, either

$$\mathcal{M}\psi_{\rm gs}(x,y,z) \equiv \psi_{\rm gs}(x,y,-z) = 1 \times \psi_{\rm gs}(x,y,z)$$

or

$$\mathcal{M}\psi_{\rm gs}(x,y,z) \equiv \psi_{\rm gs}(x,y,-z) = -1 \times \psi_{\rm gs}(x,y,z)$$

If the first possibility applies, the wave function does not change under the mirroring. So by definition it is mirror symmetric. If the second possibility applies, the wave function changes sign under the mirroring. Such a wave function is called "mirror *anti*symmetric." But the second possibility has wave function values of opposite sign at opposite values of z. That is not possible, because the previous addendum showed that the ground state wave function is everywhere positive. So it must be possibility one. That means that the ground state must indeed be mirror symmetric as claimed.

It may be noted that the state of second lowest energy will be antisymmetric. You can see the same thing happening for the eigenfunctions of the particle in a pipe. The ground state figure 3.8, (or 3.11 in three dimensions), is symmetric around the center cross-section of the pipe. The first excited state, at the top of figures 3.9, (or 3.12), is antisymmetric. (Note that the grey tones show the square wave function. If the wave function is antisymmetric, the square wave function is symmetric. But it will be zero at the symmetry plane.)

Next consider the rotational symmetry of the hydrogen molecular ion around the axis through the nuclei. The ground state of the molecular ion does not change if you rotate the ion around the z-axis through the nuclei. That makes it rotationally symmetric. The big question is again, why?

In this case, let $\mathcal{R}_\varphi$ be the operator that rotates a wave function Ψ over an angle φ around the z-axis. This operator too commutes with the Hamiltonian. After all, the only physically meaningful direction is the z-axis through the nuclei. The angular orientation of the xy axes system normal to it is a completely

arbitrary choice. So it should not make a difference at what angle around the z axis you apply the Hamiltonian.

Therefore the ground state must be an eigenfunction of the rotation operator just like it is one of the mirror operator:

$$\mathcal{R}_\varphi \psi_{\rm gs} = \lambda \psi_{\rm gs}$$

But now what is that eigenvalue λ? First note that the magnitude of all eigenvalues of $\mathcal{R}_\varphi$ must be 1. Otherwise the magnitude of the wave function would change correspondingly during the rotation. However, the magnitude of a wave function does not change if you simply rotate it. And if the eigenvalue is a complex number of magnitude 1, then it can always be written as $e^{{\rm i}\alpha}$ where α is some real number. So the rotated ground state is some multiple $e^{{\rm i}\alpha}$ of the original ground state. But the values of the rotated ground state are real and positive just like that of the original ground state. That can only be true if the multiplying factor $e^{{\rm i}\alpha}$ is real and positive too. And if you check the Euler formula (2.5), you see that $e^{{\rm i}\alpha}$ is only real and positive if it is 1. Since multiplying by 1 does not change the wave function, the ground state does not change when rotated. That then makes it rotationally symmetric around the z-axis through the nuclei as claimed.

You might of course wonder about the rotational changes of excited energy states. For those a couple of additional observations apply. First, the number α must be proportional to the rotation angle φ, since rotating Ψ twice is equivalent to rotating it once over twice the angle. That means that, more precisely, the eigenvalues are of the form $e^{{\rm i}m\varphi}$, where m is a real constant independent of φ. Second, rotating the ion over a 2π full turn puts each point back to where it came from. That should reproduce the original wave function. So an eigenvalue $e^{{\rm i}m2\pi}$ for a full turn must be 1. According to the Euler formula, that requires m to be an integer, one of $\ldots$, -2, -1, 0, 1, 2, $\ldots$. For the ground state, m will have to be zero; that is the only way to get $e^{{\rm i}m\varphi}$ equal to 1 for all angles φ. But for excited states, m can be a nonzero integer. In that case, these states do not have rotational symmetry.

Recalling the discussion of angular momentum in chapter 4.2.2, you can see that m is really the magnetic quantum number of the state. Apparently, there is a connection between rotations around the z-axis and the angular momentum in the z-direction. That will be explored in more detail in chapter 7.3.

For the neutral hydrogen molecule discussed in chapter 5.2, there is still another symmetry of relevance. The neutral molecule has two electrons, instead of just one. This allows another operation: you can swap the two electrons. That is called "particle exchange." Mathematically, what the particle exchange operator $\mathcal{P}$ does with the wave function is swap the position coordinates of electron 1 with those of electron 2. Obviously, physically this does not do anything at all; the two electrons are exactly the same. It does not make a

difference which of the two is where. So particle exchange commutes again with the Hamiltonian.

The mathematics of the particle exchange is similar to that of the mirroring discussed above. In particular, if you exchange the particles twice, they are back to where they were originally. From that, just like for the mirroring, it can be seen that swapping the particle positions does nothing to the ground state. So the ground state is symmetric under particle exchange.

It should be noted that the ground state of systems involving three or more electrons is *not* symmetric under exchanging the positions of the electrons. Wave functions for multiple electrons must satisfy special particle-exchange requirements, chapter 5.6. In fact they must be *anti*symmetric under an expanded definition of the exchange operator. This is also true for systems involving three or more protons or neutrons. However, for some particle types, like three or more helium atoms, the symmetry under particle exchange continues to apply. This is very helpful for understanding the properties of superfluid helium, [18, p. 321].

A.10 Spin inner product

In quantum mechanics, the angle between two angular momentum vectors is not really defined. That is because at least two components of a nonzero angular momentum vector are uncertain. However, the inner product of angular momentum vectors can be well defined. In some sense, that gives an angle between the two vectors.

An important case is the inner product between the spins of two particles. It is related to the square net combined spin of the particles as

$$\widehat{S}_{\rm net}^2 = \left(\vec{\widehat{S}}_1 + \vec{\widehat{S}}_2\right) \cdot \left(\vec{\widehat{S}}_1 + \vec{\widehat{S}}_2\right)$$

If you multiply out the right hand side and rearrange, you find the inner product between the spins as

$$\boxed{\vec{\widehat{S}}_1 \cdot \vec{\widehat{S}}_2 = \tfrac{1}{2}\left(\widehat{S}_{\rm net}^2 - \widehat{S}_1^2 - \widehat{S}_2^2\right)} \tag{A.29}$$

Now an elementary particle has a definite square spin angular momentum

$$\widehat{S}^2 = s(s+1)\hbar^2$$

where s is the spin quantum number. If the square combined spin also has a definite value, then so does the dot product between the spins as given above.

As an important example, consider two fermions with spin $s_1 = s_2 = \frac{1}{2}$. These fermions may be in a singlet state with combined spin $s_{\rm net} = 0$. Or they

may be in a triplet state with combined spin $s_{\rm net} = 1$. If that is plugged into the formulae above, the inner product between the spins is found to be

$$\boxed{\text{singlet:}\quad \widehat{\vec{S}}_1 \cdot \widehat{\vec{S}}_2 = -\tfrac{3}{4}\hbar^2 \qquad \text{triplet:}\quad \widehat{\vec{S}}_1 \cdot \widehat{\vec{S}}_2 = \tfrac{1}{4}\hbar^2} \tag{A.30}$$

Based on that, you could argue that in the singlet state the angle between the spin vectors is 180°. In the triplet state the angle is not zero, but about 70°.

A.11 Thermoelectric effects

This note gives additional information on thermoelectric effects.

A.11.1 Peltier and Seebeck coefficient ballparks

The approximate expressions for the semiconductor Peltier coefficients come from [29]. Straub *et al* (App. Phys. Let. 95, 052107, 2009) note that to better approximation, $\frac{3}{2}k_{\rm B}T$ should be $(\frac{5}{2}+r)k_{\rm B}T$ with r typically $-\frac{1}{2}$. Also, a phonon contribution should be added.

The estimate for the Peltier coefficient of a metal assumes that the electrons form a free-electron gas. The conduction will be assumed to be in the x-direction. To ballpark the Peltier coefficient requires the average charge flow per electron $\overline{-ev_x}$ and the average energy flow per electron $\overline{E^{\rm p}v_x}$. Here v_x is the electron velocity in the x-direction, $-e$ the electron charge, $E^{\rm p}$ the electron energy, and an overline is used to indicate an average over all electrons. To find ballparks for the two averages, assume the model of conduction of the free-electron gas as given in chapter 6.20. The conduction occurred since the Fermi sphere got displaced a bit towards the right in the wave number space figure 6.17. Call the small amount of displacement $k_{\rm d}$. Assume for simplicity that in a coordinate system $k_xk_yk_z$ with origin at the center of the *displaced* Fermi sphere, the occupation of the single-particle states by electrons is still exactly given by the equilibrium Fermi-Dirac distribution. However, due to the displacement $k_{\rm d}$ along the k_x axis, the velocities and energies of the single-particle states are now given by

$$v_x = \frac{\hbar}{m}(k_x + k_{\rm d}) \qquad E^{\rm p} = \frac{\hbar^2}{2m}\left[(k_x+k_{\rm d})^2 + k_y^2 + k_z^2\right] = \frac{\hbar^2}{2m}\left(k^2 + 2k_xk_{\rm d} + k_{\rm d}^2\right)$$

To simplify the notations, the above expressions will be abbreviated to

$$v_x = C_v(k_x + k_{\rm d}) \qquad E^{\rm p} = C_E(k^2 + 2k_xk_{\rm d} + k_{\rm d}^2)$$

In this notation, the average charge and energy flows per electron become

$$\overline{-ev_x} = \overline{-eC_v(k_x+k_{\rm d})} \qquad \overline{E^{\rm p}v_x} = \overline{C_E(k^2+2k_xk_{\rm d}+k_{\rm d}^2)C_v(k_x+k_{\rm d})}$$

Next note that the averages involving odd powers of k_x are zero, because for every state of positive k_x in the Fermi sphere there is a corresponding state of negative k_x. Also the constants, including $k_{\rm d}$, can be taken out of the averages. So the flows simplify to

$$\overline{-ev_x} = -eC_v k_{\rm d} \qquad \overline{E^{\rm p} v_x} = C_E(2\overline{k_x^2} + \overline{k^2})C_v k_{\rm d}$$

where the term cubically small in $k_{\rm d}$ was ignored. Now by symmetry the averages of k_x^2, k_y^2, and k_z^2 are equal, so each must be one third of the average of k^2. And C_E times the average of k^2 is the average energy per electron $E^{\rm p}_{\rm ave}$ in the absence of conduction. Also, by definition $C_v k_{\rm d}$ is the drift velocity $v_{\rm d}$ that produces the current. Therefore:

$$\overline{-ev_x} = -ev_{\rm d} \qquad \overline{v_x E^{\rm p}} = \tfrac{5}{3} E^{\rm p}_{\rm ave} v_{\rm d}$$

Note that if you would simply have ballparked the average of $v_x E^{\rm p}$ as the average of v_x times the average of $E^{\rm p}$ you would have missed the factor 5/3. That would produce a Peltier coefficient that would be gigantically wrong.

To get the heat flow, the energy must be taken relative to the Fermi level μ. In other words, the energy flow $\overline{v_x \mu}$ must be subtracted from $\overline{v_x E^{\rm p}}$. The Peltier coefficient is the ratio of that heat flow to the charge flow:

$$\mathcal{P} = \frac{\overline{v_x(E^{\rm p} - \mu)}}{\overline{-ev_x}} = \frac{\frac{5}{3}E^{\rm p}_{\rm ave} - \mu}{-e}$$

If you plug in the expressions for the average energy per electron and the chemical potential found in derivation {D.62}, you get the Peltier ballpark listed in the text.

To get Seebeck coefficient ballparks, simply divide the Peltier coefficients by the absolute temperature. That works because of Kelvin's second relationship discussed below. To get the Seebeck coefficient ballpark for a metal directly from the Seebeck effect, equate the increase in electrostatic potential energy of an electron migrating from hot to cold to the decrease in average electron kinetic energy. Using the average kinetic energy of derivation {D.62}:

$$-e\,{\rm d}\varphi = -\,{\rm d}\frac{\pi^2}{4}\frac{(k_{\rm B}T)^2}{E^{\rm p}_{\rm F}}$$

Divide by $e\,{\rm d}T$ to get the Seebeck coefficient.

A.11.2 Figure of merit

To compare thermoelectric materials, an important quantity is the figure of merit of the material. The figure of merit is by convention written as M^2 where

$$M = \mathcal{P}\sqrt{\frac{\sigma}{\kappa T}}$$

The temperature T of the material should typically be taken as the average temperature in the device being examined. The reason that M is important has to do with units. Number M is "nondimensional," it has no units. In SI units, the Peltier coefficient $\mathcal{P}$ is in volts, the electrical conductivity σ in ampere/volt-meter, the temperature in Kelvin, and the thermal conductivity κ in watt/Kelvin-meter with watt equal to volt ampere. That makes the combination above nondimensional.

To see why that is relevant, suppose you have a material with a low Peltier coefficient. You might consider compensating for that by, say, scaling up the size of the material or the current through it. And maybe that does give you a better device than you would get with a material with a higher Peltier coefficient. Maybe not. How do you know?

dimensional analysis can help answer that question. It says that nondimensional quantities depend only on nondimensional quantities. For example, for a Peltier cooler you might define an efficiency as the heat removed from your ice cubes per unit electrical energy used. That is a nondimensional number. It will not depend on, say, the actual size of the semiconductor blocks, but it will depend on such nondimensional parameters as their shape, and their size relative to the overall device. Those are within your complete control during the design of the cooler. But the efficiency will also depend on the nondimensional figure of merit M above, and there you are limited to the available materials. Having a material with a higher figure of merit would give you a higher thermoelectric effect for the same losses due to electrical resistance and heat leaks.

To be sure, it is somewhat more complicated than that because two different materials are involved. That makes the efficiency depend on at least two nondimensional figures of merit, one for each material. And it might also depend on other nondimensional numbers that can be formed from the properties of the materials. For example, the efficiency of a simple thermoelectric generator turns out to depend on a net figure of merit given by, [9],

$$M_{\text{net}} = M_{\text{A}} \frac{\sqrt{\kappa_{\text{A}}/\sigma_{\text{A}}}}{\sqrt{\kappa_{\text{A}}/\sigma_{\text{A}}} + \sqrt{\kappa_{\text{B}}/\sigma_{\text{B}}}} - M_{\text{B}} \frac{\sqrt{\kappa_{\text{B}}/\sigma_{\text{B}}}}{\sqrt{\kappa_{\text{A}}/\sigma_{\text{A}}} + \sqrt{\kappa_{\text{B}}/\sigma_{\text{B}}}}$$

It shows that the figures of merit M_{A} and M_{B} of the two materials get multiplied by nondimensional fractions. These fractions are in the range from 0 to 1, and they sum to one. To get the best efficiency, you would like M_{A} to be as large positive as possible, and M_{B} as large negative as possible. That is as noted in the text. But all else being the same, the efficiency also depends to some extent on the nondimensional fractions multiplying M_{A} and M_{B}. It helps if the material with the larger figure of merit $|M|$ also has the larger ratio of κ/σ. If say M_{A} exceeds $-M_{\text{B}}$ for the best materials A and B, then you could potentially replace B by a cheaper material with a much lower figure of merit, as long as that replacement material has a very low value of κ/σ relative to A. In general,

the more nondimensional numbers there are that are important, the harder it is to analyze the efficiency theoretically.

A.11.3 Physical Seebeck mechanism

The given qualitative description of the Seebeck mechanism is very crude. For example, for semiconductors it ignores variations in the number of charge carriers. Even for a free-electron gas model for metals, there may be variations in charge carrier density that offset velocity effects. Worse, for metals it ignores the exclusion principle that restricts the motion of the electrons. And it ignores that the hotter side does not just have electrons with higher energy relative to the Fermi level than the colder side, it also has electrons with lower energy that can be excited to move. If the lower energy electrons have a larger mean free path, they can come from larger distances than the higher energy ones. And for metal electrons in a lattice, the velocity might easily go down with energy instead of up. That is readily appreciated from the spectra in chapter 6.22.2.

For a much more detailed description, see "Thermoelectric Effects in Metals: Thermocouples" by S. O. Kasap, 2001. This paper is available on the web for personal study. It includes actual data for metals compared to the simple theory.

A.11.4 Full thermoelectric equations

To understand the Peltier, Seebeck, and Thomson effects more precisely, the full equations of heat and charge flow are needed. That is classical thermodynamics, not quantum mechanics. However, standard undergraduate thermodynamics classes do not cover it, and even the thick standard undergraduate text books do not provide much more than a superficial mention that thermoelectric effects exist. Therefore this subsection will describe the equations of thermoelectrics in a nutshell.

The discussion will be one-dimensional. Think of a bar of material aligned in the x-direction. If the full three-dimensional equations of charge and heat flow are needed, for isotropic materials you can simply replace the x derivatives by gradients.

Heat flow is primarily driven by variations in temperature, and electric current by variations in the chemical potential of the electrons. The question is first of all what is the precise relation between those variations and the heat flow and current that they cause.

Now the microscopic scales that govern the motion of atoms and electrons are normally extremely small. Therefore an atom or electron "sees" only a very small portion of the macroscopic temperature and chemical potential distributions. The atoms and electrons do notice that the distributions are not constant, otherwise they would not conduct heat or current at all. But they see

so little of the distributions that to them they appear to vary linearly with position. As a result it is simple gradients, i.e. first derivatives, of the temperature and potential distributions that drive heat flow and current in common solids. Symbolically:

$$q = f_1\left(\frac{\mathrm{d}T}{\mathrm{d}x}, \frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x}\right) \qquad j = f_2\left(\frac{\mathrm{d}T}{\mathrm{d}x}, \frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x}\right)$$

Here q is the "heat flux density;" "flux" is a fancy word for "flow" and the qualifier "density" indicates that it is per unit cross-sectional area of the bar. Similarly j is the current density, the current per unit cross-sectional area. If you want, it is the charge flux density. Further T is the temperature, and φ_μ is the chemical potential μ per unit electron charge $-e$. That includes the electrostatic potential (simply put, the voltage) as well as an intrinsic chemical potential of the electrons. The unknown functions f_1 and f_2 will be different for different materials and different conditions.

The above equations are not valid if the temperature and potential distributions change nontrivially on microscopic scales. For example, shock waves in supersonic flows of gases are extremely thin; therefore you cannot use equations of the type above for them. Another example is highly rarefied flows, in which the molecules move long distances without collisions. Such extreme cases can really only be analyzed numerically and they will be ignored here. It is also assumed that the materials maintain their internal integrity under the conduction processes.

Under normal conditions, a further approximation can be made. The functions f_1 and f_2 in the expressions for the heat flux and current densities would surely depend nonlinearly on their two arguments if these would appear finite *on a microscopic scale.* But on a microscopic scale, temperature and potential hardly change. (Supersonic shock waves and similar are again excluded.) Therefore, the gradients appear small in microscopic terms. And if that is true, functions f_1 and f_2 can be linearized using Taylor series expansion. That gives:

$$q = A_{11}\frac{\mathrm{d}T}{\mathrm{d}x} + A_{12}\frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x} \qquad j = A_{21}\frac{\mathrm{d}T}{\mathrm{d}x} + A_{22}\frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x}$$

The four coefficients $A_{..}$ will normally need to be determined experimentally for a given material at a given temperature. The properties of solids vary normally little with pressure.

By convention, the four coefficients are rewritten in terms of four other, more intuitive, ones:

$$\boxed{q = -(\kappa + \mathcal{P}\mathcal{S}\sigma)\frac{\mathrm{d}T}{\mathrm{d}x} - \mathcal{P}\sigma\frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x} \qquad j = -\mathcal{S}\sigma\frac{\mathrm{d}T}{\mathrm{d}x} - \sigma\frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x}} \tag{A.31}$$

This *defines* the heat conductivity κ, the electrical conductivity σ, the Seebeck coefficient $\mathcal{S}$ and the Peltier coefficient $\mathcal{P}$ of the material. (The signs of the Peltier and Seebeck coefficients vary considerably between references.)

If conditions are isothermal, the second equation is Ohm's law for a unit cube of material, with σ the usual conductivity, the inverse of the resistance of the unit cube. The Seebeck effect corresponds to the case that there is no current. In that case, the second equation produces

$$\boxed{\frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x} = -\mathcal{S}\frac{\mathrm{d}T}{\mathrm{d}x}} \tag{A.32}$$

To see what this means, integrate this along a closed circuit all the way from lead 1 of a voltmeter through a sample to the other lead 2. That gives

$$\varphi_{\mu,2} - \varphi_{\mu,1} = -\int_1^2 \mathcal{S}\mathrm{d}T \tag{A.33}$$

Assuming that the two leads of the voltmeter are at the same temperature, their intrinsic chemical potentials are the same. In that case, the difference in potentials is equal to the difference in electrostatic potentials. In other words, the integral gives the difference between the voltages inside the two leads. And that is the voltage that will be displayed by the voltmeter.

It is often convenient to express the heat flux density q in terms of the current density instead of the gradient of the potential φ_μ. Eliminating this gradient from the equations (A.31) produces

$$\boxed{q = -\kappa\frac{\mathrm{d}T}{\mathrm{d}x} + \mathcal{P}j} \tag{A.34}$$

In case there is no current, this is the well-known Fourier's law of heat conduction, with κ the usual thermal conductivity. Note that the heat flux density is often simply called the heat flux, even though it is per unit area. In the presence of current, the heat flux density is augmented by the Peltier effect, the second term.

The total energy flowing through the bar is the sum of the thermal heat flux and the energy carried along by the electrons:

$$j_E = q + j\varphi_\mu$$

If the energy flow is constant, the same energy flows out of a piece $\mathrm{d}x$ of the bar as flows into it. Otherwise the negative x-derivative of the energy flux density gives the net energy accumulation $\dot{e}$ per unit volume:

$$\dot{e} = -\frac{\mathrm{d}j_E}{\mathrm{d}x} = -\frac{\mathrm{d}q}{\mathrm{d}x} - j\frac{\mathrm{d}\varphi_\mu}{\mathrm{d}x}$$

where it was assumed that the electric current is constant as it must be for a steady state. Of course, in a steady state any nonzero $\dot{e}$ must be removed through heat conduction through the sides of the bar of material being tested,

or through some alternative means. Substituting in from (A.34) for q and from the second of (A.31) for the gradient of the potential gives:

$$\dot{e} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\kappa \frac{\mathrm{d}T}{\mathrm{d}x}\right) + \frac{j^2}{\sigma} - \mathcal{K}\frac{\mathrm{d}T}{\mathrm{d}x} j \qquad \mathcal{K} \equiv \frac{\mathrm{d}\mathcal{P}}{\mathrm{d}T} - \mathcal{S}$$

The final term in the energy accumulation is the Thomson effect or Kelvin heat. The Kelvin (Thomson) coefficient $\mathcal{K}$ can be cleaned up using the second Kelvin relationship given in a later subsection.

The equations (A.31) are often said to be representative of nonequilibrium thermodynamics. However, they correspond to a vanishingly small perturbation from thermodynamical equilibrium. The equations would more correctly be called quasi-equilibrium thermodynamics. Nonequilibrium thermodynamics is what you have inside a shock wave.

A.11.5 Charge locations in thermoelectrics

The statement that the charge density is neutral inside the material comes from [[8]].

A simplified macroscopic derivation can be given based on the thermoelectric equations (A.31). The derivation assumes that the temperature and chemical potential are almost constant. That means that derivatives of thermodynamic quantities and electric potential are small. That makes the heat flux and current also small.

Next, in three dimensions replace the x derivatives in the thermoelectric equations (A.31) by the gradient operator ∇. Now under steady-state conditions, the divergence of the current density must be zero, or there would be an unsteady local accumulation or depletion of net charge, chapter 13.2. Similarly, the divergence of the heat flux density must be zero, or there would be an accumulation or depletion of thermal energy. (This ignores local heat generation as an effect that is quadratically small for small currents and heat fluxes.)

Therefore, taking the divergence of the equations (A.31) and ignoring the variations of the coefficients, which give again quadratically small contributions, it follows that the Laplacians of both the temperature and the chemical potential are zero.

Now the chemical potential includes both the intrinsic chemical potential and the additional electrostatic potential. The intrinsic chemical potential depends on temperature. Using again the assumption that quadratically small terms can be ignored, the Laplacian of the intrinsic potential is proportional to the Laplacian of the temperature and therefore zero.

Then the Laplacian of the electrostatic potential must be zero too, to make the Laplacian of the total potential zero. And that then implies the absence of net charge inside the material according to Maxwell's first equation, chapter 13.2. Any net charge must accumulate at the surfaces.

A.11.6 Kelvin relationships

This subsection gives an explanation of the definition of the thermal heat flux in thermoelectrics. It also explains that the Kelvin (or Thomson) relationships are a special case of the more general "Onsager reciprocal relations." If you do not know what thermodynamical entropy is, you should not be reading this subsection. Not before reading chapter 11, at least.

For simplicity, the discussion will again assume one-dimensional conduction of heat and current. The physical picture is therefore conduction along a bar aligned in the x-direction. It will be assumed that the bar is in a steady state, in other words, that the temperature and chemical potential distributions, heat flux and current through the bar all do not change with time.

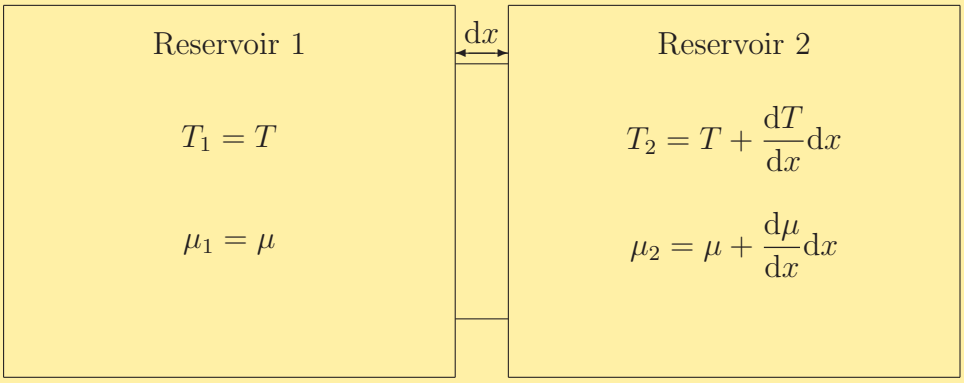


Figure A.1: Analysis of conduction.

The primary question is what is going on in a single short segment $\mathrm{d}x$ of such a bar. Here $\mathrm{d}x$ is assumed to be small on a macroscopic scale, but large on a microscopic scale. To analyze the segment, imagine it taken out of the bar and sandwiched between two big idealized "reservoirs" 1 and 2 of the same material, as shown in figure A.1. The idealized reservoirs are assumed to remain at uniform, thermodynamically reversible, conditions. Reservoir 1 is at the considered time at the same temperature and chemical potential as the start of the segment, and reservoir 2 at the same temperature and chemical potential as the end of the segment. The reservoirs are assumed to be big enough that their properties change slowly in time. Therefore it is assumed that their time variations do not have an effect on what happens inside the bar segment at the considered time. For simplicity, it will also be assumed that the material consists of a single particle type. Some of these particles are allowed to move through the bar segment from reservoir 1 to reservoir 2.

In other words, there is a flow, or flux, of particles through the bar segment. The corresponding particle flux density j_I is the particle flow per unit area. For simplicity, it will be assumed that the bar has unit area. Then there is no difference between the particle flow and the particle flux density. Note that the

same flow of particles j_I must enter the bar segment from reservoir 1 as must exit from the segment into reservoir 2. If that was not the case, there would be a net accumulation or depletion of particles inside the bar segment. That is not possible, because the bar segment is assumed to be in a steady state. Therefore the flow of particles through the bar segment decreases the number of particles I_1 in reservoir 1, but increases the number I_2 in reservoir 2 correspondingly:

$$j_I = -\frac{\mathrm{d}I_1}{\mathrm{d}t} = \frac{\mathrm{d}I_2}{\mathrm{d}t}$$

Further, due to the energy carried along by the moving particles, as well as due to thermal heat flow, there will be a net energy flow j_E through the bar segment. Like the particle flow, the energy flow comes out of reservoir 1 and goes into reservoir 2:

$$j_E = -\frac{\mathrm{d}E_1}{\mathrm{d}t} = \frac{\mathrm{d}E_2}{\mathrm{d}t}$$

Here E_1 is the total energy inside reservoir 1, and E_2 that inside reservoir 2. It is assumed that the reservoirs are kept at constant volume and are thermally insulated except at the junction with the bar, so that no energy is added due to pressure work or heat conduction elsewhere. Similarly, the sides of the bar segment are assumed thermally insulated.

One question is how to define the heat flux through the bar segment. In the absence of particle motion, the second law of thermodynamics allows an unambiguous answer. The heat flux q through the bar enters reservoir 2, and the second law of thermodynamics then says:

$$q_2 = T_2 \frac{\mathrm{d}S_2}{\mathrm{d}t}$$

Here S_2 is the entropy of the reservoir 2. In the presence of particles moving through the bar, the definition of thermal energy, and so the corresponding heat flux, becomes more ambiguous. The particles also carry along nonthermal energy. The question then becomes what should be counted as thermal energy, and what as nonthermal. To resolve that, the heat flux into reservoir 2 will be *defined* by the expression above. Note that the heat flux out of reservoir 1 might be slightly different because of variations in energy carried by the particles. It is the total energy flow j_E, not the heat flow q, that must be exactly constant.

To understand the relationship between heat flux and energy flux more clearly, some basic thermodynamics can be used. See chapter 11.12 for more details, including generalization to more than one particle type. A combination of the first and second laws of thermodynamics produces

$$T\,\mathrm{d}\bar{s} = \mathrm{d}\bar{e} + P\,\mathrm{d}\bar{v} \qquad S = \bar{s}I \quad E = \bar{e}I \quad V = \bar{v}I$$

in which $\bar{s}$, $\bar{e}$, and $\bar{v}$ are the entropy, internal energy, and volume per particle, and P is the pressure. That can be used to rewrite the derivative of entropy in

the definition of the heat flux above:

$$T\,\mathrm{d}S = T\mathrm{d}(\bar{s}I) = T(\mathrm{d}\bar{s})I + T\bar{s}(\mathrm{d}I) = (\mathrm{d}\bar{e} + P\,\mathrm{d}\bar{v})I + T\bar{s}(\mathrm{d}I)$$

That can be rewritten as

$$T\,\mathrm{d}S = \mathrm{d}E + P\mathrm{d}V - (\bar{e} + P\bar{v} - T\bar{s})\mathrm{d}I$$

as can be verified by writing E and V as $\bar{e}I$ and $\bar{v}I$ and differentiating out. The parenthetical expression in the above equation is in thermodynamics known as the Gibbs free energy. Chapter 11.13 explains that it is the same as the chemical potential μ in the distribution laws. Therefore:

$$T\,\mathrm{d}S = \mathrm{d}E + P\mathrm{d}V - \mu\mathrm{d}I$$

(Chapter 11.13 does not include an additional electrostatic energy due to an ambient electric field. But an intrinsic chemical potential can be defined by subtracting the electrostatic potential energy. The corresponding intrinsic energy also excludes the electrostatic potential energy. That makes the expression for the chemical potential the same in terms of intrinsic quantities as in terms of nonintrinsic ones. See also the discussion in chapter 6.14.)

Using the above expression for the change in entropy in the definition of the heat flux gives, noting that the volume is constant,

$$q_2 = \frac{\mathrm{d}E_2}{\mathrm{d}t} - \mu_2\frac{\mathrm{d}I_2}{\mathrm{d}t} = j_E - \mu_2 j_I$$

It can be concluded from this that the nonthermal energy carried along per particle is μ. The rest of the net energy flow is thermal energy.

The Kelvin relationships are related to the net entropy generated by the segment of the bar. The second law implies that irreversible processes always increase the net entropy in the universe. And by definition, the complete system figure A.1 examined here is isolated. It does not exchange work nor heat with its surroundings. Therefore, the entropy of this system must increase in time due to irreversible processes. More specifically, the net system entropy must go up due to the irreversible heat conduction and particle transport in the segment of the bar. The reservoirs are taken to be thermodynamically reversible; they do not create entropy out of nothing. But the heat conduction in the bar is irreversible; it goes from hot to cold, not the other way around, in the absence of other effects. Similarly, the particle transport goes from higher chemical potential to lower.

While the conduction processes in the bar create net entropy, the entropy of the bar still does not change. The bar is assumed to be in a steady state. Instead the entropy created in the bar causes a net increase in the combined entropy of the reservoirs. Specifically,

$$\frac{\mathrm{d}S_{\mathrm{net}}}{\mathrm{d}t} = \frac{\mathrm{d}S_2}{\mathrm{d}t} + \frac{\mathrm{d}S_1}{\mathrm{d}t}$$

By definition of the heat flux,

$$\frac{\mathrm{d}S_{\text{net}}}{\mathrm{d}t} = \frac{q_2}{T_2} - \frac{q_1}{T_1}$$

Substituting in the expression for the heat flux in terms of the energy and particle fluxes gives

$$\frac{\mathrm{d}S_{\text{net}}}{\mathrm{d}t} = \left(\frac{1}{T_2} j_E - \frac{\mu_2}{T_2} j_I\right) - \left(\frac{1}{T_1} j_E - \frac{\mu_1}{T_1} j_I\right)$$

Since the area of the bar is one, its volume is $\mathrm{d}x$. Therefore, the entropy generation per unit volume is:

$$\frac{1}{\mathrm{d}x}\frac{\mathrm{d}S_{\text{net}}}{\mathrm{d}t} = \frac{\mathrm{d}1/T}{\mathrm{d}x} j_E + \frac{\mathrm{d}{-}\mu/T}{\mathrm{d}x} j_I \tag{A.35}$$

That used that any expression of the form $(f_2 - f_1)/\mathrm{d}x$ is by definition the derivative of f.

The above expression for the entropy generation implies that a nonzero derivative of $1/T$ must cause an energy flow of the same sign. Otherwise the entropy of the system would decrease if the derivative in the second term is zero. Similarly, a nonzero derivative of $-\mu/T$ must cause a particle flow of the same sign. Of course, that does not exclude that the derivative of $1/T$ may also cause a particle flow as a secondary effect, or a derivative of $-\mu/T$ an energy flow. Using the same reasoning as in an earlier subsection gives:

$$j_E = L_{11}\frac{\mathrm{d}1/T}{\mathrm{d}x} + L_{12}\frac{\mathrm{d}{-}\mu/T}{\mathrm{d}x} \qquad j_I = L_{21}\frac{\mathrm{d}1/T}{\mathrm{d}x} + L_{22}\frac{\mathrm{d}{-}\mu/T}{\mathrm{d}x} \tag{A.36}$$

where the $L_{..}$ are again coefficients to be determined experimentally. But in this case, the coefficients L_{11} and L_{22} must necessarily be positive. That can provide a sanity check on the experimental results. It is an advantage gained from taking the flows and derivatives directly from the equation of entropy generation. In fact, somewhat stronger constraints apply. If the expressions for j_E and j_I are plugged into the expression for the entropy generation, the result must be positive regardless of what the values of the derivatives are. That requires not just that L_{11} and L_{22} are positive, but also that the average of L_{12} and L_{21} is smaller in magnitude than the geometric average of L_{11} and L_{22}.

The so-called Onsager reciprocal relations provide a further, and much more specific constraint. They say that the coefficients of the secondary effects, L_{12} and L_{21}, must be equal. In the terms of linear algebra, matrix $L_{..}$ must be symmetric and positive definite. In real life, it means that only three, not four coefficients have to be determined experimentally. That is very useful because the experimental determination of secondary effects is often difficult.

The Onsager relations remain valid for much more general systems, involving flows of other quantities. Their validity can be argued based on experimental evidence, or also theoretically based on the symmetry of the microscopic dynamics with respect to time reversal. If there is a magnetic field involved, a coefficient L_{ij} will only equal L_{ji} after the magnetic field has been reversed: time reversal causes the electrons in your electromagnet to go around the opposite way. A similar observation holds if Coriolis forces are a factor in a rotating system.

The equations (A.36) for j_E and j_I above can readily be converted into expressions for the heat flux density $q = j_E - \mu j_I$ and the current density $j = -ej_I$. If you do so, then differentiate out the derivatives, and compare with the thermoelectric equations (A.31) given earlier, you find that the Onsager relation $L_{12} = L_{21}$ translates into the second Kelvin relation

$$\mathcal{P} = \mathcal{S}T$$

That allows you to clean up the Kelvin coefficient to the first Kelvin relationship:

$$\mathcal{K} \equiv \frac{\mathrm{d}\mathcal{P}}{\mathrm{d}T} - \mathcal{S} = T\frac{\mathrm{d}\mathcal{S}}{\mathrm{d}T} = \frac{\mathrm{d}\mathcal{S}}{\mathrm{d}\ln T}$$

It should be noted that while the second Kelvin relationship is named after Kelvin, he never gave a valid proof of the relationship. Neither did many other authors that tried. It was Onsager who first succeeded in giving a more or less convincing theoretical justification. Still, the most convincing support for the reciprocal relations remains the overwhelming experimental data. See Miller (Chem. Rev. 60, 15, 1960) for examples. Therefore, the reciprocal relationships are commonly seen as an additional axiom to be added to thermodynamics to allow quasi-equilibrium systems to be treated.

A.12 Heisenberg picture

This book follows the formulation of quantum mechanics as developed by Schrödinger. However, there is another, earlier, formulation due to Heisenberg. This subsection gives a brief description so that you are aware of it when you run into it in literature.

In the Schrödinger picture, physical observables like position and momentum are represented by time-independent operators. The time dependence is in the wave function. This is somewhat counterintuitive because classically position and momentum are time dependent quantities. The Heisenberg picture removes the time dependence from the wave function and absorbs it into the operator.

To see how that works out, consider first the general form of the wave function. It can be written as

$$\Psi(\ldots;t) = e^{-\mathrm{i}Ht/\hbar}\Psi(\ldots;0) \tag{A.37}$$

where the exponential of an operator is defined through its Taylor series:

$$e^{-\mathrm{i}Ht/\hbar} = 1 - \mathrm{i}\frac{t}{\hbar}H - \frac{t^2}{2!\hbar^2}H^2 + \ldots \tag{A.38}$$

(To check the above expression for the wave function, take the initial wave function to be any energy eigenfunction of energy E. You get the correct $e^{-\mathrm{i}Et/\hbar}$ time dependence, chapter 7.1.2. Every H becomes an E. And if the expression works for any eigenfunction, it works for all their combinations too. That means that it works for any wave function, because the eigenfunctions are complete. To be sure, the above form of the wave function applies only if the Hamiltonian is independent of time. Even if it is not, the transformation from the initial wave function $\Psi(\ldots;0)$ to a later one $\Psi(\ldots;t)$ still remains a "unitary" one; one that keeps the wave function normalized. But then you will need to use the Schrödinger equation directly to figure out the time dependence.)

Now consider an arbitrary Schrödinger operator $\widehat{A}$. The physical effects of the operator can be characterized by inner products, as in

$$\langle \Psi_1(\ldots;t) | \widehat{A} \Psi_2(\ldots;t) \rangle \tag{A.39}$$

Such a dot product tells you what amount of a wave function Ψ_1 is produced by applying the operator on a wave function Ψ_2. Knowing these inner products for all wave functions is equivalent to knowing the operator.

If the time-dependent exponentials are now peeled off Ψ_1 and Ψ_2 and absorbed into the operator, you get the time-dependent Heisenberg operator

$$\widetilde{A} \equiv e^{\mathrm{i}Ht/\hbar} \widehat{A} e^{-\mathrm{i}Ht/\hbar} \tag{A.40}$$

Heisenberg operators will be indicated with a tilde instead of a hat. Note that the argument of the first exponential changed sign because it was taken to the other side of the inner product.

The operator $\widetilde{A}$ depends on time. To see how it evolves, differentiate the product with respect to time:

$$\frac{\mathrm{d}\widetilde{A}}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar} H e^{\mathrm{i}Ht/\hbar} \widehat{A} e^{-\mathrm{i}Ht/\hbar} + e^{\mathrm{i}Ht/\hbar} \frac{\partial \widehat{A}}{\partial t} e^{-\mathrm{i}Ht/\hbar} - e^{\mathrm{i}Ht/\hbar} \widehat{A} e^{-\mathrm{i}Ht/\hbar} \frac{\mathrm{i}}{\hbar} H$$

The first and third terms can be recognized as a multiple of the commutator of H and $\widetilde{A}$, while the middle term is the Heisenberg version of the time derivative of $\widehat{A}$, in case $\widehat{A}$ does happen to depend on time. So the evolution equation for the Heisenberg operator becomes

$$\frac{\mathrm{d}\widetilde{A}}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar} \left[H, \widetilde{A} \right] + \widetilde{\frac{\partial \widehat{A}}{\partial t}} \qquad \left[H, \widetilde{A} \right] = e^{\mathrm{i}Ht/\hbar} \left[H, \widehat{A} \right] e^{-\mathrm{i}Ht/\hbar} \tag{A.41}$$

(Note that there is no difference between the Hamiltonians $\widehat{H}$ and $\widetilde{H}$ because H commutes with itself, hence with its exponentials.)

For example, consider the Schrödinger $\widehat{x}$ position and $\widehat{p}_x$ linear momentum operators of a particle. These do not depend on time. Using the commutators as figured out in chapter 7.2.1, the corresponding Heisenberg operators evolve as:

$$\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}t} = \frac{1}{m}\widetilde{p}_x \qquad \frac{\mathrm{d}\widetilde{p}_x}{\mathrm{d}t} = -\widetilde{\frac{\partial V}{\partial x}}$$

Those have the exact same form as the equations for the classical position and momentum of the particle.

In fact, the equivalent of the general equation (A.41) is also found in classical physics: it is derived in advanced mechanics, with the so-called "Poisson bracket" taking the place of the commutator. As a simple example, consider one-dimensional motion of a particle. Any variable a that depends on the position and linear momentum of the particle, and maybe also explicitly on time, has a time derivative given by

$$\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{\partial a}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial a}{\partial p_x}\frac{\mathrm{d}p_x}{\mathrm{d}t} + \frac{\partial a}{\partial t}$$

according to the total differential of calculus. And from the classical Hamiltonian

$$H = \frac{p_x^2}{2m} + V$$

it is seen that the time derivatives of position and momentum obey the classical "Hamiltonian dynamics"

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial p_x} \qquad \frac{\mathrm{d}p_x}{\mathrm{d}t} = -\frac{\partial H}{\partial x}$$

Substituting this into the time derivative of a gives

$$\frac{\mathrm{d}a}{\mathrm{d}t} = \frac{\partial a}{\partial x}\frac{\partial H}{\partial p_x} - \frac{\partial a}{\partial p_x}\frac{\partial H}{\partial x} + \frac{\partial a}{\partial t}$$

The first two terms in the right hand side are by definition minus the Poisson bracket $\{H, a\}_{\mathrm{P}}$, so

$$\frac{\mathrm{d}a}{\mathrm{d}t} = -\{H, a\}_{\mathrm{P}} + \frac{\partial a}{\partial t} \qquad \{H, a\}_{\mathrm{P}} \equiv \frac{\partial H}{\partial x}\frac{\partial a}{\partial p_x} - \frac{\partial a}{\partial x}\frac{\partial H}{\partial p_x}$$

Note that the Poisson bracket, like the commutator, is antisymmetric under exchange of H and a. Apparently, formally identifying the Poisson bracket with the commutator divided by $\mathrm{i}\hbar$ brings you from classical mechanics to Heisenberg's quantum mechanics.

More generally, the classical Hamiltonian can depend on multiple and non-Cartesian coordinates, generically called "generalized coordinates." In that case, in the Poisson bracket you must sum over all generalized coordinates and their

associated so-called "canonical" momenta. For a Cartesian position coordinate, the canonical momentum is the corresponding linear momentum. For an angular coordinate, it is the corresponding angular momentum. In general, using the so-called Lagrangian formulation usually covered in an engineering education, and otherwise found in addendum {A.1}, the canonical momentum is the derivative of the Lagrangian with respect to the time derivative of the coordinate.

The bottom line is that the Heisenberg equations are usually not easy to solve unless you return to the Schrödinger picture by peeling off the time dependence. In relativistic applications however, time joins space as an additional coordinate, and the Heisenberg picture becomes more helpful. It can also make it easier to identify the correspondence between classical equations and the corresponding quantum operators.

Key Points

- In the Heisenberg picture, operators evolve in time just like their physical variables do in classical physics.

A.13 Integral Schrödinger equation

The Hamiltonian eigenvalue problem, or time-independent Schrödinger equation, is the central equation of quantum mechanics. It reads

$$\frac{\hbar^2}{2m}\nabla^2\psi(\vec{r}) + V(\vec{r})\psi(\vec{r}) = E\psi(\vec{r})$$

Here ψ is the wave function, E is the energy of the state described by the wave function, V is the potential energy, m is the mass of the particle, and $\hbar$ is the scaled Planck constant.

The equation also involves the Laplacian operator, defined as

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

Therefore the Hamiltonian eigenvalue problem involves partial derivatives, and it is called a partial differential equation.

However, it is possible to manipulate the equation so that the wave function ψ appears inside an integral rather than inside partial derivatives. The equation that you get this way is called the "integral Schrödinger equation." It takes the form, {D.31}:

$$\boxed{\psi(\vec{r}) = \psi_0(\vec{r}) - \frac{m}{2\pi\hbar^2}\int_{\text{all } \vec{r}\,'} \frac{e^{ik|\vec{r}-\vec{r}\,'|}}{|\vec{r}-\vec{r}\,'|} V(\vec{r}\,')\psi(\vec{r}\,')\,\mathrm{d}^3\vec{r}\,' \qquad k = \frac{\sqrt{2mE}}{\hbar}} \tag{A.42}$$

Here ψ_0 is any wave function of energy E in free space. In other words ψ_0 is any wave function for the particle in the absence of the potential V. The constant k is a measure of the energy of the particle. It also corresponds to a wave number far from the potential. While not strictly required, the integral Schrödinger equation above tends to be most suited for particles in infinite space.

A.14 The Klein-Gordon equation

The Schrödinger equation for the quantum wave function is based on the non-relativistic expression for the energy of a particle. This addendum looks at the simplest relativistic equation for wave functions, called the Klein-Gordon equation. The discussion will largely be restricted to a spinless particle in empty space, where there is no potential energy. However, the Klein-Gordon equation is the first step to more complex relativistic equations.

Recall first how the Schrödinger equation arose. If there is no potential energy, classical physics says that the energy E is just the kinetic energy $p^2/2m$ of the particle. Here p is the linear momentum of the particle and m its mass. Quantum mechanics replaces the energy E by the operator $\mathrm{i}\hbar\partial/\partial t$ and the momentum $\vec{p}$ by $\hbar\nabla/\mathrm{i}$, where

$$\nabla = \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

Then it applies the resulting operators on the wave function Ψ. That then produces the Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi$$

Solutions with a definite value E for the energy take the form $\Psi = ce^{-\mathrm{i}Et/\hbar}\psi$. Substituting that into the Schrödinger equation and rearranging produces the so-called Hamiltonian eigenvalue problem

$$-\frac{\hbar^2}{2m}\nabla^2\psi = E\psi$$

Here ψ is called the energy eigenfunction.

According to classical relativity however, the energy E of a particle in empty space is not just kinetic energy, but also rest mass energy mc^2, where c is the speed of light. In particular, chapter 1.1.2 (1.2),

$$E = \sqrt{(mc^2)^2 + p^2c^2}$$

The momentum can be identified with the same operator as before. But square roots of operators are very ugly. So the smart thing to do is to square both

sides above. Making the same substitutions as for the Schrödinger equation and cleaning up then gives the "Klein-Gordon equation"

$$\boxed{-\frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2}+\nabla^2\Psi=\left(\frac{mc^2}{\hbar c}\right)^2\Psi} \tag{A.43}$$

Solutions $ce^{-\mathrm{i}Et/\hbar}\psi$ with definite energy E satisfy the "time-independent Klein-Gordon equation" or square Hamiltonian eigenvalue problem

$$-\hbar^2c^2\nabla^2\psi+(mc^2)^2\psi=E^2\psi$$

This may be rewritten in a form so that both the Schrödinger equation and the Klein-Gordon equation are covered:

$$\boxed{\text{empty space:}\quad -\nabla^2\psi=k^2\psi\qquad\begin{cases}\text{Schrödinger:}\quad k=\dfrac{\sqrt{2mE}}{\hbar}\\[2ex]\text{Klein-Gordon:}\quad k=\dfrac{\sqrt{E^2-(mc^2)^2}}{\hbar c}\end{cases}} \tag{A.44}$$

Here the constant k is called the "wave number." Note that the nonrelativistic energy does not include the rest mass energy. When that is taken into account, the Schrödinger expression for k above is the nonrelativistic approximation for the Klein-Gordon k as it should be.

Further note that relativistic or not, the magnitude of linear momentum p is given by the "de Broglie relation" $p=\hbar k$. That is because relativistic or not the momentum operator is $\hbar\nabla/\mathrm{i}$, so $\widehat{p}^2=-\hbar^2\nabla^2$. Similarly, relativistic or not, the energy is associated the operator $\mathrm{i}\hbar\partial/\partial t$. That means that the time-dependent factor in states of definite energy is $e^{-\mathrm{i}Et/\hbar}$. That allows the energy to be associated with an "angular frequency" ω by writing the exponential as $e^{-\mathrm{i}\omega t}$. The relationship between energy and frequency is then $E=\hbar\omega$. That is known as the "Planck-Einstein relation" when applied to photons. In short, relativistic or not,

$$\boxed{p=\hbar k\qquad E=\hbar\omega} \tag{A.45}$$

The wave number k is the quantum number of linear momentum, and the angular frequency ω is the one of energy. See addendum {A.19} for more on how these numbers arise physically from symmetries of nature.

It may be noted that Schrödinger wrote down the Klein-Gordon equation first. But when he added the Coulomb potential, he was not able to get the energy levels of the hydrogen atom. To fix that problem, he wrote down the simpler nonrelativistic equation that bears his name. The problem in the relativistic case is that after you add the Coulomb potential to the energy, you can no longer square away the square root. Eventually, Dirac figured out how to get around that problem, chapter 12.12 and {D.81}. In brief, he assumed

that the wave function for the electron is not a scalar, but a four-dimensional vector, (two spin states for the electron, plus two spin states for the associated antielectron, or positron.) Then he assumed that the square root takes a simple form for that vector.

Since this addendum assumes a particle in empty space, the problem with the Coulomb potential does not arise. But there are other issues. The good news is that according to the Klein-Gordon equation, effects do not propagate at speeds faster than the speed of light. That is known from the theory of partial differential equations. In classical physics, effects cannot propagate faster than the speed of light, so it is somewhat reassuring that the Klein-Gordon equation respects that.

Also, all inertial observers agree about the Klein-Gordon equation, regardless of the motion of the observer. That is because all inertial observers agree about the rest mass m of a particle and the value of the speed of light c. So they all agree about the right hand side in the Klein-Gordon equation (A.43). And the left hand side in the Klein-Gordon equation is also the same for all inertial observers. You can crunch that out using the Lorentz transform as given in chapter 1.2.1 (1.6). (Those familiar with index notation as briefly described in chapter 1.2.5 recognize the entire left hand side as being simply $\partial^\mu \partial_\mu \Psi$. That is unchanged going from one observer to the next, because the upper index transforms under the Lorentz transform and the lower index under the inverse Lorentz transform. The operator $\partial^\mu \partial_\mu$ is called the "D'Alembertian," much like ∇^2 is called the Laplacian.)

But the bad news is that the Klein-Gordon equation does not necessarily preserve the integral of the square magnitude of the wave function. The Schrödinger equation implies that, {D.32},

$$\int_{\text{all}} |\Psi|^2 \, \mathrm{d}^3\vec{r} = \text{constant, the same for all time}$$

The wave function is then normalized so that the constant is 1. According to the Born statistical interpretation, chapter 3.1, the integral above represents the probability of finding the particle if you look at all possible positions. That must be 1 at whatever time you look; the particle must be somewhere. Because the Schrödinger equation ensures that the integral above stays 1, it ensures that the particle cannot just disappear, and that no second particle can show up out of nowhere.

But the Klein-Gordon equation does not preserve the integral above. Therefore the number of particles is not necessarily preserved. That is not as bad as it looks, anyway, because in relativity the mass-energy equivalence allows particles to be created or destroyed, chapter 1.1.2. But certainly, the interpretation of the wave function is not a priori obvious. The integral that the Klein-Gordon

equation does preserve is, {D.32},

$$\int_{\text{all}} \left| \frac{1}{c} \frac{\partial \Psi}{\partial t} \right|^2 + |\nabla \Psi|^2 + \left| \frac{mc^2}{\hbar c} \Psi \right|^2 \, \mathrm{d}^3\vec{r} = \text{constant}$$

It is maybe somewhat comforting that according to this expression, the integral of $|\Psi|^2$ must at least remain bounded. That does assume that the rest mass m of the particle is not zero. Photons need not apply.

Another problem arises because even though the square energy E^2 is normally positive, the energy E itself can still be both positive or negative. That is a problem, because then there is no lower limit to the energy, there is no ground state. The particle can then transition to states of lower and lower energy tending to minus infinity. That would release unbounded amounts of energy. (Since the kinetic energy can be arbitrarily large, the positive value of the energy can be arbitrarily large. That makes the negative value of the energy also arbitrarily large in magnitude.)

You might say, just ignore the negative energy possibility. But Dirac found that that does not work; you need both positive and negative energy states to explain such things as the hydrogen energy levels. The way Dirac solved the problem for electrons is to assume that all negative states are already filled with electrons. Unfortunately, that does not work for bosons, since any number of bosons can go into a state.

The modern view is to consider the negative energy solutions to represent antiparticles. In that view, antiparticles have positive energy, but move backwards in time. For example, Dirac's negative energy states are not electrons with negative energy, but positrons with positive energy. Positrons are then electrons that move backward in time. To illustrate the idea, consider two hypothetical wave functions of the form

$$e^{-\mathrm{i}Et/\hbar}\psi_1 \quad \text{and} \quad e^{\mathrm{i}Et/\hbar}\psi_2$$

where E is the positive root for the energy. The first wave function is no problem; it is of the form of a wave function that you would get for a nonrelativistic particle of energy E. The second wave function is the problem. It is not considered to be a particle of negative energy $-E$. Instead it is considered an antiparticle of positive energy E that moves backward in time. It is the reversal of the relevant direction of time that causes the sign change in the argument of the exponential.

You see why so much quantum physics is done using nonrelativistic equations.

A.15 Quantum Field Theory in a Nanoshell

The "classical" quantum theory discussed in this book runs into major difficulties with truly relativistic effects. In particular, relativity allows particles to

be created or destroyed. For example, a very energetic photon near a heavy nucleus might create an electron and a positron. Einstein's $E = mc^2$ implies that that is possible because mass is equivalent to energy. The photon energy is converted into the electron and positron masses. Similarly, an electron and positron can annihilate each other, releasing their energy as photons. The quantum formalism in this book cannot deal with particles that appear out of nothing or disappear. A modified formulation called "quantum field theory" is needed.

And quantum field theory is not just for esoteric conditions like electron-positron pair creation. The photons of light are routinely created and destroyed under normal conditions. Still more basic to an engineer, so are their equivalents in solids, the phonons of crystal vibrations. Then there is the band theory of semiconductors: electrons are "created" within the conduction band, if they pick up enough energy, or "annihilated" when they lose it. And the same happens for the real-life equivalent of positrons, holes in the valence band.

Such phenomena are routinely described within the framework of quantum field theory. Almost unavoidably you will run into it in literature, [18, 29]. Electron-phonon interactions are particularly important for engineering applications, leading to electrical resistance (along with crystal defects and impurities), and to the combination of electrons into Cooper pairs that act as bosons and so give rise to superconductivity.

This addendum explains some of the basic ideas of quantum field theory. It should allow you to recognize it when you see it. Addendum {A.23} uses the ideas to explain the quantization of the electromagnetic field. That then allows the quantum description of spontaneous emission of radiation by excited atoms or nuclei in {A.24}. Here a photon is created.

Unfortunately a full discussion of quantum field theory is far outside the scope of this book. Especially the fully relativistic theory is very involved. To explain quantum field theory in a nutshell takes Zee 500 pages, [53]. Tong ⟦17⟧ writes: "This is charming book, where emphasis is placed on physical understanding and the author isn't afraid to hide the ugly truth when necessary. It contains many gems." But you first need to learn linear algebra, at the minimum read all of chapter 1 on relativity, chapter 1.2.5 and {A.4} on index notation, chapter 12.12 and {A.36} on the Dirac equation, addendum {A.14} on the Klein-Gordon equation, {A.1} on Lagrangian mechanics, {A.12} on the Heisenberg interpretation, and pick up enough group theory. Learning something about the path integral approach to quantum mechanics, like from [22], cannot hurt either. In the absence of 1 000 pages and a willing author, the following discussion will truly be quantum field theory in a nanoshell.

If you want to get a start on a more advanced treatment of quantum field theory of elementary particles at a relatively low level of mathematics, Griffiths [24] is recommended.

And if you are just interested in relativistic quantum mechanics from an

intellectual point of view, there is good news. Feynman gave a set of lectures on "quantum electrodynamics" for a general audience around 1983, and the text is readily available at low cost. Without doubt, this is the best exposition of the fundamentals of quantum mechanics that has ever been written, or ever will. The subject is reduced to its bare abstract axioms, and no more can be said. If the human race is still around a millennium or so from now, artificial intelligence may take care of the needed details of quantum mechanics. But those who need or want to understand what it means will still reach for Feynman. The 2006 edition, [19], has a foreword by Zee that gives a few hints how to relate the basic concepts in the discussion to more conventional mathematics like the complex numbers found in this book. It will not be much help applying quantum field theory to engineering problems, however.

A.15.1 Occupation numbers

The first concept that must be understood in quantum field theory is occupation numbers. They will be the new way to represent quantum wave functions.

Recall first the form of wave functions in "classical" quantum mechanics, as normally covered in this book. Assume a system of independent, or maybe weakly interacting particles. The energy eigenfunctions of such a system can be written in terms of whatever are the *single-particle* energy eigenfunctions

$$\psi^{\mathrm{p}}_1(\vec{r}, S_z), \psi^{\mathrm{p}}_2(\vec{r}, S_z), \psi^{\mathrm{p}}_3(\vec{r}, S_z), \ldots$$

For each single-particle eigenfunction, $\vec{r}$ indicates the position of the particle and S_z its spin angular momentum in the chosen z-direction.

Now consider a system of, say, 36 particles. A completely arbitrary example of an energy eigenfunction for such a system would be:

$$\begin{aligned}\psi^{\mathrm{S}}(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \vec{r}_3, S_{z3}, \vec{r}_4, S_{z4}, \vec{r}_5, S_{z5}, \ldots, \vec{r}_{36}, S_{z36}) = \\ \psi^{\mathrm{p}}_{24}(\vec{r}_1, S_{z1})\psi^{\mathrm{p}}_4(\vec{r}_2, S_{z2})\psi^{\mathrm{p}}_7(\vec{r}_3, S_{z3})\psi^{\mathrm{p}}_1(\vec{r}_4, S_{z4})\psi^{\mathrm{p}}_6(\vec{r}_5, S_{z5}) \ldots \psi^{\mathrm{p}}_{54}(\vec{r}_{36}, S_{z36})\end{aligned} \tag{A.46}$$

This system eigenfunction has particle 1 in the single-particle state ψ^{p}_{24}, particle 2 in ψ^{p}_4, etcetera. The system energy is the sum of the separate energies of the 36 single-particle states involved:

$$E^{\mathrm{S}} = E^{\mathrm{p}}_{\psi_{24}} + E^{\mathrm{p}}_{\psi_4} + E^{\mathrm{p}}_{\psi_7} + E^{\mathrm{p}}_{\psi_1} + E^{\mathrm{p}}_{\psi_6} + \ldots + E^{\mathrm{p}}_{\psi_{54}}$$

Instead of writing out the example eigenfunction mathematically as done in (A.46) above, it can be graphically depicted as in figure A.2. In the figure the single-particle states are shown as boxes, and the particles that are in those particular single-particle states are shown inside the boxes. In the example, particle 1 is inside the ψ^{p}_{24} box, particle 2 is inside the ψ^{p}_4 one, etcetera. It is just the reverse from the mathematical expression (A.46): the mathematical

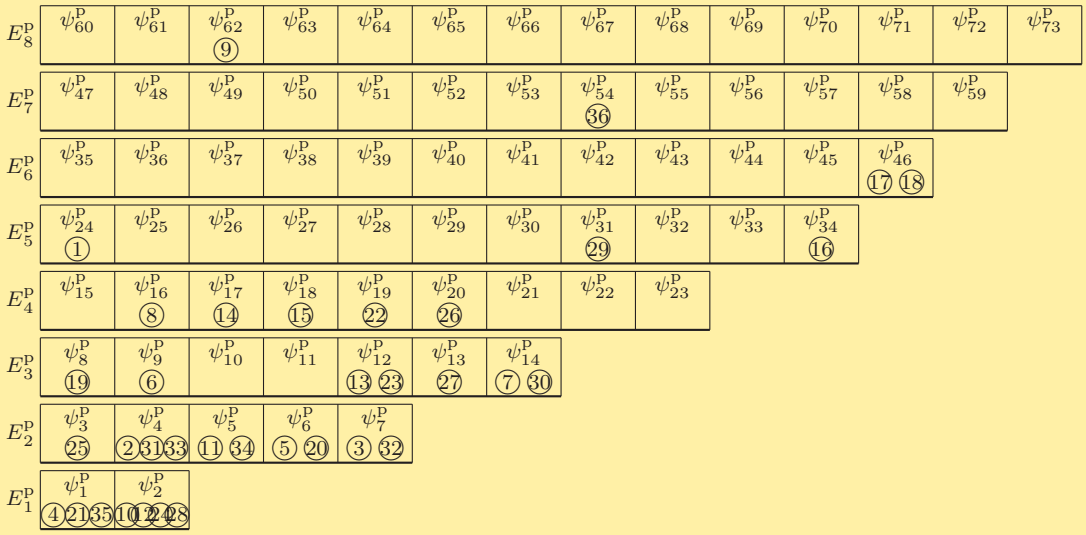


Figure A.2: Graphical depiction of an arbitrary system energy eigenfunction for 36 distinguishable particles.

expression shows for each particle in turn what the single-particle eigenstate of that particle is. The figure shows for each single-particle eigenstate in turn what particles are in that eigenstate.

However, if the 36 particles are identical bosons, (like photons or phonons), the example mathematical eigenfunction (A.46) and corresponding depiction figure A.2 is unacceptable. As chapter 5.7 explained, wave functions for bosons must be unchanged if two particles are swapped. But if, for example, particles 2 and 5 in eigenfunction (A.46) above are exchanged, it puts 2 in state 6 and 5 in state 4:

$$\psi^{\rm S}_{2\leftrightarrow 5}(\vec{r}_1,S_{z1},\vec{r}_2,S_{z2},\vec{r}_3,S_{z3},\vec{r}_4,S_{z4},\vec{r}_5,S_{z5},\ldots,\vec{r}_{36},S_{z36}) = \\ \psi^{\rm p}_{24}(\vec{r}_1,S_{z1})\psi^{\rm p}_{6}(\vec{r}_2,S_{z2})\psi^{\rm p}_{7}(\vec{r}_3,S_{z3})\psi^{\rm p}_{1}(\vec{r}_4,S_{z4})\psi^{\rm p}_{4}(\vec{r}_5,S_{z5})\ldots\psi^{\rm p}_{54}(\vec{r}_{36},S_{z36})$$

That is simply a different energy eigenfunction. So neither (A.46) nor this swapped form are acceptable by themselves. To fix up the problem, eigenfunctions must be combined. To get a valid energy eigenfunction for bosons out of (A.46), all the different eigenfunctions that can be formed by swapping the 36 particles must be summed together. The normalized sum gives the correct eigenfunction for bosons. But note that there is a humongous number of different eigenfunctions that can be obtained by swapping the particles. Over 10^{37} if you care to count them. As a result, there is no way that the gigantic expression for the resulting 36-boson energy eigenfunction could ever be written out here.

It is much easier in terms of the graphical depiction figure A.2: graphically all these countless system eigenfunctions differ only with respect to the numbers in the particles. And since in the final eigenfunction, all particles are present in exactly the same way, then so are their numbers within the particles. Every

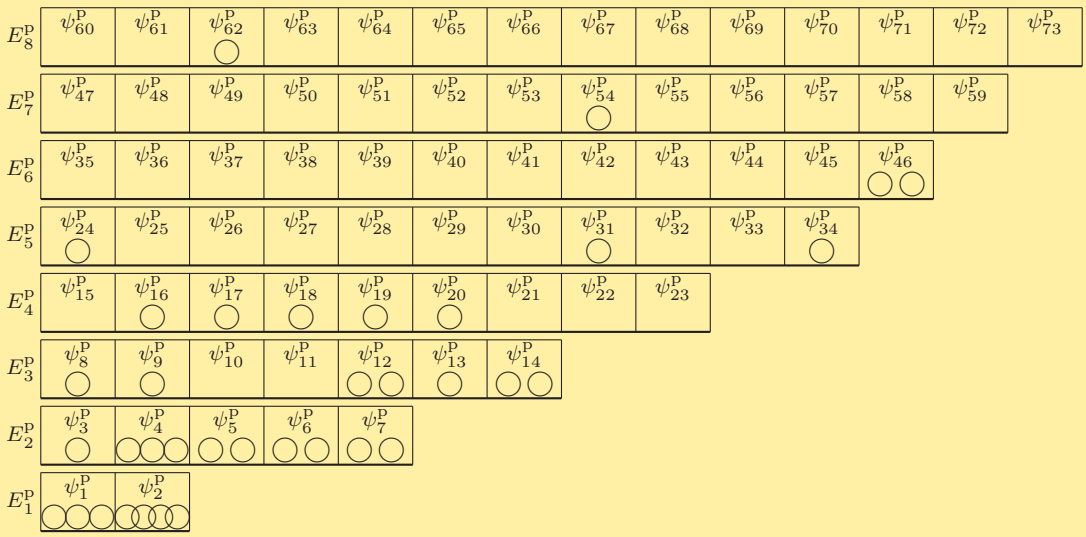


Figure A.3: Graphical depiction of an arbitrary system energy eigenfunction for 36 identical bosons.

number appears equally in every particle. So the numbers do no longer add distinguishing information and can be left out. That makes the graphical depiction of the example eigenfunction for a system of identical bosons as in figure A.3. It illustrates why identical particles are commonly called "indistinguishable."

For a system of identical fermions, (like electrons or quarks), the eigenfunctions must change sign if two particles are swapped. As chapter 5.7 showed, that is very restrictive. It means that you cannot create an eigenfunction for a system of 36 fermions from the example eigenfunction (A.46) and the swapped versions of it. Various single-particle eigenfunctions appear multiple times in (A.46), like $\psi_4^{\rm p}$, which is occupied by particles 2, 31, and 33. That cannot happen for fermions. A system eigenfunction for 36 identical fermions requires 36 different single-particle eigenfunctions.

It is the same graphically. The example figure A.3 for bosons is impossible for a system of identical fermions; there cannot be more than one fermion in a single-particle state. A depiction of an arbitrary energy eigenfunction that is acceptable for a system of 33 identical fermions is in figure A.4.

As explained in chapter 5.7, a neat way of writing down the system energy eigenfunction of the pictured example is to form a Slater determinant from the "occupied states"

$$\psi_1^{\rm p}, \psi_2^{\rm p}, \psi_3^{\rm p}, \ldots, \psi_{43}^{\rm p}, \psi_{45}^{\rm p}, \psi_{56}^{\rm p}.$$

It is good to meet old friends again, isn't it?

Now consider what happens in relativistic quantum mechanics. For example, suppose that an electron and positron annihilate each other. What are you going to do, leave holes in the parameter list of your wave function, where the electron

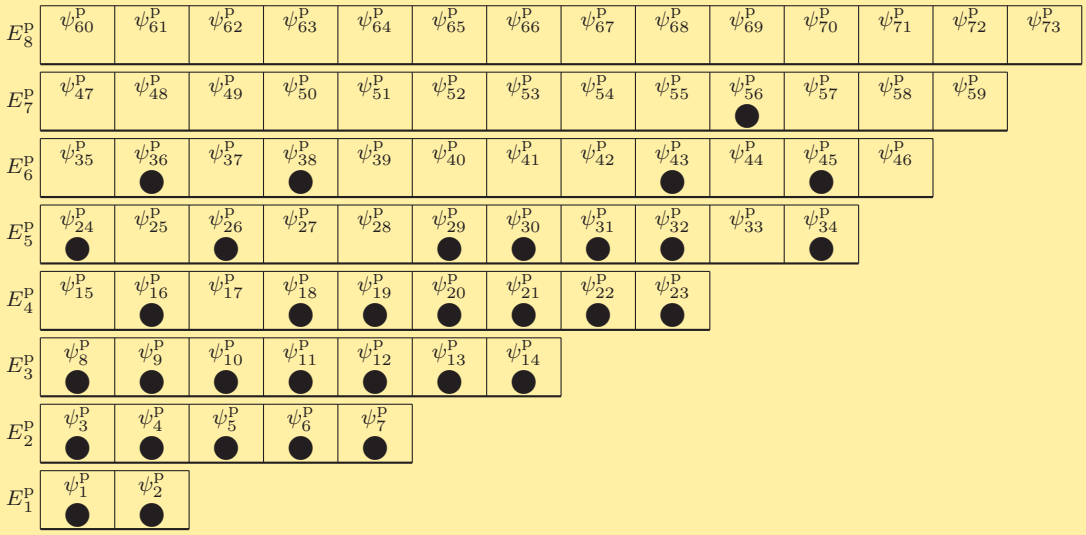


Figure A.4: Graphical depiction of an arbitrary system energy eigenfunction for 33 identical fermions.

and positron used to be? Like

$$\Psi(\vec{r}_1, S_{z1}, [\text{gone}], \vec{r}_3, S_{z3}, [\text{gone}], \vec{r}_5, S_{z5}, \ldots, \vec{r}_{36}, S_{z36}; t)$$

say? Or worse, what if a photon with very high energy hits an heavy nucleus and creates an electron-positron pair in the collision from scratch? Are you going to scribble in a set of additional parameters for the new particles into your parameter list? Scribble in another row and column in the Slater determinant for your electrons? That is voodoo mathematics. The classical way of writing wave functions fails.

And if positrons are too weird for you, consider photons, the particles of electromagnetic radiation, like ordinary light. As chapters 6.8 and 7.8 showed, the electrons in hot surfaces create and destroy photons readily when the thermal equilibrium shifts. Moving at the speed of light, with zero rest mass, photons are as relativistic as they come. Good luck scribbling in trillions of new states for the photons into your wave function when your black box heats up. Then there are solids; as chapter 11.14.6 shows, the phonons of crystal vibrational waves are the equivalent of the photons of electromagnetic waves.

One of the key insights of quantum field theory is to do away with classical mathematical forms of the wave function such as (A.46) and the Slater determinants. Instead, the graphical depictions, such as the examples in figures A.3 and A.4, are captured in terms of mathematics. How do you do that? By listing how many particles are in each type of single-particle state. In other words, you do it by listing the single-state "occupation numbers."

Consider the example bosonic eigenfunction of figure A.3. The occupation

numbers for that state would be

$$|3, 4, 1, 3, 2, 2, 2, 1, 1, 0, 0, 2, 1, 2, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, \ldots\rangle$$

indicating that there are 3 bosons in single-particle state $\psi_1^{\rm p}$, 4 in $\psi_2^{\rm p}$, 1 in $\psi_3^{\rm p}$, etcetera. Knowing those numbers is completely equivalent to knowing the classical system energy eigenfunction; it could be reconstructed from them. Similarly, the occupation numbers for the example fermionic eigenfunction of figure A.4 would be

$$|1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, \ldots\rangle$$

Such sets of occupation numbers are called "Fock basis states." Each describes one system energy eigenfunction.

General wave functions can be described by taking linear combinations of these basis states. The most general Fock wave function for a classical set of exactly I particles is a linear combination of all the basis states whose occupation numbers add up to I. But Fock states make it also possible to describe systems like photons in a box with varying numbers of particles. Then the most general wave function is a linear combination of all the Fock basis states, regardless of the total number of particles. The set of all possible wave functions that can be formed from linear combinations of the Fock basis states is called the "Fock space."

How about the case of distinguishable particles as in figure A.2? In that case, the numbers inside the particles also make a difference, so where do they go?? The answer of quantum field theory is to deny the existence of generic particles that take numbers. There are no generic particles in quantum field theory. There is a field of electrons, there is a field of protons, (or quarks, actually), there is a field of photons, etcetera, and each of these fields is granted its own set of occupation numbers. There is no way to describe a generic particle using a number. For example, if there is an electron in a single-particle state, in quantum field theory it means that the electron field has a particle in that energy state. The particle has no number.

Some physicist feel that this is a strong point in favor of believing that quantum field theory is the way nature really works. In the classical formulation of quantum mechanics, the (anti) symmetrization requirements under particle exchange are an additional ingredient, added to explain the data. In quantum field theory, it comes naturally: particles that are distinguishable simply cannot be described by the formalism. Still, our convenience in describing it is an uncertain motivator for nature.

The successful analysis of the blackbody spectrum in chapter 6.8 already testified to the usefulness of the Fock space. If you check the derivations in chapter 11 leading to it, they were all conducted based on occupation numbers. A classical wave function for the system of photons was never written down; that simply cannot be done.



Figure A.5: Example wave functions for a system with just one type of single particle state. Left: identical bosons; right: identical fermions.

There is a lot more involved in quantum field theory than just the blackbody spectrum, of course. To explain some of the basic ideas, simple examples can be helpful. The simplest example that can be studied involves just *one* single-particle state, say just a single-particle ground state. The graphical depiction of an arbitrary example wave function is then as in figure A.5. There is just one single-particle box. In nonrelativistic quantum mechanics, this would be a completely trivial quantum system. In the case of I identical bosons, shown to the left in the figure, all of them would have to go into the only state there is. In the case of identical fermions, shown to the right, there can only be one fermion, and it has to go into the only state there is.

But when particles can be created or destroyed, things get more interesting. When there is no given number of particles, there can be any number of identical bosons within that single particle state. That allows $|0\rangle$ (no particles,) $|1\rangle$ (1 particle), $|2\rangle$ (2 particles), etcetera. And the general wave function can be a linear combination of those possibilities. It is the same for identical fermions, except that there are now only the states $|0\rangle$ (no particles) and $|1\rangle$ (1 particle). The wave function can still be a combination of these two possibilities.

A relativistic system with just one type of single-particle state does seem very artificial. It raises the question how esoteric such an example is. But there are in fact two very well established classical systems that behave just like this:

1. The one-dimensional harmonic oscillator of chapter 4.1 has energy levels that happen to be exactly equally spaced. It can pick up an energy above the ground state that is any whole multiple of $\hbar\omega$, where ω is its natural frequency. If you are willing to accept the "particles" to be quanta of energy of size $\hbar\omega$, then it provides a model of a bosonic system with just one single-particle state. The ground state, h_0 in the notations of chapter 4.1, is the state $|0\rangle$. The first excited state h_1 is $|1\rangle$; it has one additional energy quantum $\hbar\omega$. The second excited state h_2 is $|2\rangle$, with two quanta more than the ground state, etcetera.

   Recall from chapter 4.1 that there is an additional ground state energy of half a $\hbar\omega$ quantum. In a quantum field theory, this additional energy that exists even when there are no particles is called the "vacuum energy."

   The general wave function of a harmonic oscillator is a linear combination of the energy states. In terms of chapter 4.1, that expresses an uncertainty in energy. In the present context, it expresses an

uncertainty in the *number* of these energy particles!

2. A single electron has exactly two spin states. It can pick up exactly one unit $\hbar$ of z-momentum above the spin-down state. If you accept the "particles" to be single quanta of z-momentum of size $\hbar$, then it provides an example of a fermionic system with just one single-particle state. There can be either 0 or 1 quantum $\hbar$ of angular momentum in that single-particle state. The general wave function is a linear combination of the state with one angular momentum "particle" and the state with no angular momentum "particle".

   This example is less intuitive, since normally when you talk about a particle, you talk about an amount of energy, like in Einstein's mass-energy relation. If it bothers you, think of the electron as being confined inside a magnetic field; then the spin-up state is associated with a corresponding increase in energy.

While the above two examples of "relativistic" systems with only one single-particle state are obviously made up, they do provide a very valuable sanity check on any relativistic analysis.

Not only that, the two examples are also very useful to understand the difference between a zero wave function and the so-called "vacuum state"

$$\boxed{|\vec{0}\rangle \equiv |0,0,0,\ldots\rangle} \tag{A.47}$$

in which all occupation numbers are zero. The vacuum state is a normalized, nonzero, wave function just like the other possible sets of occupation numbers. It describes that there are no particles with certainty. You can see it from the two examples above. For the harmonic oscillator, the state $|0\rangle$ is the ground state h_0 of the oscillator. For the electron-spin example, it is the spin-down state of the electron. These are completely normal eigenstates that the system can be in. They are *not* zero wave functions, which would be unable to describe a system.

Fock basis kets are taken to be orthonormal; an inner product between kets is zero unless all occupation numbers are equal. If they are all equal, the inner product is 1. In short:

$$\boxed{\langle \ldots, \underline{i}_3, \underline{i}_2, \underline{i}_1 | i_1, i_2, i_3, \ldots\rangle = \begin{cases} 1 \text{ if } \underline{i}_1 = i_1 \text{ and } \underline{i}_2 = i_2 \text{ and } \underline{i}_3 = i_3 \ \ldots \\ 0 \text{ otherwise} \end{cases}} \tag{A.48}$$

If the two kets have the same total number of particles, this orthonormality is required because the corresponding classical wave functions are orthonormal. Inner products between classical eigenfunctions that have even a single particle in a different state are zero. That is easily verified if the wave functions are simple products of single-particle ones. But then it also holds for sums of such eigenfunctions, as you have for bosons and fermions.

If the two kets have different total numbers of particles, the inner product between the classical wave functions does not exist. But basis kets are still orthonormal. To see that, take the two simple examples given above. For the harmonic oscillator example, different occupation numbers for the "particles" correspond to different energy eigenfunctions of the actual harmonic oscillator. These are orthonormal. It is similar for the spin example. The state of 0 "particles" is the spin-down state of the electron. The state of 1 "particle" is the spin-up state. These spin states are orthonormal states of the actual electron.

A.15.2 Creation and annihilation operators

The key to relativistic quantum mechanics is that particles can be created and annihilated. So it may not be surprising that it is very helpful to define operators that "create" and "annihilate" particles .

To keep the notations relatively simple, it will initially be assumed that there is just one type of single-particle state. Graphically that means that there is just one single-particle state box, like in figure A.5. However, there can be an arbitrary number of particles in that box.

The desired actions of the creation and annihilation operators are sketched in figure A.6. An annihilation operator $\widehat{a}$ turns a state $|i\rangle$ with i particles into a state $|i{-}1\rangle$ with $i-1$ particles. A creation operator $\widehat{a}^\dagger$ turns a state $|i\rangle$ with i particles into a state $|i{+}1\rangle$ with $i+1$ particles.

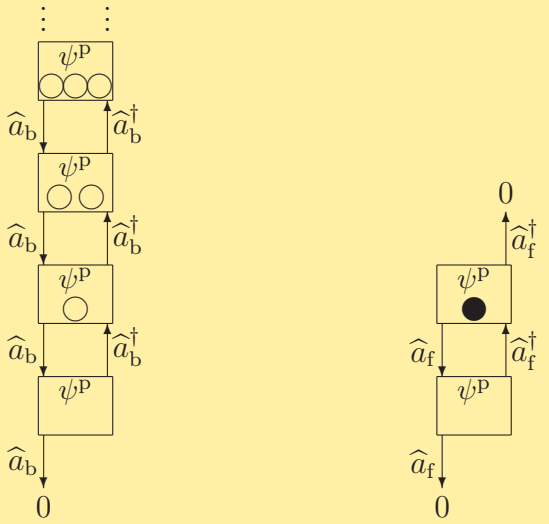


Figure A.6: Creation and annihilation operators for a system with just one type of single particle state. Left: identical bosons; right: identical fermions.

The operators are therefore *defined* by the relations

$$\widehat{a}|i\rangle = \alpha_i|i{-}1\rangle \text{ but } \widehat{a}|0\rangle = 0 \qquad \widehat{a}^\dagger|i\rangle = \alpha_i^\dagger|i{+}1\rangle \text{ but } \widehat{a}^\dagger|1\rangle = 0 \text{ for fermions} \tag{A.49}$$

Here the α_i and $\alpha_i^\dagger$ are numerical constants still to be chosen.

Note that the above relations only specify what the operators $\widehat{a}$ and $\widehat{a}^\dagger$ do to basis kets. But that is enough information to define them. To figure out what these operators do to linear combinations of basis kets, just apply them to each term in the combination separately.

Mathematically you can always define whatever operators you want. But you must hope that they will turn out to be operators that are physically helpful. To help achieve that, you want to chose the numerical constants α_i and $\alpha_i^\dagger$ appropriately. Consider what happens if the operators are applied in sequence:

$$\widehat{a}^\dagger\widehat{a}|i\rangle = \widehat{a}^\dagger\alpha_i|i{-}1\rangle = \alpha_{i-1}^\dagger\alpha_i|i\rangle$$

Reading from right to left, the order in which the operators act on the state, first $\widehat{a}$ destroys a particle, then $\widehat{a}^\dagger$ restores it again. It gives the same state back, except for the numerical factor $\alpha_{i-1}^\dagger\alpha_i$. That makes every state $|i\rangle$ an eigenvector of the operator $\widehat{a}^\dagger\widehat{a}$ with eigenvalue $\alpha_{i-1}^\dagger\alpha_i$.

If the constants $\alpha_{i-1}^\dagger$ and α_i are chosen to make the eigenvalue a real number, then the operator $\widehat{a}^\dagger\widehat{a}$ will be Hermitian. More specifically, if they are chosen to make the eigenvalue equal to i, then $\widehat{a}^\dagger\widehat{a}$ will be the "particle number operator" whose eigenvalues are the number of particles in the single-particle state. The most logical choice for the constants to achieve that is clearly

$$\alpha_i = \sqrt{i} \qquad \alpha_{i-1}^\dagger = \sqrt{i} \quad \Longrightarrow \quad \alpha_i^\dagger = \sqrt{i+1}$$

The full definition of the annihilation and creation operators can now be written in a nice symmetric way as

$$\boxed{\widehat{a}|i\rangle = \sqrt{i}\,|i{-}1\rangle \qquad \widehat{a}^\dagger|i{-}1\rangle = \sqrt{i}\,|i\rangle \quad \text{except } \widehat{a}^\dagger|1\rangle = 0 \text{ for fermions}} \tag{A.50}$$

In words, the annihilation operator $\widehat{a}$ kills off one particle and adds a factor $\sqrt{i}$. The operator $\widehat{a}^\dagger$ puts the particle back in and adds another factor $\sqrt{i}$.

These operators are particularly convenient since they are Hermitian conjugates. That means that if you take them to the other side in an inner product, they turn into each other. In particular, for inner products between basis kets,

$$\Big\langle |\underline{i}\rangle \Big| \widehat{a}|i\rangle \Big\rangle = \Big\langle \widehat{a}^\dagger|\underline{i}\rangle \Big| |i\rangle \Big\rangle \qquad \Big\langle |i\rangle \Big| \widehat{a}^\dagger|\underline{i}\rangle \Big\rangle = \Big\langle \widehat{a}|i\rangle \Big| |\underline{i}\rangle \Big\rangle$$

Note that if such relations apply for basis kets, they also apply for all linear combinations of basis kets.

To verify that the above relations apply, recall from the previous subsection that kets are orthonormal. In the equalities above, the inner products are only nonzero if $\underline{i} = i-1$: after lowering the particle number with $\widehat{a}$, or raising it with $\widehat{a}^\dagger$, the particle numbers must be the same at both sides of the inner product. And when $\underline{i} = i-1$, according to the definitions (A.50) of $\widehat{a}$ and $\widehat{a}^\dagger$ all inner products above equal $\sqrt{i}$, so the equalities still apply.

It remains true for fermions that $\widehat{a}$ and $\widehat{a}^\dagger$ are Hermitian conjugates, even though $\widehat{a}^\dagger|1\rangle = 0$ instead of $\sqrt{2}\,|2\rangle$. The reason is that the latter would only make a difference if there was a $|2\rangle$ state in the other side of the inner product, and such a state does not exist.

The inner products are usually written in the more esthetic form

$$\langle \underline{i}|\widehat{a}|i\rangle = \langle \underline{i}|\Big(\widehat{a}|i\rangle\Big) = \Big(\langle \underline{i}|\widehat{a}\Big)|i\rangle \qquad \langle i|\widehat{a}^\dagger|\underline{i}\rangle = \langle i|\Big(\widehat{a}^\dagger|\underline{i}\rangle\Big) = \Big(\langle i|\widehat{a}^\dagger\Big)|\underline{i}\rangle$$

Here it is to be understood that, say, $\langle \underline{i}|\widehat{a}$ stands for $\widehat{a}^\dagger|\underline{i}\rangle$ pushed into the left hand side of an inner product, chapter 2.7.1.

You may well wonder why $\widehat{a}^\dagger\widehat{a}$ is the particle count operator; why not $\widehat{a}\widehat{a}^\dagger$? The reason is that $\widehat{a}\widehat{a}^\dagger$ would not work for the state $|0\rangle$ unless you took $\widehat{a}^\dagger|0\rangle$ to be zero or $\widehat{a}|1\rangle$ to be zero, and then they could no longer create or annihilate $|1\rangle$.

Still, it is interesting to see what the effect of $\widehat{a}\widehat{a}^\dagger$ is. It turns out that this depends on the type of particle. For bosons, using (A.50),

$$\widehat{a}_{\rm b}\widehat{a}_{\rm b}^\dagger|i\rangle = \widehat{a}_{\rm b}\sqrt{i+1}\,|i{+}1\rangle = \sqrt{i+1}\sqrt{i+1}\,|i\rangle = (i+1)|i\rangle$$

So the operator $\widehat{a}_{\rm b}\widehat{a}_{\rm b}^\dagger$ has eigenvalues one greater than the number of particles. That means that if you subtract $\widehat{a}_{\rm b}\widehat{a}_{\rm b}^\dagger$ and $\widehat{a}_{\rm b}^\dagger\widehat{a}_{\rm b}$, you get the unit operator that leaves all states unchanged. And the difference between $\widehat{a}_{\rm b}\widehat{a}_{\rm b}^\dagger$ and $\widehat{a}_{\rm b}^\dagger\widehat{a}_{\rm b}$ is by definition the commutator of $\widehat{a}_{\rm b}$ and $\widehat{a}_{\rm b}^\dagger$, indicated by square brackets:

$$\boxed{[\widehat{a}_{\rm b}, \widehat{a}_{\rm b}^\dagger] \equiv \widehat{a}_{\rm b}\widehat{a}_{\rm b}^\dagger - \widehat{a}_{\rm b}^\dagger\widehat{a}_{\rm b} = 1} \tag{A.51}$$

Isn't that cute! Of course, $[\widehat{a}_{\rm b}, \widehat{a}_{\rm b}]$ and $[\widehat{a}_{\rm b}^\dagger, \widehat{a}_{\rm b}^\dagger]$ are zero since everything commutes with itself. It turns out that you can learn a lot from these commutators, as seen in later subsections.

The same commutator does not apply to fermions, because if you apply $\widehat{a}_{\rm f}\widehat{a}_{\rm f}^\dagger$ to $|1\rangle$, you get zero instead of $2|1\rangle$. But for fermions, the only state for which $\widehat{a}_{\rm f}\widehat{a}_{\rm f}^\dagger$ produces something nonzero is $|0\rangle$ and then it leaves the state unchanged. Similarly, the only state for which $\widehat{a}_{\rm f}^\dagger\widehat{a}_{\rm f}$ produces something nonzero is $|1\rangle$ and then it leaves that state unchanged. That means that if you *add* $\widehat{a}_{\rm f}\widehat{a}_{\rm f}^\dagger$ and $\widehat{a}_{\rm f}^\dagger\widehat{a}_{\rm f}$ together, instead of subtract them, it reproduces the same state state whether it is $|0\rangle$ or $|1\rangle$ (or any combination of them). The sum of $\widehat{a}_{\rm f}\widehat{a}_{\rm f}^\dagger$ and $\widehat{a}_{\rm f}^\dagger\widehat{a}_{\rm f}$ is called the "anticommutator" of $\widehat{a}_{\rm f}$ and $\widehat{a}_{\rm f}^\dagger$; it is indicated by curly brackets:

$$\boxed{\{\widehat{a}_{\rm f}, \widehat{a}_{\rm f}^\dagger\} \equiv \widehat{a}_{\rm f}\widehat{a}_{\rm f}^\dagger + \widehat{a}_{\rm f}^\dagger\widehat{a}_{\rm f} = 1} \tag{A.52}$$

Isn't that neat? Note also that $\{\widehat{a}_\mathrm{f}, \widehat{a}_\mathrm{f}\}$ and $\{\widehat{a}_\mathrm{f}, \widehat{a}_\mathrm{f}\}$ are zero since applying either operator twice ends up in a nonexisting state.

How about the Hamiltonian for the energy of the system of particles? Well, for noninteracting particles the energy of i particles is i times the single particle energy E^p. And since the operator that gives the number of particles is $\widehat{a}^\dagger\widehat{a}$, that is $E^\mathrm{p}\widehat{a}^\dagger\widehat{a}$. The total Hamiltonian for noninteracting particles becomes therefore:

$$\boxed{H = E^\mathrm{p}\widehat{a}^\dagger\widehat{a} + E_\mathrm{ve}} \tag{A.53}$$

Here E_ve stands for any additional "vacuum energy" that exists even if there are no particles. That is the ground state energy of the system. The above Hamiltonian allows the Schrödinger equation to be written in terms of occupation numbers and creation and annihilation operators.

A.15.3 The caHermitians

It is important to note that the creation and annihilation operators $\widehat{a}^\dagger$ and $\widehat{a}$ are not Hermitian. They cannot be taken unchanged to the other side of an inner product. And their eigenvalues are not real. Therefore they cannot correspond to physically observable quantities. But since they are Hermitian conjugates, it is easy to form operators from them that are Hermitian. For example, their products $\widehat{a}^\dagger\widehat{a}$ and $\widehat{a}\widehat{a}^\dagger$ are Hermitian. The Hamiltonian for noninteracting particles (A.53) given in the previous subsection illustrates that.

Hermitian operators can also be formed from linear combinations of the creation and annihilation operators. Two combinations that are often physically relevant are

$$\widehat{P} \equiv \tfrac{1}{2}(\widehat{a} + \widehat{a}^\dagger) \qquad \widehat{Q} \equiv \tfrac{1}{2}\mathrm{i}(\widehat{a} - \widehat{a}^\dagger)$$

In lack of a better name that the author knows of, this book will call $\widehat{P}$ and $\widehat{Q}$ the caHermitians.

Conversely, the annihilation and creation operators can be written in terms of the caHermitians as

$$\boxed{\widehat{a} = \widehat{P} - \mathrm{i}\widehat{Q} \qquad \widehat{a}^\dagger = \widehat{P} + \mathrm{i}\widehat{Q}} \tag{A.54}$$

This can be verified by substituting in the definitions of $\widehat{P}$ and $\widehat{Q}$.

The Hamiltonian (A.53) for noninteracting particles can be written in terms of $\widehat{P}$ and $\widehat{Q}$ as

$$H = E^\mathrm{p}\left(\widehat{P}^2 + \widehat{Q}^2 - \mathrm{i}[\widehat{P}, \widehat{Q}]\right) + E_\mathrm{ve}$$

Here E^p is again the single-particle energy and E_ve the vacuum energy. The square brackets indicate again the commutator of the enclosed operators.

What this Hamiltonian means depends on whether the particles being described are bosons or fermions. They have different commutators $[\widehat{P}, \widehat{Q}]$.

Consider first the case that the particles are bosons. The previous subsection showed that the commutator $[\widehat{a}_\mathrm{b}, \widehat{a}_\mathrm{b}^\dagger]$ is 1. From that the commutator of P_b and Q_b is readily found using the rules of chapter 4.5.4. It is:

$$\boxed{[\widehat{P}_\mathrm{b}, \widehat{Q}_\mathrm{b}] = -\tfrac{1}{2}\mathrm{i}} \tag{A.55}$$

So the commutator is an imaginary constant. That is very much like Heisenberg's canonical commutator between position and linear momentum in classical quantum mechanics. It implies a similar uncertainty principle, chapter 4.5.2 (4.46). In particular, P_b and Q_b cannot have definite values at the same time. Their values have uncertainties σ_{P_b} and σ_{Q_b} that are at least so big that

$$\sigma_{P_\mathrm{b}} \sigma_{Q_\mathrm{b}} \geqslant \tfrac{1}{4}$$

The Hamiltonian for bosons becomes, using the commutator above,

$$H_\mathrm{b} = E^\mathrm{p} \left(\widehat{P}_\mathrm{b}^2 + \widehat{Q}_\mathrm{b}^2 \right) + E_\mathrm{ve} - \tfrac{1}{2} E^\mathrm{p} \tag{A.56}$$

Often, the Hamiltonian is simply the first term in the right hand side. In that case, the vacuum energy is half a particle.

For fermions, the following useful relations follow from the anticommutators for the creation and annihilation operators given in the previous subsection:

$$\widehat{P}_\mathrm{f}^2 = \tfrac{1}{4} \qquad \widehat{Q}_\mathrm{f}^2 = \tfrac{1}{4} \tag{A.57}$$

The Hamiltonian then becomes

$$H = E^\mathrm{p} \left(\tfrac{1}{2} - \mathrm{i}[\widehat{P}_\mathrm{f}, \widehat{Q}_\mathrm{f}] \right) + E_\mathrm{ve} \tag{A.58}$$

A.15.4 Recasting a Hamiltonian as a quantum field one

The arguments of the previous subsection can be reversed. Given a suitable Hamiltonian, it can be recast in terms of annihilation and creation operators. This is often useful. It provides a way to quantize systems such as a harmonic oscillator or electromagnetic radiation.

Assume that some system has a Hamiltonian with the following properties:

$$\boxed{H = E^\mathrm{p} \left(\widehat{P}^2 + \widehat{Q}^2 \right) + E_\mathrm{ref} \qquad \left[\widehat{P}, \widehat{Q} \right] = -\tfrac{1}{2}\mathrm{i}} \tag{A.59}$$

Here $\widehat{P}$ and $\widehat{Q}$ must be Hermitian operators and E^p and E_ref must be constants with units of energy.

It may be noted that typically E_ref is zero. It may also be noted that it suffices that the commutator is an imaginary constant. A different magnitude of the constant can be accommodated by rescaling $\widehat{P}$ and $\widehat{Q}$, and absorbing the

scaling factor in $E^{\rm p}$. A sign change can be accommodated by swapping $\widehat{P}$ and $\widehat{Q}$.

From the given apparently limited amount of information, all of the following conclusions follow:

1. The observable quantities P and Q corresponding to the Hermitian operators are always uncertain. As explained in chapter 4.4, if you measure an uncertain quantity, say P, for a lot of identical systems, you do get some average value. That average value is called the expectation value $\langle P \rangle$. However, the individual measured values will deviate from that expectation value. The average deviation is called the standard deviation or uncertainty σ_P. For the system above, the uncertainties in P and Q must satisfy the relation

$$\sigma_P \sigma_Q \geqslant \tfrac{1}{4}$$

   Neither uncertainty can be zero, because that would make the other uncertainty infinite.
2. The expectation values of the observables P and Q satisfy the equations

$$\frac{{\rm d}\langle P \rangle}{{\rm d}t} = -\omega \langle Q \rangle \qquad \frac{{\rm d}\langle Q \rangle}{{\rm d}t} = \omega \langle P \rangle \qquad \text{where } \omega \equiv \frac{E^{\rm p}}{\hbar}$$

   That means that the expectation values vary harmonically with time,

$$\langle P \rangle = A \cos(\omega t + \alpha) \qquad \langle Q \rangle = A \sin(\omega t + \alpha)$$

   Here the "amplitude" A and the "phase angle" α are arbitrary constants.
3. In energy eigenstates, the expectation values $\langle P \rangle$ and $\langle Q \rangle$ are always zero.
4. The ground state energy of the system is

$$E_0 = \tfrac{1}{2} E^{\rm p} + E_{\rm ref}$$

   For now it will be assumed that the ground state is unique. It will be indicated as $|0\rangle$. It is often called the vacuum state.
5. The higher energy states will be indicated by $|1\rangle$, $|2\rangle$, ... in order of increasing energy E_1, E_2, ... The states are unique and their energy is

$$\text{wave function: } |i\rangle \qquad \text{energy: } E_i = (i + \tfrac{1}{2}) E_p + E_{\rm ref}$$

   So a state $|i\rangle$ has i additional "quanta" of energy $E^{\rm p}$ more than the vacuum state. In particular that means that the energy levels are equally spaced. There is no maximum energy.

6. In energy eigenstates,

$$\left\langle E^{\rm p} P^2 \right\rangle = \left\langle E^{\rm p} Q^2 \right\rangle = \tfrac{1}{2}(i+\tfrac{1}{2})E_p$$

So the expectation values of these two terms in the Hamiltonian are equal. Each contributes half to the energy of the quanta.

7. In the ground state, the two expectation energies above are the absolute minimum allowed by the uncertainty relation. Each expectation energy is then $\frac{1}{4}E^{\rm p}$.

8. Annihilation and creation operators can be defined as

$$\widehat{a} \equiv \widehat{P} - {\rm i}\widehat{Q} \qquad \widehat{a}^\dagger \equiv \widehat{P} + {\rm i}\widehat{Q}$$

These have the following effects on the energy states:

$$\widehat{a}|i\rangle = \sqrt{i}|i{-}1\rangle \qquad \widehat{a}^\dagger|i{-}1\rangle = \sqrt{i}|i\rangle$$

(This does assume that the normalization factors in the energy eigenstates are chosen consistently. Otherwise there might be additional factors of magnitude 1.) The commutator $[\widehat{a}, \widehat{a}^\dagger]$ is 1.

9. The Hamiltonian can be rewritten as

$$H = E^{\rm p}\widehat{a}^\dagger\widehat{a} + \tfrac{1}{2}E_p + E_{\rm ref}$$

Here the operator $\widehat{a}^\dagger\widehat{a}$ gives the number of energy quanta of the state it acts on.

10. If the ground state is not unique, each independent ground state gives rise to its own set of energy eigenfunctions, with the above properties. Consider the example that the system describes an electron, and that the energy does not depend on the spin. In that case, there will be a spin-up and a spin-down version of the ground state, $|0\rangle\uparrow$ and $|0\rangle\downarrow$. These will give rise to two families of energy states $|i\rangle\uparrow$ respectively $|i\rangle\downarrow$. Each family will have the properties described above.

The derivation of the above properties is really quite simple and elegant. It can be found in {D.33}.

Note that various properties above are exactly the same as found in the analysis of bosons starting with the annihilation and creation operators. The difference in this subsection is that the starting point was a Hamiltonian in terms of two square Hermitian operators; and those merely needed to have a purely imaginary commutator.

A.15.5 The harmonic oscillator as a boson system

This subsection will illustrate the power of the introduced quantum field ideas by example. The objective is to use these ideas to rederive the one-dimensional

harmonic oscillator from scratch. The derivation will be much cleaner than the elaborate algebraic derivation of chapter 4.1, and in particular {D.12}.

The Hamiltonian of a harmonic oscillator in classical quantum mechanics is, chapter 4.1,

$$H = \frac{1}{2m}\widehat{p}_x^2 + \frac{m}{2}\omega^2 x^2$$

Here the first term is the kinetic energy and the second the potential energy.

According to the previous subsection, a system like this can be solved immediately if the commutator of $\widehat{p}_x$ and x is an imaginary constant. It is, that is the famous "canonical commutator" of Heisenberg:

$$[x, \widehat{p}_x] = \mathrm{i}\hbar$$

To use the results of the previous subsection, first the Hamiltonian must be rewritten in the form

$$H = E^{\mathrm{p}} \left(\widehat{P}^2 + \widehat{Q}^2 \right)$$

where $\widehat{P}$ and $\widehat{Q}$ satisfy the commutation relationship for bosonic caHermitians:

$$\left[\widehat{P}, \widehat{Q} \right] = -\tfrac{1}{2}\mathrm{i}$$

That requires that you define

$$E^{\mathrm{p}} = \hbar\omega \qquad \widehat{P} = \sqrt{\frac{1}{2\hbar m\omega}}\, \widehat{p}_x \qquad \widehat{Q} = \sqrt{\frac{m\omega}{2\hbar}}\, x$$

According to the previous subsection, the energy eigenvalues are

$$E_i = (i + \tfrac{1}{2})\hbar\omega$$

So the spectrum has already been found.

And various other interesting properties of the solution may also be found in the previous subsection. Like the fact that there is half a quantum of energy left in the ground state. True, the zero level of energy is not important for the dynamics. But this half quantum does have a physical meaning. Assume that you have a lot of identical harmonic oscillators in the ground state, and that you do a measurement of the kinetic energy for each. You will not get zero kinetic energy. In fact, the average kinetic energy measured will be a quarter quantum, half of the total energy. The other quarter quantum is what you get on average if you do potential energy measurements.

Another observation of the previous subsection is that the expectation position of the particle will vary harmonically with time. It is a harmonic oscillator, after all.

The energy eigenfunctions will be indicated by h_i, rather than $|i\rangle$. What has not yet been found are specific expressions for these eigenfunctions. However,

as figure A.6 shows, if you apply the annihilation operator $\widehat{a}$ on the ground state h_0, you get zero:

$$\widehat{a}h_0 = 0$$

And also according to the previous subsection

$$\widehat{a} = \widehat{P} - \mathrm{i}\widehat{Q}$$

Putting in the expressions for $\widehat{P}$ and $\widehat{Q}$ above, with $\widehat{p}_x = \hbar\partial/\mathrm{i}\partial x$, and rearranging gives

$$\frac{1}{h_0}\frac{\partial h_0}{\partial x} = -\frac{m\omega}{\hbar}x$$

This can be simplified by defining a scaled x coordinate:

$$\frac{1}{h_0}\frac{\partial h_0}{\partial \xi} = -\xi \qquad \xi \equiv \frac{x}{\ell} \qquad \ell \equiv \sqrt{\frac{\hbar}{m\omega}}$$

Integrating both sides with respect to ξ and cleaning up by taking an exponential gives the ground state as

$$h_0 = Ce^{-\xi^2/2}$$

The integration constant C can be found from normalizing the wave function. The needed integral can be found under "!" in the notations section. That gives the final ground state as

$$h_0 = \frac{1}{(\pi\ell^2)^{1/4}}e^{-\xi^2/2}$$

To get the other eigenfunctions h_i for $i = 1, 2, \ldots$, apply the creation operator $\widehat{a}^\dagger$ repeatedly:

$$h_i = \frac{1}{\sqrt{i}}\widehat{a}^\dagger h_{i-1}$$

According to the previous subsection, the creation operator is

$$\widehat{a}^\dagger = \widehat{P} + \mathrm{i}\widehat{Q} = \sqrt{\frac{1}{2\hbar m\omega}}\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x} + \mathrm{i}\sqrt{\frac{m\omega}{2\hbar}}x = \frac{\mathrm{i}}{\sqrt{2}}\left(\xi - \frac{\partial}{\partial \xi}\right)$$

So the entire process involves little more than a single differentiation for each energy eigenfunction found. In particular, unlike in {D.12}, no table books are needed. Note that factors i do not make a difference in eigenfunctions. So the i in the final expression for $\widehat{a}^\dagger$ may be left out to get real eigenfunctions. That gives table 4.1.

That was easy, wasn't it?

A.15.6 Canonical (second) quantization

"Canonical quantization" is a procedure to turn a classical system into the proper quantum one. If it is applied to a field, like the electromagnetic field, it is often called "second quantization."

Recall the quantum analysis of the harmonic oscillator in the previous subsection. The key to the correct solution was the canonical commutator between position and momentum. Apparently, if you get the commutators right in quantum mechanics, you get the quantum mechanics right. That is the idea behind canonical quantization.

The basic idea can easily be illustrated for the harmonic oscillator. The standard harmonic oscillator in classical physics is a simple spring-mass system. The classical governing equations are:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v_x \qquad m\frac{\mathrm{d}v_x}{\mathrm{d}t} = -kx$$

Here x is the position of the oscillating mass m and k is the spring constant. The first of these equations is merely the definition of velocity. The second is Newton's second law.

As you can readily check by substitution, the most general solution is

$$x = A\sin(\omega t + \alpha) \qquad v_x = A\omega\cos(\omega t + \alpha) \qquad \omega \equiv \sqrt{\frac{k}{m}}$$

Here the "amplitude" A and the "phase angle" α are arbitrary constants. The "frequency" ω is given in terms of the known spring constant and mass.

This system is now to be quantized using canonical quantization. The process is somewhat round-about. First a "canonical momentum," or "conjugate momentum," or "generalized momentum," p_x is defined by taking the derivative of the kinetic energy, $\frac{1}{2}mv_x^2$, (or more generally, of the Lagrangian {A.1}), with respect to the time derivative of x. Since the time derivative of x is v_x, the momentum is mv_x. That is the usual linear momentum.

Next a classical Hamiltonian is defined. It is the total energy of the system expressed in terms of position and momentum:

$$H_{\mathrm{cl}} = \frac{p_x^2}{2m} + \frac{m}{2}\omega^2 x^2$$

Here the first term is the kinetic energy, with v_x rewritten in terms of the momentum. The second term is the potential energy in the spring. The spring constant in it was rewritten as $m\omega^2$ because m and ω are physically more important variables, and the symbol k is already greatly overworked in quantum mechanics as it is. See {A.1} for more on classical Hamiltonians.

To quantize the system, the momentum and position in the Hamiltonian must be turned into operators. Actual values of momentum and position are

then the eigenvalues of these operators. Basically, you just put a hat on the momentum and position in the Hamiltonian:

$$H = \frac{\widehat{p}_x^{\,2}}{2m} + \frac{m}{2}\omega^2 \widehat{x}^{\,2}$$

Note that the hat on x is usually omitted. However, it is still an operator in the sense that it is supposed to multiply wave functions now. Now all you need is the right commutator between $\widehat{p}_x$ and $\widehat{x}$.

In general, you identify commutators in quantum mechanics with so-called "Poisson brackets" in classical mechanics. Assume that A and B are any two quantities that depend on x and p_x. Then their Poisson bracket is defined as, {A.12},

$$\{A, B\}_{\rm P} \equiv \frac{\partial A}{\partial x}\frac{\partial B}{\partial p_x} - \frac{\partial B}{\partial x}\frac{\partial A}{\partial p_x}$$

From that it is immediately seen that

$$\{x, p_x\}_{\rm P} = 1 \qquad \{x, x\}_{\rm P} = 0 \qquad \{p_x, p_x\}_{\rm P} = 0$$

Correspondingly, in quantum mechanics you take

$$[x, \widehat{p}_x] = {\rm i}\hbar \qquad [x, x] = 0 \qquad [\widehat{p}_x, \widehat{p}_x] = 0$$

In this way the nonzero Poisson brackets bring in Planck's constant that defines quantum mechanics. (In case of fermions, anticommutators take the place of commutators.)

Because of reasons discussed for the Heisenberg picture of quantum mechanics, {A.12}, the procedure ensures that the quantum mechanics is consistent with the classical mechanics. And indeed, the results of the previous subsection confirmed that. You can check that the expectation position and momentum had the correct classical harmonic dependence on time.

Fundamentally, quantization of a classical system is just an educated guess. Classical mechanics is a special case of quantum mechanics, but quantum mechanics is not a special case of classical mechanics. For the material covered in this book, there are simpler ways to make an educated guess than canonical quantization. Being less mathematical, they are more understandable and intuitive. That might make them maybe more convincing too.

A.15.7 Spin as a fermion system

There is, of course, not much analysis that can be done with a fermion system with only one single-particle state. There are only two independent system states; no fermion or one fermion.

However, there is at least one physical example of such a simple system. As noted in subsection A.15.1, a particle with spin ½ like an electron can be

considered to be a model for it. The vacuum state $|0\rangle$ is the spin-down state of the electron. The state $|1\rangle$ is the spin-up state. This state has one unit $\hbar$ more angular momentum in the z-direction. If the electron is in a magnetic field, that additional momentum corresponds to a quantum of energy.

One reasonable question that can now be asked is whether the annihilation and creation operators, and the caHermitians, have some physical meaning for this system. They do.

Recall that for fermions, the Hamiltonian was given in terms of the caHermitians $\widehat{P}_\mathrm{f}$ and $\widehat{Q}_\mathrm{f}$ as

$$H = E^\mathrm{p} \left(\tfrac{1}{2} - \mathrm{i}[\widehat{P}_\mathrm{f}, \widehat{Q}_\mathrm{f}] \right) + E_\mathrm{ve}$$

The expression between parentheses is the particle count operator, equal to zero for the spin-down state and 1 for the spin up state. So the second term within parentheses in the Hamiltonian must be the spin in the z-direction, nondimensionalized by $\hbar$. (Recall that the spin in the z-direction has the values $\pm\frac{1}{2}\hbar$.) So apparently

$$[\widehat{P}_\mathrm{f}, \widehat{Q}_\mathrm{f}] = \mathrm{i}\frac{\widehat{S}_z}{\hbar}$$

Reasonably speaking then, the caHermitians themselves should be the nondimensional components of spin in the x and y directions,

$$\widehat{P}_\mathrm{f} = \frac{\widehat{S}_x}{\hbar} \qquad \widehat{Q}_\mathrm{f} = \frac{\widehat{S}_y}{\hbar}$$

What other variables are there in this problem? And so it is. The commutator above, with the caHermitians equal to the nondimensional spin components, is known as the "fundamental commutation relation." Quantum field analysis is one way to understand that this relation applies.

Recall another property of the caHermitians for fermions:

$$\widehat{P}_\mathrm{f}^2 = \tfrac{1}{4} \qquad \widehat{Q}_\mathrm{f}^2 = \tfrac{1}{4}$$

Apparently then, the square spin components are just constants with no uncertainty. Of course, that is no surprise since the only spin values in any direction are $\pm\frac{1}{2}\hbar$.

Finally consider the annihilation and creation operators, multiplied by $\hbar$:

$$\hbar\widehat{a} = \widehat{S}_x - \mathrm{i}\widehat{S}_y \qquad \hbar\widehat{a}^\dagger = \widehat{S}_x + \mathrm{i}\widehat{S}_y$$

Apparently these operators can remove, respectively add a unit $\hbar$ of angular momentum in the z-direction. That is often important in relativistic applications where a fermion emits or absorbs angular momentum in the z-direction. This changes the spin of the fermion and that can be expressed by the operators

above. So you will usually find x and y spin operators in the analysis of such processes.

Obviously, you can learn a lot by taking a quantum field type approach. To be sure, the current analysis applies only to particles with spin ½. But advanced analysis of angular momentum in general is very similar to quantum field analysis, chapter 12. It resembles some mixture of the boson and fermion cases.

A.15.8 More single particle states

The previous subsections discussed quantum field theory when there is just one type of single-particle state for the particles. This subsection considers the case that there is more than one. An index n will be used to number the states.

Graphically, the case of multiple single-particle states was illustrated in figures A.3 and A.4. There is now more than one box that particles can be in. Each box corresponds to one type of single-particle state $\psi^{\rm p}_n$.

Each such single-particle state has an occupation number i_n that gives the number of particles in that state. A complete set of such occupation numbers form a Fock space basis ket

$$|i_1, i_2, i_3, i_4, \ldots\rangle$$

An annihilation operator $\widehat{a}_n$ and a creation operator $\widehat{a}^\dagger_n$ must be defined for every occupation number. The mathematical definition of these operators for bosons is

$$\boxed{\begin{array}{l}\widehat{a}_{{\rm b},n}|i_1, i_2, \ldots, i_{n-1}, i_n, i_{n+1}, \ldots\rangle = \sqrt{i_n}|i_1, i_2, \ldots, i_{n-1}, i_n{-}1, i_{n+1}, \ldots\rangle \\ \\ \widehat{a}^\dagger_{{\rm b},n}|i_1, i_2, \ldots, i_{n-1}, i_n{-}1, i_{n+1}, \ldots\rangle = \sqrt{i_n}|i_1, i_2, \ldots, i_{n-1}, i_n, i_{n+1}, \ldots\rangle\end{array}} \tag{A.60}$$

The commutator relations are

$$\boxed{[\widehat{a}_{{\rm b},n}, \widehat{a}_{{\rm b},\underline{n}}] = 0 \qquad \left[\widehat{a}^\dagger_{{\rm b},n}, \widehat{a}^\dagger_{{\rm b},\underline{n}}\right] = 0 \qquad \left[\widehat{a}_{{\rm b},n}, \widehat{a}^\dagger_{{\rm b},\underline{n}}\right] = \delta_{n\underline{n}}} \tag{A.61}$$

Here $\delta_{n\underline{n}}$ is the Kronecker delta, equal to one if $n = \underline{n}$, and zero in all other cases. These commutator relations apply for $n \neq \underline{n}$ because then the operators do unrelated things to different single-particle states; in that case it does not make a difference in what order you apply them. That makes the commutator zero. For $n = \underline{n}$, the commutator relations are unchanged from the case of just one single-particle state.

For fermions it is a bit more complex. The graphical representation of the example fermionic energy eigenfunction figure A.4 cheats a bit, because it suggests that there is only one classical wave function for a given set of occupation

numbers. Actually, there are two variations, based on how the particles are ordered. The two are the same except that they have the opposite sign. Suppose that you create a particle in a state n; classically you would want to call that particle 1, and then create a particle in a state $\underline{n}$, classically you would want to call it particle 2. Do the particle creation in the opposite order, and it is particle 1 that ends up in state $\underline{n}$ and particle 2 that ends up in state n. That means that the classical wave function will have changed sign. However, the Fock space ket will not unless you do something.

What you can do is define the annihilation and creation operators for fermions as follows:

$$\boxed{\begin{aligned}
&\widehat{a}_{\mathrm{f},n}|i_1, i_2, \ldots, i_{n-1}, 0, i_{n+1}, \ldots\rangle = 0 \\
&\widehat{a}_{\mathrm{f},n}|i_1, i_2, \ldots, i_{n-1}, 1, i_{n+1}, \ldots\rangle = (-1)^{i_1+i_2+\ldots+i_{n-1}}|i_1, i_2, \ldots, i_{n-1}, 0, i_{n+1}, \ldots\rangle \\
&\widehat{a}^\dagger_{\mathrm{f},n}|i_1, i_2, \ldots, i_{n-1}, 0, i_{n+1}, \ldots\rangle = (-1)^{i_1+i_2+\ldots+i_{n-1}}|i_1, i_2, \ldots, i_{n-1}, 1, i_{n+1}, \ldots\rangle \\
&\widehat{a}^\dagger_{\mathrm{f},n}|i_1, i_2, \ldots, i_{n-1}, 1, i_{n+1}, \ldots\rangle = 0
\end{aligned}} \tag{A.62}$$

The only difference from the annihilation and creation operators for just one type of single-particle state is the potential sign changes due to the $(-1)^{\cdots}$. It adds a minus sign whenever you swap the order of annihilating/creating two particles in different states. For the annihilation and creation operators of the same state, it may change both their signs, but that does nothing much: it leaves the important products such as $\widehat{a}^\dagger_n \widehat{a}_n$ and the anticommutators unchanged.

Of course, you can *define* the annihilation and creation operators with whatever sign you want, but putting in the sign pattern above may produce easier mathematics. In fact, there is an immediate benefit already for the anticommutator relations; they take the same form as for bosons, except with anticommutators instead of commutators:

$$\boxed{\{\widehat{a}_{\mathrm{f},n}, \widehat{a}_{\mathrm{f},\underline{n}}\} = 0 \qquad \left\{\widehat{a}^\dagger_{\mathrm{f},n}, \widehat{a}^\dagger_{\mathrm{f},\underline{n}}\right\} = 0 \qquad \left\{\widehat{a}_{\mathrm{f},n}, \widehat{a}^\dagger_{\mathrm{f},\underline{n}}\right\} = \delta_{n\underline{n}}} \tag{A.63}$$

These relationships apply for $n \neq \underline{n}$ exactly because of the sign change caused by swapping the order of the operators. For $n = \underline{n}$, they are unchanged from the case of just one single-particle state.

The Hamiltonian for a system of noninteracting particles is like the one for just one single-particle state, except that you must now sum over all single-particle states:

$$\boxed{H = \sum_n E^{\mathrm{p}}_n \widehat{a}^\dagger_n \widehat{a}_n + E_{\mathrm{ve},n}} \tag{A.64}$$

A.15.9 Field operators

As noted at the start of this section, quantum field theory is particularly suited for relativistic applications because the number of particles can vary. However, in relativistic applications, it is often necessary to work in terms of position coordinates instead of single-particle energy eigenfunctions. To be sure, practical quantum field computations are usually worked out in terms of relativistic energy-momentum states. But to understand them requires consideration of position and time. Relativistic applications must make sure that coordinate systems moving at different speeds are physically equivalent and related through the Lorentz transformation. There is also the "causality problem," that an event at one location and time may not affect an event at another location and time that is not reachable with the speed of light. These conditions are posed in terms of position and time.

To handle such problems, the annihilation and creation operators can be converted into so-called "field operators" $\widehat{a}(\underline{\vec{r}})$ and $\widehat{a}^\dagger(\underline{\vec{r}})$ that annihilate respectively create particles at a given position $\underline{\vec{r}}$ in space. At least, roughly speaking that is what they do.

Now in classical quantum mechanics, a particle at a given position $\underline{\vec{r}}$ corresponds to a wave function that is nonzero at only that single point. And if the wave function is concentrated at the single point $\underline{\vec{r}}$, it must then be infinitely large at that point. Relaxing the normalization condition a bit, the appropriate infinitely concentrated mathematical function is called the "delta function," $\Psi = \delta^3(\vec{r}-\underline{\vec{r}})$, chapter 7.9. Here $\underline{\vec{r}}$ is the position of the particle and $\vec{r}$ the position at which the delta function is evaluated. If $\vec{r}$ is not equal to $\underline{\vec{r}}$, the delta function is zero; but at $\vec{r} = \underline{\vec{r}}$ it is infinite. A delta function by itself integrates to 1; its square magnitude would integrate to infinity. So it is definitely not normalized.

Like any function, a delta function can be written in terms of the single-particle energy eigenfunctions ψ_n as

$$\delta^3(\vec{r} - \underline{\vec{r}}) = \sum_{\text{all } n} c_n \psi_n(\vec{r})$$

Here the coefficients c_n can be found by taking inner products of both sides with an arbitrary eigenfunction $\psi_{\underline{n}}$. That gives, noting that orthonormality of the eigenfunctions only leaves $c_{\underline{n}}$ in the right-hand side,

$$c_{\underline{n}} = \int \psi^*_{\underline{n}}(\vec{r}) \delta^3(\vec{r} - \underline{\vec{r}}) \, \mathrm{d}^3\vec{r}$$

The integral is over all space. The index $\underline{n}$ can be renotated as n since the above expression applies for all possible values of $\underline{n}$. Also, an inner product with a delta function can easily be evaluated. The inner product above simply picks out the value of ψ^*_n at $\underline{\vec{r}}$. So

$$c_n = \psi^*_n(\underline{\vec{r}})$$

After all, $\underline{\vec{r}}$ is the only position where the delta function is nonzero. So finally

$$\delta^3(\vec{r}-\underline{\vec{r}}) = \sum_{\text{all } n} \psi_n^*(\underline{\vec{r}})\psi_n(\vec{r})$$

Since $\psi_n^*(\underline{\vec{r}})$ is the amount of eigenfunction ψ_n that must be created to create the delta function at $\underline{\vec{r}}$, the annihilation and creation field operators should presumably be

$$\widehat{a}(\underline{\vec{r}}) = \sum_n \psi_n(\underline{\vec{r}})\widehat{a}_n \qquad \widehat{a}^\dagger(\underline{\vec{r}}) = \sum_n \psi_n^*(\underline{\vec{r}})\widehat{a}_n^\dagger \tag{A.65}$$

The annihilation operator is again the Hermitian conjugate of the creation operator.

In the case of noninteracting particles in free space, the energy eigenfunctions are the momentum eigenfunctions $e^{\mathrm{i}\vec{p}\cdot\vec{r}/\hbar}$. The combination $\vec{k} = \vec{p}/\hbar$ is commonly referred to as the "wave number vector." Note that in infinite free space, the sums become integrals called Fourier transforms; see chapter 7.9 and 7.10.1 for more details.

To check the appropriateness of the creation field operator as defined above, consider its consistency with classical quantum mechanics. A classical wave function Ψ can always be written as a combination of the energy eigenfunctions:

$$\Psi(\vec{r}) = \sum_n c_n\psi_n(\vec{r}) \quad \text{where} \quad c_n = \int \psi_n^*(\vec{r})\Psi(\vec{r})\,\mathrm{d}^3\vec{r}$$

That is the same as for the delta function case above. However, any normal function also always satisfies

$$\Psi(\vec{r}) = \int \Psi(\underline{\vec{r}})\delta(\vec{r}-\underline{\vec{r}})\,\mathrm{d}^3\underline{\vec{r}}$$

That is because the delta function picks out the value of $\Psi(\underline{\vec{r}})$ at $\underline{\vec{r}} = \vec{r}$ as also already noted above. You can look at the expression above as follows: $\Psi(\vec{r})$ is a combination of position states $\delta(\vec{r}-\underline{\vec{r}})\mathrm{d}^3\underline{\vec{r}}$ with coefficients $\Psi(\underline{\vec{r}})$. So here the classical wave function is written as a combination of position states instead of energy states.

Now this needs to be converted to quantum field form. The classical wave function then becomes a combination $|\Psi\rangle$ of Fock space kets. But by definition, the creation field operator $\widehat{a}^\dagger(\underline{\vec{r}})$ applied on the vacuum state $|0\rangle$ should produce the Fock space equivalent of a delta function at $\underline{\vec{r}}$. So the above classical wave function should convert to a Fock space wave function as

$$|\Psi\rangle = \int \Psi(\underline{\vec{r}})\widehat{a}^\dagger(\underline{\vec{r}})|0\rangle\,\mathrm{d}^3\underline{\vec{r}}$$

To check that, substitute in the definition of the creation field operator:

$$|\Psi\rangle = \sum_n \int \psi_n^*(\vec{\underline{r}})\Psi(\vec{\underline{r}})\,\mathrm{d}^3\vec{\underline{r}}\;\; \widehat{a}_n^\dagger|0\rangle$$

But $\widehat{a}_n^\dagger|0\rangle$ is the Fock space equivalent of the classical energy eigenfunction ψ_n. The reason is that $\widehat{a}_n^\dagger$ puts exactly one particle in the state ψ_n. And the integral is the same coefficient c_n of this energy eigenstate as in the classical case. So the creation field operator as defined does produce the correct combination of energy states.

As a check on the appropriateness of the annihilation field operator, consider the Hamiltonian. The Hamiltonian of noninteracting particles satisfies

$$H|\Psi\rangle = \sum_n \widehat{a}_n^\dagger E_n^{\mathrm{p}} \widehat{a}_n|\Psi\rangle$$

Here E_n^{p} is the single-particle energy and $|\Psi\rangle$ stands for a state described by Fock space kets. The ground state energy was taken zero for simplicity. Note the critical role of the trailing $\widehat{a}_n$. States with no particles should not produce energy. The trailing $\widehat{a}_n$ ensures that they do not; it produces 0 when state n has no particles.

In terms of annihilation and creation field operators, you would like the Hamiltonian to be defined similarly:

$$H|\Psi\rangle = \int \widehat{a}^\dagger(\vec{r}) H^{\mathrm{p}} \widehat{a}(\vec{r})|\Psi\rangle\,\mathrm{d}^3\vec{r}$$

Note that the sum has become an integral, as $\vec{r}$ is a continuous variable. Also, the single-particle energy E^{p} has become the single-particle Hamiltonian; that is necessary because position states are not energy eigenstates with definite energy. The trailing $\widehat{a}(\vec{r})$ ensures that positions with no particles do not contribute to the Hamiltonian.

Now, if the definitions of the field operators are right, this Hamiltonian should still produce the same answer as before. Substituting in the definitions of the field operators gives

$$H|\Psi\rangle = \int \sum_{\underline{n}} \psi_{\underline{n}}^*(\vec{r})\widehat{a}_{\underline{n}}^\dagger H^{\mathrm{p}} \sum_n \psi_n(\vec{r})\widehat{a}_n|\Psi\rangle\,\mathrm{d}^3\vec{r}$$

The single-particle Hamiltonian H^{p} applied on ψ_n gives a factor E_n^{p}. And orthonormality of the eigenfunctions implies that the integral is zero unless $\underline{n} = n$. And in that case, the square energy eigenfunction magnitude integrates to 1. That then implies that the Hamiltonian is indeed the same as before.

The above argument roughly follows [43, pp. 22-29], but note that this source puts a tilde on $\widehat{a}_n^\dagger$ and $\widehat{a}_n$ as defined here. See also [35, pp. 19-24] for a somewhat

different approach, with a somewhat different definition of the annihilation and creation field operators.

One final question that is much more messy is in what sense these operators really create or annihilate a particle localized at $\underline{\vec{r}}$. An answer can be given using arguments like those used for the electromagnetic magnetic field in {A.23.4}. In particular, you want to leave some uncertainty in the number of particles created at position $\underline{\vec{r}}$. Then the expectation values for the observable field do become strongly localized near position $\underline{\vec{r}}$. The details will be skipped. But qualitatively, the fact that in quantum field theory there is uncertainty in the number of particles does of course add to the uncertainty in the measured quantities.

A big advantage of the way the annihilation and creation operators were defined now shows up: the annihilation and creation field operators satisfy essentially the same (anti)commutation relations. In particular

$$\boxed{\left[\widehat{a}_{\rm b}(\vec{r})\widehat{a}_{\rm b}(\underline{\vec{r}})\right] = 0 \qquad \left[\widehat{a}^\dagger_{\rm b}(\vec{r})\widehat{a}^\dagger_{\rm b}(\underline{\vec{r}})\right] = 0 \qquad \left[\widehat{a}_{\rm b}(\vec{r})\widehat{a}^\dagger_{\rm b}(\underline{\vec{r}})\right] = \delta^3(\vec{r}-\underline{\vec{r}})} \tag{A.66}$$

$$\boxed{\left\{\widehat{a}_{\rm f}(\vec{r})\widehat{a}_{\rm f}(\underline{\vec{r}})\right\} = 0 \qquad \left\{\widehat{a}^\dagger_{\rm f}(\vec{r})\widehat{a}^\dagger_{\rm f}(\underline{\vec{r}})\right\} = 0 \qquad \left\{\widehat{a}_{\rm f}(\vec{r})\widehat{a}^\dagger_{\rm f}(\underline{\vec{r}})\right\} = \delta^3(\vec{r}-\underline{\vec{r}})} \tag{A.67}$$

In other references you might see an additional constant multiplying the three-dimensional delta function, depending on how the position and momentum eigenfunctions were normalized.

To check these commutators, plug in the definitions of the field operators. Then the zero commutators above follow immediately from the ones for a_n and $\widehat{a}^\dagger_n$, (A.61) and (A.63). For the nonzero commutator, multiply by a completely arbitrary function $f(\vec{r})$ and integrate over $\vec{r}$. That gives $f(\underline{\vec{r}})$, which is the same result as obtained from integrating against $\delta^3(\vec{r}-\underline{\vec{r}})$. That can only be true for every function f if the commutator *is* the delta function. (In fact, producing $f(\underline{\vec{r}})$ for any $f(\vec{r})$ is exactly the way how a delta function would be defined by a conscientious mathematician.)

Field operators help solve a vexing problem for relativistic quantum mechanics: how to put space and time on equal footing, [43, p. 7ff]. Relativity unavoidably mixes up position and time. But classical quantum mechanics, as covered in this book, needs to keep them rigidly apart.

Right at the beginning, this book told you that observable quantities are the eigenvalues of Hermitian operators. That was not completely true, there is an exception. Spatial coordinates are indeed the eigenvalues of Hermitian position operators, chapter 7.9. But time is *not* an eigenvalue of an operator. When this book wrote a wave function as, say, $\Psi(\vec{r}, S_z; t)$ the time t was just a *label*. It indicated that at any given time, you have some wave function. Then you can apply purely spatial operators like x, $\widehat{p}_x$, H, etcetera to find out things about the measurable position, momentum, energy, etcetera at that time. At a different time you have a different wave function, for which you can do the same things. Time itself is left out in the cold.

Correspondingly, the classical Schrödinger equation $\mathrm{i}\hbar\partial\Psi/\partial t = H\Psi$ treats space and time quite different. The spatial derivatives, in H, are second order, but the time derivative is first order. The first-order time derivative describes the change from one spatial wave function to the next one, a time ∂t later. Of course, you cannot think of the spatial derivatives in the same way. Even if there was only one spatial coordinate instead of three, the second order spatial derivatives would not represent a change of wave function from one position to the next.

The different treatment of time and space causes problems in generalizing the Schrödinger equation to the relativistic case.

For spinless particles, the simplest generalization of the Schrödinger equation is the Klein-Gordon equation, {A.14}. However, this equation brings in states with negative energies, including negative rest mass energies. That is a problem. For example, what prevents a particle from transitioning to states of more and more negative energy, releasing infinite amounts of energy in the process? There is no clean way to deal with such problems within the bare context of the Klein-Gordon equation.

There is also the matter of what to make of the Klein-Gordon wave function. It appears as if a wave function for a single particle is being written down, like it would be for the Schrödinger equation. But for the Schrödinger equation the integrated square magnitude of the wave function is 1 and stays 1. That is taken to mean that the probability of finding the particle is 1 if you look everywhere. But the Klein-Gordon equation does not preserve the integrated square magnitude of the wave function in time. That is not surprising, since in relativity particles can be created out of energy or annihilated. But if that is so, in what sense could the Klein-Gordon equation possibly describe a wave function for a single, (i.e. exactly 1), particle?

(Of course, this is not a problem for single-particle energy eigenstates. Energy eigenstates are stationary, chapter 7.1.4. It is also not a problem if there are only particle states, or only antiparticle states, {D.32}. The real problems start when you try to add perturbations to the equation.)

For fermions with spin ½, the appropriate generalization of the Schrödinger equation is the Dirac equation, chapter 12.12. However, there are still those negative-energy solutions. Dirac postulated that all, infinitely many, negative energy states in the universe are already filled with electrons. That is obviously a rather ugly assumption. Worse, it would not work for bosons. Any number of bosons can go into a single state, they cannot fill them.

Quantum field theory can put space and time on a more equal footing, especially in the Heisenberg formulation, {A.12}. This formulation pushes time from the wave function onto the operator. To see how this works, consider some arbitrary inner product involving a Schrödinger operator $\widehat{A}$:

$$\langle \Phi | \widehat{A} \Psi \rangle$$

(Why look at inner products? Simply put, if you get all inner products right, you get the quantum mechanics right. Anything in quantum mechanics can be found by taking the right inner product.) Now recall that if a wave function Ψ has definite energy E, it varies in time as $e^{-\mathrm{i}Et/\hbar}\Psi_0$ where Ψ_0 is independent of time, chapter 7.1.2. If Ψ does not have definite energy, you can replace E in the exponential by the Hamiltonian H. (Exponentials of operators are defined by their Taylor series.) So the inner product becomes

$$\langle \Phi_0 | e^{\mathrm{i}Ht/\hbar} \widehat{A} e^{-\mathrm{i}Ht/\hbar} \Psi_0 \rangle$$

(Recall that i changes sign when taken to the other side of an inner product.) The Heisenberg $\widetilde{A}$ operator absorbs the exponentials:

$$\widetilde{A} \equiv e^{\mathrm{i}Ht/\hbar} \widehat{A} e^{-\mathrm{i}Ht/\hbar}$$

Now note that if $\widehat{A}$ is a field operator, the position coordinates in it are *not* Hamiltonian operators. They are labels just like time. They label what position the particle is annihilated or created at. So space and time are now treated much more equally.

Here is where the term "field" in "quantum field theory" comes from. In classical physics, a field is a numerical function of position. For example, a pressure field in a moving fluid has a value, the pressure, at each position. An electric field has three values, the components of the electric field, at each position. However, in quantum field theory, a "field" does not consist of values, but of operators. Each position has one or more operator associated with it. Each particle type is associated with a "field." This field will involve both creation and annihilation operators of that particle, or the associated antiparticle, at each position.

Within the quantum field framework, equations like the Klein-Gordon and Dirac ones can be given a clear meaning. The eigenfunctions of these equations give states that particles *can* be in. Since energy eigenfunctions are stationary, conservation of probability is not an issue.

It may be mentioned that there is an alternate way to put space and time on an equal footing, [43, p. 10]. Instead of turning spatial coordinates into labels, time can be turned into an operator. However, clearly wave functions do evolve with time, even if different observers may disagree about the details. So what to make of the time parameter in the Schrödinger equation? Relativity offers an answer. The time in the Schrödinger equation can be associated with the "proper" time of the considered particle. That is the time measured by an observer moving along with the particle, chapter 1.2.2. The time measured by an observer in an inertial coordinate system is then promoted to an operator. All this can be done. In fact, it is the starting point of the so-called "string theory." In string theory, a second parameter is added to proper time. You might think of the second parameter as the arc length along a string that wiggles around in

time. However, approaches along these lines are extremely complicated. Quantum field theory remains the workhorse of relativistic quantum mechanics.

A.15.10 Nonrelativistic quantum field theory

This example exercise from Srednicki [43, p. 11] uses quantum field theory to describe nonrelativistic quantum mechanics. It illustrates some of the mathematics that you will encounter in quantum field theories.

The objective is to convert the classical nonrelativistic Schrödinger equation for I particles,

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = H_{\rm cl}\Psi \tag{A.68}$$

into quantum field form. The classical wave function has the positions of the numbered particles and time as arguments:

$$\text{classical quantum mechanics:}\quad \Psi = \Psi(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots, \vec{r}_I; t) \tag{A.69}$$

where $\vec{r}_1$ is the position of particle 1, $\vec{r}_2$ is the position of particle 2, etcetera. (You could include particle spin within the vectors $\vec{r}$ if you want. But particle spin is in fact relativistic, chapter 12.12.) The classical Hamiltonian is

$$H_{\rm cl} = \sum_{i=1}^{I}\left(\frac{\hbar^2}{2m}\nabla_i^2 + V_{\rm ext}(\vec{r}_i)\right) + \tfrac{1}{2}\sum_{i=1}^{I}\sum_{\substack{\underline{i}=1\\ \underline{i}\neq i}}^{I} V(\vec{r}_i - \vec{r}_{\underline{i}}) \tag{A.70}$$

The ∇_i^2 term represents the kinetic energy of particle number i. The potential $V_{\rm ext}$ represents forces on the particles by external sources, while the potential V represents forces between particles.

In quantum field theory, the wave function for exactly I particles takes the form

$$|\Psi\rangle = \int_{\text{all }\vec{r}_1}\ldots\int_{\text{all }\vec{r}_I}\Psi(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_I; t)\,\widehat{a}^\dagger(\vec{r}_1)\widehat{a}^\dagger(\vec{r}_2)\ldots\widehat{a}^\dagger(\vec{r}_I)\,|\vec{0}\rangle\,\mathrm{d}^3\vec{r}_1\ldots\mathrm{d}^3\vec{r}_I \tag{A.71}$$

Here the ket $|\Psi\rangle$ in the left hand side is the wave function expressed as a Fock space ket. The ket $\left|\vec{0}\right\rangle$ to the far right is the vacuum state where there are no particles. However, the preceding creation operators then put in the particles at positions $\vec{r}_1$, $\vec{r}_2$, ... That produces a ket state with the particles at these positions.

The quantum amplitude of that ket state is the preceding Ψ, a function, not a ket. This is the classical nonrelativistic wave function, the one found in the nonrelativistic Schrödinger equation. After all, the classical wave function is supposed to give the quantum amplitude for the particles to be at given

positions. In particular, its square magnitude gives the probability for them to be at given positions.

So far, all this gives just the ket for one particular set of particle positions. But then it is integrated over all possible particle positions.

The Fock space Schrödinger equation for $|\Psi\rangle$ takes the form

$$\mathrm{i}\hbar\frac{\mathrm{d}|\Psi\rangle}{\mathrm{d}t} = H|\Psi\rangle \tag{A.72}$$

That looks just like the classical case. However, the Fock space Hamiltonian H is defined by

$$\begin{aligned} H|\Psi\rangle &= \int_{\text{all }\vec{r}} \widehat{a}^\dagger(\vec{r})\left[-\frac{\hbar^2}{2m}\nabla^2_{\vec{r}} + V_{\text{ext}}(\vec{r})\right]\widehat{a}(\vec{r})|\Psi\rangle\,\mathrm{d}^3\vec{r} \\ &\quad +\tfrac{1}{2}\int_{\text{all }\vec{r}}\int_{\text{all }\underline{\vec{r}}} \widehat{a}^\dagger(\vec{r})\widehat{a}^\dagger(\underline{\vec{r}})V(\vec{r}-\underline{\vec{r}})\widehat{a}(\underline{\vec{r}})\widehat{a}(\vec{r})|\Psi\rangle\,\mathrm{d}^3\vec{r}\mathrm{d}^3\underline{\vec{r}} \end{aligned} \tag{A.73}$$

In order for this to make some sense, note that the Fock space ket $|\Psi\rangle$ is an object that allows you to annihilate or create a particle at any arbitrary location $\vec{r}$. That is because it is a linear combination of basis kets that allow the same thing.

The goal is now to show that the Schrödinger equation (A.72) for the Fock space ket $|\Psi\rangle$ produces the classical Schrödinger equation (A.68) for classical wave function $\Psi(\ldots)$. This needs to be shown whether it is a system of identical bosons or a system of identical fermions.

Before trying to tackle this problem, it is probably a good idea to review representations of functions using delta functions. As the simplest example, a wave function $\Psi(x)$ of just one spatial coordinate can be written as

$$\Psi(x) = \int_{\text{all }\underline{x}} \underbrace{\Psi(\underline{x})}_{\text{coefficients}} \underbrace{\delta(x-\underline{x})\mathrm{d}\underline{x}}_{\text{basis states}}$$

The way to think about the above integral expression for $\Psi(x)$ is just like you would think about a vector in three dimensions being written as $\vec{v} = v_1\hat{\imath} + v_2\hat{\jmath} + v_3\hat{k}$ or a vector in 30 dimensions as $\vec{v} = \sum_{i=1}^{30} v_i\hat{\imath}_i$. The $\Psi(\underline{x})$ are the coefficients, corresponding to the v_i-components of the vectors. The $\delta(x-\underline{x})\mathrm{d}\underline{x}$ are the basis states, just like the unit vectors $\hat{\imath}_i$. If you want a graphical illustration, each $\delta(x-\underline{x})\mathrm{d}\underline{x}$ would correspond to one spike of unit height at a position $\underline{x}$ in figure 2.3, and you need to sum (integrate) over them all, with their coefficients, to get the total vector.

Now assume that H_1 is the one-dimensional classical Hamiltonian. Then $H_1\Psi(x)$ is just another function of x, so it can be written similarly:

$$H_1\Psi(x) = \int_{\text{all }\underline{x}} H_1\Psi(\underline{x})\delta(x-\underline{x})\,\mathrm{d}\underline{x}$$

$$= \int_{\text{all } \underline{x}} \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(\underline{x})}{\partial \underline{x}^2} + V_{\text{ext}}(\underline{x}) \Psi(\underline{x}) \right] \delta(x - \underline{x}) \, \mathrm{d}\underline{x}$$

Note that the Hamiltonian acts on the *coefficients*, not on the basis states.

You may be surprised by this, because if you straightforwardly apply the Hamiltonian H_1, in terms of x, on the integral expression for $\Psi(x)$, you get:

$$H_1 \Psi(x) = \int_{\text{all } \underline{x}} \Psi(\underline{x}) \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \delta(x - \underline{x})}{\partial x^2} + V_{\text{ext}}(x) \delta(x - \underline{x}) \right] \mathrm{d}\underline{x}$$

Here the Hamiltonian acts on the *basis states*, not the coefficients.

However, the two expressions are indeed the same. Whether there is an x or $\underline{x}$ in the potential does not make a difference, because the multiplying delta function is only nonzero when $x = \underline{x}$. And you can use a couple of integrations by parts to get the derivatives off the delta function and on $\Psi(\underline{x})$. Note here that differentiation of the delta function with respect to x or $\underline{x}$ is the same save for a sign change.

The bottom line is that you do not want to use the expression in which the Hamiltonian is applied to the basis states, because derivatives of delta functions are highly singular objects that you should not touch with a ten foot pole. (And if you have mathematical integrity, you would not really want to use delta functions either. At least not the way that they do it in physics. But in that case, you better forget about quantum field theory.)

It may here be noted that if you do have to differentiate an integral for a function $\Psi(x)$ in terms of delta functions, there is a much better way to do it. If you first make a change of integration variable to $u = \underline{x} - x$, the differentiation is no longer on the nasty delta functions.

Still, there is an important observation here: you might either know what an operator does to the coefficients, leaving the basis states untouched, or what it does to the basis states, leaving the coefficients untouched. Either one will tell you the final effect of the operator, but the mathematics is different.

Now that the general terms of engagement have been discussed, it is time to start solving Srednicki's problem. The Fock space wave function ket can be thought of the same way as the example:

$$|\Psi\rangle = \int_{\text{all } \vec{r}_1} \cdots \int_{\text{all } \vec{r}_I} \underbrace{\Psi(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_I; t)}_{\text{coefficients}} \underbrace{\widehat{a}^\dagger(\vec{r}_1) \widehat{a}^\dagger(\vec{r}_2) \ldots \widehat{a}^\dagger(\vec{r}_I) |\vec{0}\rangle \mathrm{d}^3\vec{r}_1 \ldots \mathrm{d}^3\vec{r}_I}_{\text{Fock space basis state kets}}$$

The basis states are Fock space kets in which a particle called 1 is in a delta function at a position $\vec{r}_1$, a particle called 2 in a delta function at position $\vec{r}_2$, etcetera. The classical wave function $\Psi(\ldots)$ gives the quantum amplitude of each such ket. The integration gives $|\Psi\rangle$ as a combined ket.

Note that Fock states do not know about particle numbers. A Fock basis state is the same regardless what the classical wave function calls the particles.

It means that the *same* Fock basis state ket reappears in the integration above at all swapped positions of the particles. (For fermions read: the same except possibly a sign change, since swapping the order of application of any two $\widehat{a}^\dagger$ creation operators flips the sign, compare subsection A.15.2.) This will become important at the end of the derivation.

The left hand side of the Fock space Schrödinger equation (A.72) is evaluated by pushing the time derivative inside the above integral for $|\Psi\rangle$:

$$\mathrm{i}\hbar\frac{\mathrm{d}|\Psi\rangle}{\mathrm{d}t} = \int_{\text{all }\vec{r}_1} \ldots \int_{\text{all }\vec{r}_I} \mathrm{i}\hbar\frac{\partial\Psi(\vec{r}_1,\vec{r}_2,\ldots,\vec{r}_I;t)}{\partial t}\widehat{a}^\dagger(\vec{r}_1)\ldots\widehat{a}^\dagger(\vec{r}_I)|\vec{0}\rangle\,\mathrm{d}^3\vec{r}_1\ldots\mathrm{d}^3\vec{r}_I$$

so the time derivative drops down on the classical wave function in the normal way.

Applying the Fock-space Hamiltonian (A.73) on the wave function is quite a different story, however. It is best to start with just a single particle:

$$H|\Psi\rangle = \int_{\text{all }\vec{r}}\int_{\text{all }\vec{r}_1} \widehat{a}^\dagger(\vec{r})\left[-\frac{\hbar^2}{2m}\nabla^2_{\vec{r}} + V_{\text{ext}}(\vec{r})\right]\widehat{a}(\vec{r})\Psi(\vec{r}_1;t)\widehat{a}^\dagger(\vec{r}_1)|\vec{0}\rangle\,\mathrm{d}^3\vec{r}_1\mathrm{d}^3\vec{r}$$

The field operator $\widehat{a}(\vec{r})$ may be pushed past the classical wave function $\Psi(\ldots)$; $\widehat{a}(\vec{r})$ is defined by what it does to the Fock basis states while leaving their coefficients, here $\Psi(\ldots)$, unchanged. That gives:

$$H|\Psi\rangle = \int_{\text{all }\vec{r}}\int_{\text{all }\vec{r}_1} \widehat{a}^\dagger(\vec{r})\left[-\frac{\hbar^2}{2m}\nabla^2_{\vec{r}} + V_{\text{ext}}(\vec{r})\right]\Psi(\vec{r}_1;t)\widehat{a}(\vec{r})\widehat{a}^\dagger(\vec{r}_1)|\vec{0}\rangle\,\mathrm{d}^3\vec{r}_1\mathrm{d}^3\vec{r}$$

It is now that the (anti)commutator relations become useful. The fact that for bosons $[\widehat{a}(\vec{r})\widehat{a}^\dagger(\vec{r}_1)]$ or for fermions $\{\widehat{a}(\vec{r})\widehat{a}^\dagger(\vec{r}_1)\}$ equals $\delta^3(\vec{r}-\vec{r}_1)$ means that you can swap the order of these operators as long as you add a delta function term:

$$\widehat{a}_{\mathrm{b}}(\vec{r})\widehat{a}^\dagger_{\mathrm{b}}(\vec{r}_1) = \widehat{a}^\dagger_{\mathrm{b}}(\vec{r}_1)\widehat{a}_{\mathrm{b}}(\vec{r}) + \delta^3(\vec{r}-\vec{r}_1)$$

$$\widehat{a}_{\mathrm{f}}(\vec{r})\widehat{a}^\dagger_{\mathrm{f}}(\vec{r}_1) = -\widehat{a}^\dagger_{\mathrm{f}}(\vec{r}_1)\widehat{a}_{\mathrm{f}}(\vec{r}) + \delta^3(\vec{r}-\vec{r}_1)$$

But when you swap the order of these operators, you get a factor $\widehat{a}(\vec{r})|\vec{0}\rangle$. That is zero, because applying an annihilation operator on the vacuum state produces zero, figure A.6. So the delta function term is all that remains:

$$H|\Psi\rangle = \int_{\text{all }\vec{r}}\int_{\text{all }\vec{r}_1} \widehat{a}^\dagger(\vec{r})\left[-\frac{\hbar^2}{2m}\nabla^2_{\vec{r}} + V_{\text{ext}}(\vec{r})\right]\Psi(\vec{r}_1;t)\delta^3(\vec{r}-\vec{r}_1)|\vec{0}\rangle\,\mathrm{d}^3\vec{r}_1\mathrm{d}^3\vec{r}$$

Integration over $\vec{r}_1$ now picks out the value $\Psi(\vec{r},t)$ from function $\Psi(\vec{r}_1,t)$, as delta functions do, so

$$H|\Psi\rangle = \int_{\text{all }\vec{r}} \widehat{a}^\dagger(\vec{r})\left[-\frac{\hbar^2}{2m}\nabla^2_{\vec{r}} + V_{\text{ext}}(\vec{r})\right]\Psi(\vec{r};t)|\vec{0}\rangle\,\mathrm{d}^3\vec{r}$$

Note that the term in square brackets is the classical Hamiltonian $H_{\rm cl}$ for a single particle. The creation operator $\widehat{a}^\dagger(\vec{r})$ can be pushed over the coefficient $H_{\rm cl}\Psi(\vec{r};t)$ of the vacuum state ket for the same reason that $\widehat{a}(\vec{r})$ could be pushed over $\Psi(\vec{r}_1;t)$; these operators do not affect the coefficients of the Fock states, just the states themselves.

Then, renotating $\vec{r}$ to $\vec{r}_1$, the grand total Fock state Schrödinger equation for a system of one particle becomes

$$\int_{\text{all } \vec{r}_1} \mathrm{i}\hbar \frac{\partial \Psi(\vec{r}_1;t)}{\partial t} \widehat{a}^\dagger(\vec{r}_1)|\vec{0}\rangle \, \mathrm{d}^3\vec{r}_1 =$$

$$\int_{\text{all } \vec{r}_1} \left[-\frac{\hbar^2}{2m} \nabla^2_{\vec{r}_1} + V_{\rm ext}(\vec{r}_1) \right] \Psi(\vec{r}_1;t) \widehat{a}^\dagger(\vec{r}_1)|\vec{0}\rangle \, \mathrm{d}^3\vec{r}_1$$

It is now seen that if the classical wave function $\Psi(\vec{r}_1;t)$ satisfies the classical Schrödinger equation, the Fock-space Schrödinger equation above is also satisfied. And so is the converse: if the Fock-space equation above is satisfied, the classical wave function must satisfy the classical Schrödinger equation. The reason is that Fock states can only be equal if the coefficients of all the basis states are equal, just like vectors can only be equal if all their components are equal. Here that means that the coefficient of $\widehat{a}^\dagger(\vec{r}_1)|\vec{0}\rangle$ must be the same at both sides, for every single value of $\vec{r}_1$.

If there is more than one particle, however, the equivalent latter conclusion is not justified. Remember that the *same* Fock space kets reappear in the integration at swapped positions of the particles. It now makes a difference. The following example from basic vectors illustrates the problem: yes, $a\hat{\imath} = a'\hat{\imath}$ implies that $a = a'$, but no, $(a+b)\hat{\imath} = (a'+b')\hat{\imath}$ does not imply that $a = a'$ and $b = b'$; it merely implies that $a + b = a' + b'$. However, if additionally it is postulated that the classical wave function has the symmetry properties appropriate for bosons or fermions, then the Fock-space Schrödinger equation does imply the classical one. In terms of the example from vectors, $(a+a)\hat{\imath} = (a'+a')\hat{\imath}$ does imply that $a = a'$.

In any case, the problem has been solved for a system with one particle. Doing it for I particles will be left as an exercise for your mathematical skills.

A.16 The adiabatic theorem

An adiabatic system is a system whose Hamiltonian changes slowly in time. Despite the time dependence of the Hamiltonian, the wave function can still be written in terms of the energy eigenfunctions $\psi_{\vec{n}}$ of the Hamiltonian, because the eigenfunctions are complete. But since the Hamiltonian changes with time, so do the energy eigenfunctions. And that affects how the coefficients of the eigenfunctions evolve in time.

In particular, in the adiabatic approximation, the wave function of a system can be written as, {D.34}:

$$\boxed{\Psi = \sum_{\vec{n}} c_{\vec{n}}(0) e^{\mathrm{i}\theta_{\vec{n}}} e^{\mathrm{i}\gamma_{\vec{n}}} \psi_{\vec{n}} \qquad \theta_{\vec{n}} = -\frac{1}{\hbar}\int E_{\vec{n}}\,\mathrm{d}t \qquad \gamma_{\vec{n}} = \mathrm{i}\int \langle \psi_{\vec{n}} | \psi'_{\vec{n}} \rangle\,\mathrm{d}t} \tag{A.74}$$

where the $c_{\vec{n}}(0)$ are constants. The angle $\theta_{\vec{n}}$ is called the "dynamic phase" while the angle $\gamma_{\vec{n}}$ is called the "geometric phase." Both phases are real. The prime on $\psi_{\vec{n}}$ indicates the time derivative of the eigenfunction.

Note that if the Hamiltonian does not depend on time, the above expression simplifies to the usual solution of the Schrödinger equation as given in chapter 7.1.2. In particular, in that case the geometric phase is zero and the dynamic phase is the usual $-E_{\vec{n}}t/\hbar$.

Even if the Hamiltonian depends on time, the geometric phase is still zero as long as the Hamiltonian is real. The reason is that real Hamiltonians have real eigenfunctions; then $\gamma_{\vec{n}}$ can only be real, as it must be, if it is zero.

If the geometric phase is nonzero, you may be able to play games with it. Suppose first that Hamiltonian changes with time because some single parameter λ that it depends on changes with time. Then the geometric phase can be written as

$$\gamma_{\vec{n}} = \mathrm{i}\int \langle \psi_{\vec{n}} | \frac{\partial \psi_{\vec{n}}}{\partial \lambda} \rangle\,\mathrm{d}\lambda \equiv \int f(\lambda)\,\mathrm{d}\lambda$$

It follows that if you bring the system back to the state it started out at, the total geometric phase is zero, because the limits of integration will be equal.

But now suppose that not one, but a set of parameters $\vec{\lambda} = (\lambda_1, \lambda_2, \ldots)$ changes during the evolution. Then the geometric phase is

$$\gamma_{\vec{n}} = \mathrm{i}\int \langle \psi_{\vec{n}} | \nabla_{\vec{\lambda}} \psi_{\vec{n}} \rangle \cdot \mathrm{d}\vec{\lambda} \equiv \int f_1(\lambda_1, \lambda_2, \ldots)\,\mathrm{d}\lambda_1 + f_2(\lambda_1, \lambda_2, \ldots)\,\mathrm{d}\lambda_2 + \ldots$$

and that is not necessarily zero when the system returns to the same state it started out at. In particular, for two or three parameters, you can immediately see from the Stokes' theorem that the integral along a closed path will not normally be zero unless $\nabla_{\vec{\lambda}} \times \vec{f} = 0$. The geometric phase that an adiabatic system picks up during such a closed path is called "Berry's phase."

You might assume that it is irrelevant since the phase of the wave function is not observable anyway. But if a beam of particles is sent along two different paths, the phase *difference* between the paths will produce interference effects when the beams merge again.

Systems that do not return to the same state when they are taken around a closed loop are not just restricted to quantum mechanics. A classical example is the Foucault pendulum, whose plane of oscillation picks up a daily angular deviation when the motion of the earth carries it around a circle. Such systems are called "nonholonomic" or "anholonomic."

A.17 The virial theorem

The virial theorem relates the expectation kinetic energy of a quantum system to the potential. That is of theoretical interest, as well as important for computational methods like "density functional theory."

Consider a quantum system in a state of definite energy E. In other words, consider a quantum system in a stationary state. It does not have to be the ground state. The quantum system will be assumed to be in infinite space.

To keep it simple, for now assume that there is a single particle with position vector $\vec{r}$ in a potential $V(\vec{r})$. That covers our previous examples of the harmonic oscillator and the hydrogen atom.

Then the virial theorem relates the expectation kinetic energy $\langle T \rangle$ to the potential V as follows:

$$\boxed{2\langle T \rangle = \langle \vec{r} \cdot \nabla V \rangle} \tag{A.75}$$

(Recall that nabla, ∇, is just the multi-dimensional derivative $\partial/\partial\vec{r}$.) The above formula can be very useful.

For example, consider the harmonic oscillator. There

$$V = \frac{1}{2}c_x x^2 + \frac{1}{2}c_y y^2 + \frac{1}{2}c_z z^2$$

so in Cartesian coordinates

$$\vec{r} \cdot \nabla V = x\frac{\partial V}{\partial x} + y\frac{\partial V}{\partial y} + z\frac{\partial V}{\partial z} = 2V$$

Then according to the virial theorem $2\langle T \rangle = \langle 2V \rangle$. So the expectation kinetic energy and the expectation potential energy are the same. Compute whichever is easiest, or just take half of the total energy E if you know it.

Also consider the hydrogen atom. There

$$V = -\frac{e^2}{4\pi\epsilon_0 r}$$

so in polar coordinates

$$\vec{r} \cdot \nabla V = r\frac{\partial V}{\partial r} = -V$$

Then according to the virial theorem the expectation potential energy is minus twice the expectation kinetic energy. And their sum, the total energy E, is then minus the expectation kinetic energy. In short, $\langle T \rangle = -E$ and $\langle V \rangle = 2E$ with E negative.

The virial theorem does not apply to the particle in a pipe, as that particle is in a bounded space. (You can assume infinite space if you take the potential infinite outside the pipe, but obviously by itself that does not help much. You could assume infinite space with a potential

$$V = (x/\ell_x)^p + (y/\ell_y)^p + (z/\ell_z)^p \qquad p \text{ even}$$

if you then take the limit $p \to \infty$ to get infinite potential outside the pipe and zero inside. That gives the correct but trivial result that all the energy is kinetic.)

But the virial theorem does apply to any number of particles, not just to one. Just sum over all the particles:

$$2\left\langle \sum_i T_i \right\rangle = \left\langle \sum_i \vec{r}_i \cdot \nabla_i V \right\rangle$$

where i is the particle number.

For example, consider the hydrogen molecule, where there are four particles, two protons and two electrons. Here

$$V = \sum_{i \neq j} \frac{q_i q_j}{4\pi\epsilon_0 |\vec{r}_i - \vec{r}_j|}$$

where q_i is e if particle i is a proton and $-e$ if it is an electron. Like for the simple hydrogen atom,

$$\left\langle \sum_i \vec{r}_i \cdot \nabla_i V \right\rangle = -\left\langle V \right\rangle$$

so the total expectation potential energy of the system is still twice the total energy E and the total kinetic energy is still minus E. And this continues to hold for much bigger systems of nuclei and electrons, which is why it is of interest for computational methods.

In some computations you might need to assume that the electrons are in a state of definite energy, like in the ground state, but the nuclei are not. In such computations the nuclei are at an assumed position and you will only compute the state of the electrons. So the summation in

$$\left\langle \sum_i \vec{r}_i \cdot \nabla_i V \right\rangle$$

now extends only over the electrons. But this summation does includes potentials of the electrons due to the attraction by the nuclei, and those terms are no longer equal to minus the corresponding potentials. You may need to evaluate these terms explicitly. But that is not too bad, as these potentials are now known functions of the individual electron positions only. The difficult term, due to the electron-electron interaction, is still given by minus the corresponding potential.

Finally, you might wonder where the virial theorem comes from. Well, one way to prove the virial theorem, as found in quantum textbooks and on Wikipedia, is to work out the commutator in

$$\frac{\mathrm{d}\langle \vec{r} \cdot \vec{p} \rangle}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar} \langle [H, \vec{r} \cdot \vec{p}] \rangle$$

using the formulae in chapter 4.5.4, to give

$$\frac{\mathrm{d}\langle \vec{r} \cdot \vec{p} \rangle}{\mathrm{d}t} = 2\left\langle T \right\rangle - \left\langle \vec{r} \cdot \nabla V \right\rangle,$$

and then note that the left hand side above is zero for stationary states, (in other words, for states with a precise total energy). This follows the classical way of deriving the classical virial theorem, but requires a messy purely mathematical derivation. The theorem then pops up out of the complex mathematics without any plausible physical reason why there would be such a theorem in the first place.

The original derivation by Fock in 1930 is much more physically appealing and more instructive. The idea is to slightly stretch the given quantum system: replace every position coordinate coordinate $\vec{r}$ by a slightly larger one $\vec{r}_{\rm s} = (1+\varepsilon)\vec{r}$. Here ε is assumed to be a vanishingly small number. We are interested in what the expectation potential and kinetic energy are in this slightly stretched system.

First however, recall that the square magnitude of the wave function gives the probability of that state, and that all probabilities must integrate together to 1, certainty. Phrased differently, the expectation value of one must be one; $\langle 1 \rangle = 1$, what else? But clearly, if you integrate the same square wave function magnitude over a slightly larger domain, you will get a value slightly greater than one. This problem is easily fixed, however, by multiplying the wave function in the stretched system by a suitable constant slightly less than one. Then $\langle 1 \rangle_{\rm s} = 1$ too. (The precise value of the constant depends on the number of particles and is not important.)

Next, the expectation kinetic energy consists of terms like $-\hbar^2/2m_i$ times $\left\langle \partial^2/\partial x_{{\rm s},{\rm i}}^2 \right\rangle$ because of the form of the kinetic energy operator. Because of the stretching of the coordinate in the bottom of the derivative, each of these terms changes by a factor $1/(1+\varepsilon)^2$, so

$$\langle T \rangle_{\rm s} = \langle T \rangle \frac{1}{(1+\varepsilon)^2} \approx \langle T \rangle - 2\varepsilon \langle T \rangle + \ldots$$

For the potential energy we can use a linear Taylor series to figure out how it changes:

$$V(\vec{r}_1 + \varepsilon\vec{r}_1, \vec{r}_2 + \varepsilon\vec{r}_2, \ldots) \approx V + \nabla_1 V \cdot \varepsilon\vec{r}_1 + \nabla_2 V \cdot \varepsilon\vec{r}_2 + \ldots$$

where in the right hand side V and its derivatives are evaluated at $(\vec{r}_1, \vec{r}_2, \ldots)$. From that

$$\langle V \rangle_s = \langle V \rangle + \varepsilon \left\langle \textstyle\sum_i \vec{r}_i \cdot \nabla_i V \right\rangle + \ldots$$

From the above expressions, it is seen that compared to the unstretched system, in the stretched system the sum of expectation kinetic and potential energies is different by an amount

$$\varepsilon \left(-2 \langle T \rangle + \left\langle \textstyle\sum_i \vec{r}_i \cdot \nabla_i V \right\rangle \right) + \ldots$$

But, as described in {A.7}, if you mess up an energy wave function by an amount of order ε, the expectation energy should only be messed up by an

amount proportional to ε^2, not ε. (In brief, the amounts of the other energy eigenfunctions found in the messed up wave function are proportional to ε. However, the probabilities of their energies are proportional to the squares of the amounts.) So the factor between parentheses in the expression above must be zero, and that is the virial theorem.

A.18 The energy-time uncertainty relationship

As mentioned in chapter 4.5.3, Heisenberg's formulae

$$\Delta p_x \Delta x \geqslant \frac{1}{2}\hbar$$

relating the typical uncertainties in momentum and position is often very convenient for qualitative descriptions of quantum mechanics, especially if you misread $\geqslant$ as $\approx$.

So, taking a cue from relativity, people would like to write a similar expression for the uncertainty in the time coordinate,

$$\Delta E \Delta t \geqslant \frac{1}{2}\hbar$$

The energy uncertainty can reasonably be defined as the standard deviation σ_E in energy. However, if you want to formally justify the energy-time relationship, it is not at all obvious what to make of that uncertainty in time Δt.

To arrive at one definition, assume that the variable of real interest in a given problem has a time-invariant operator A. The generalized uncertainty relationship of chapter 4.5.2 between the uncertainties in energy and A is then:

$$\sigma_E \sigma_A \geqslant \frac{1}{2}|\langle [H, A] \rangle|.$$

But according to chapter 7.2 $|\langle [H, A] \rangle|$ is just $\hbar |\mathrm{d}\langle A \rangle / \mathrm{d}t|$.

So the Mandelshtam-Tamm version of the energy-time uncertainty relationship just *defines* the uncertainty in time to be

$$\Delta t = \sigma_A \Big/ \left| \frac{\mathrm{d}\,\langle A \rangle}{\mathrm{d}t} \right|.$$

That corresponds to the typical time in which the expectation value of A changes by one standard deviation. In other words, it is the time that it takes for A to change to a value sufficiently different that it will clearly show up in measurements.

A.19 Conservation Laws and Symmetries

This note has a closer look at the relation between conservation laws and symmetries. As an example it derives the law of conservation of angular momentum directly from the rotational symmetry of physics. It then briefly explains how the arguments carry over to other conservation laws like linear momentum and parity. A simple example of a local gauge symmetry is also given. The final subsection has a few remarks about the symmetry of physics with respect to time shifts.

A.19.1 An example symmetry transformation

The mathematician Weyl gave a simple definition of a symmetry. A symmetry exists if you do something and it does not make a difference. A circular cylinder is an axially symmetric object because if you rotate it around its axis over some arbitrary angle, it still looks exactly the same. However, this note is not concerned with symmetries of objects, but of physics. That are symmetries where you do something, like place a system of particles at a different position or angle, and the physics stays the same. The system of particles itself does not necessarily need to be symmetric here.

As an example, this subsection and the next ones will explore one particular symmetry and its conservation law. The symmetry is that the physics is the same if a system of particles is placed under a different angle in otherwise empty space. There are no preferred directions in empty space. The angle that you place a system under does not make a difference. The corresponding conservation law will turn out to be conservation of angular momentum.

First a couple of clarifications. Empty space should really be understood to mean that there are no external effects on the system. A hydrogen atom in a vacuum container on earth is effectively in empty space. Or at least it is as far as its electronic structure is concerned. The energies associated with the gravity of earth and with collisions with the walls of the vacuum container are negligible. Atomic nuclei are normally effectively in empty space because the energies to excite them are so large compared to electronic energies. As a macroscopic example, to study the internal motion of the solar system the rest of the galaxy can presumably safely be ignored. Then the solar system too can be considered to be in empty space.

Further, placing a system under a different angle may be somewhat awkward. Don't burn your fingers on that hot sun when placing the solar system under a different angle. And there always seems to be a vague suspicion that you will change something nontrivially by placing the system under a different angle.

There is a different, better, way. Note that you will always need a coordinate system to describe the evolution of the system of particles mathematically. Instead of putting the system of particles under an different angle, you can put

that coordinate system under a different angle. It has the same effect. In empty space there is no reference direction to say which one got rotated, the particle system or the coordinate system. And rotating the coordinate system leaves the system truly untouched. That is why the view that the coordinate system gets rotated is called the "passive view." The view that the system itself gets rotated is called the "active view."

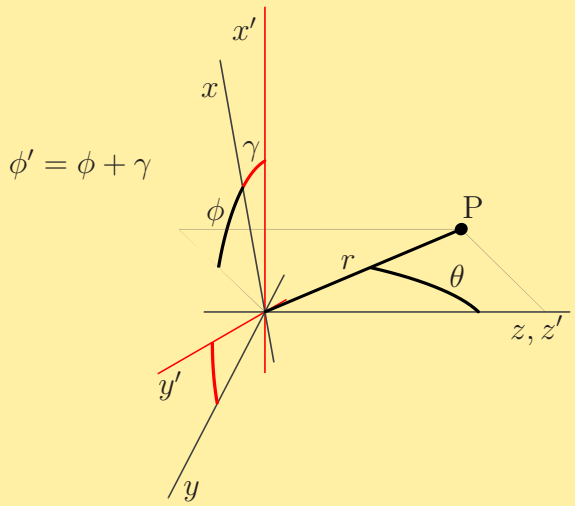


Figure A.7: Effect of a rotation of the coordinate system on the spherical coordinates of a particle at an arbitrary location P.

Figure A.7 shows graphically what happens to the position coordinates of a particle if the coordinate system gets rotated. The original coordinate system is indicated by primes. The z'-axis has been chosen along the axis of the desired rotation. Rotation of this coordinate system over an angle γ produces a new coordinate system indicated without primes. In terms of spherical coordinates, the radial position r of the particle does not change. And neither does the "polar" angle θ. But the "azimuthal" angle ϕ does change. As the figure shows, the relation between the azimuthal angles is

$$\phi' = \phi + \gamma$$

That is the basic mathematical description of the symmetry transformation.

However, it must still be applied to the description of the physics. And in quantum mechanics, the physics is described by a wave function Ψ that depends on the position coordinates of the particles;

$$\Psi(r_1, \theta_1, \phi_1, r_2, \theta_2, \phi_2, \ldots; t)$$

where 1, 2, ..., is the numbering of the particles. Particle spin will be ignored for now.

Physically absolutely nothing changes if the coordinate system is rotated. So the *values* Ψ of the wave function in the rotated coordinate system are exactly the same as the values Ψ' in the original coordinate system. But the particle coordinates corresponding to these values do change:

$$\Psi(r_1,\theta_1,\phi_1,r_2,\theta_2,\phi_2,\ldots;t) = \Psi'(r_1',\theta_1',\phi_1',r_2',\theta_2',\phi_2',\ldots;t)$$

Therefore, considered as *functions*, Ψ' and Ψ are different. However, only the azimuthal angles change. In particular, putting in the relation between the azimuthal angles above gives:

$$\Psi(r_1,\theta_1,\phi_1,r_2,\theta_2,\phi_2,\ldots;t) = \Psi'(r_1,\theta_1,\phi_1+\gamma,r_2,\theta_2,\phi_2+\gamma,\ldots;t)$$

Mathematically, changes in functions are most conveniently written in terms of an appropriate operator, chapter 2.4. The operator here is called the "generator of rotations around the z-axis." It will be indicated as $\mathcal{R}_{z,\gamma}$. What it does is add γ to the azimuthal angles of the function. By definition:

$$\mathcal{R}_{z,\gamma}\Psi'(r_1,\theta_1,\phi_1,r_2,\theta_2,\phi_2,\ldots;t) \equiv \Psi'(r_1,\theta_1,\phi_1+\gamma,r_2,\theta_2,\phi_2+\gamma,\ldots;t)$$

In terms of this operator, the relationship between the wave functions in the rotated and original coordinate systems can be written concisely as

$$\Psi = \mathcal{R}_{z,\gamma}\Psi'$$

Using $\mathcal{R}_{z,\gamma}$, there is no longer a need for using primes on one set of coordinates. Take any wave function in terms of the original coordinates, written without primes. Application of $\mathcal{R}_{z,\gamma}$ will turn it into the corresponding wave function in the rotated coordinates, also written without primes.

So far, this is all mathematics. The above expression applies whether or not there is symmetry with respect to rotations. It even applies whether or not Ψ is a wave function.

A.19.2 Physical description of a symmetry

The next question is what it means in terms of physics that empty space has no preferred directions. According to quantum mechanics, the Schrödinger equation describes the physics. It says that the time derivative of the wave function can be found as

$$\frac{\partial\Psi}{\partial t} = \frac{1}{\mathrm{i}\hbar}H\Psi$$

where H is the Hamiltonian. If space has no preferred directions, then the Hamiltonian must be the same regardless of angular orientation of the coordinate system used.

In particular, consider the two coordinate systems of the previous subsection. The second system differed from the first by a rotation over an arbitrary angle γ around the z-axis. If one system had a different Hamiltonian than the other, then systems of particles would be observed to evolve in a different way in that coordinate system. That would provide a fundamental distinction between the two coordinate system orientations right there.

A couple of very basic examples can make this more concrete. Consider the electronic structure of the hydrogen atom as analyzed in chapter 4.3. The electron was *not* in empty space in that analysis. It was around a proton, which was assumed to be at rest at the origin. However, the electric field of the proton has no preferred direction either. (Proton spin was ignored). Therefore the current analysis *does* apply to the electron of the hydrogen atom. In terms of Cartesian coordinates, the Hamiltonian in the original x', y', z' coordinate system is

$$H' = -\frac{\hbar^2}{2m_{\rm e}}\left[\frac{\partial}{\partial x'^2}+\frac{\partial}{\partial y'^2}+\frac{\partial}{\partial z'^2}\right] - \frac{e^2}{4\pi\epsilon_0}\frac{1}{\sqrt{x'^2+y'^2+z'^2}}$$

The first term is the kinetic energy operator. It is proportional to the Laplacian operator, inside the square brackets. Standard vector calculus says that this operator is independent of the angular orientation of the coordinate system. So to get the corresponding operator in the rotated x, y, z coordinate system, simply leave away the primes. The second term is the potential energy in the field of the proton. It is inversely proportional to the distance of the electron from the origin. The expression for the distance from the origin is the same in the rotated coordinate system. Once again, just leave away the primes. The bottom line is that you cannot see a difference between the two coordinate systems by looking at their Hamiltonians. The expressions for the Hamiltonians are identical.

As a second example, consider the analysis of the complete hydrogen atom as described in addendum {A.5}. The complete atom was assumed to be in empty space; there were no external effects on the atom included. The analysis still ignored all relativistic effects, including the electron and proton spins. However, it did include the motion of the proton. That meant that the kinetic energy of the proton had to be added to the Hamiltonian. But that too is a Laplacian, now in terms of the proton coordinates $x'_{\rm p}, y'_{\rm p}, z'_{\rm p}$. Its expression too is the same regardless of angular orientation of the coordinate system. And in the potential energy term, the distance from the origin now becomes the distance between electron and proton. But the formula for the distance between two points is the same regardless of angular orientation of the coordinate system. So once again, the expression for the Hamiltonian does not depend on the angular orientation of the coordinate system.

The equality of the Hamiltonians in the original and rotated coordinate systems has a consequence. It leads to a mathematical requirement for the operator

$\mathcal{R}_{z,\gamma}$ of the previous subsection that describes the effect of a coordinate system rotation on wave functions. This operator must commute with the Hamiltonian:

$$H\mathcal{R}_{z,\gamma} = \mathcal{R}_{z,\gamma}H$$

That follows from examining the wave function of a system as seen in both the original and the rotated coordinate system. There are two ways to find the time derivative of the wave function in the rotated coordinate system. One way is to rotate the original wave function using $\mathcal{R}_{z,\gamma}$ to get the one in the rotated coordinate system. Then you can apply the Hamiltonian on that. The other way is to apply the Hamiltonian on the wave function in the original coordinate system to find the time derivative in the original coordinate system. Then you can use $\mathcal{R}_{z,\gamma}$ to convert that time derivative to the rotated system. The Hamiltonian and $\mathcal{R}_{z,\gamma}$ get applied in the opposite order, but the result must still be the same.

This observation can be inverted to define a symmetry of physics in general:

> *A symmetry of physics is described by a unitary operator that commutes with the Hamiltonian.*

If an operator commutes with the Hamiltonian, then the same Hamiltonian applies in the changed coordinate system. So there is no physical difference in how systems evolve between the two coordinate systems.

The qualification "unitary" means that the operator should not change the magnitude of the wave function. The wave function should remain normalized. It does for the transformations of interest in this note, like rotations of the coordinate system, shifts of the coordinate system, time shifts, and spatial coordinate inversions. All of these transformations are unitary. Like Hermitian operators, unitary operators have a complete set of orthonormal eigenfunctions. However, the eigenvalues are normally not real numbers.

For those who wonder, time reversal is somewhat of a special case. To understand the difficulty, consider first the operation "take the complex conjugate of the wave function." This operator preserves the magnitude of the wave function. And it commutes with the Hamiltonian, assuming a basic real Hamiltonian. But taking complex conjugate is not a linear operator. For a linear operator $(\mathrm{i}\Psi)' = \mathrm{i}(\Psi)'$. But $(\mathrm{i}\Psi)^* = -\mathrm{i}\Psi^*$. If constants come out of an operator as complex conjugates, the operator is called "antilinear." So taking complex conjugate is antilinear. Another issue: a linear unitary operator preserves the inner products between any two wave functions Ψ_1 and Ψ_2. (That can be verified by expanding the square magnitudes of $\Psi_1 + \Psi_2$ and $\Psi_1 + \mathrm{i}\Psi_2$). However, taking complex conjugate changes inner products into their complex conjugates. Operators that do that are called "antiunitary." So taking complex conjugate is both antilinear and antiunitary. (Of course, in normal language it is neither. The appropriate terms would have been conjugate-linear and conjugate-unitary. But if you got

this far in this book, you know how much chance appropriate terms have of being used in physics.)

Now the effect of time-reversal on wave functions turns out to be antilinear and antiunitary too, [49, p. 76]. One simple way to think about it is that a straightforward time reversal would change $e^{-\mathrm{i}Et/\hbar}$ into $e^{\mathrm{i}Et/\hbar}$. Then an additional complex conjugate will take things back to positive energies. For the same reason you do *not* want to add a complex conjugate to spatial transformations or time shifts.

A.19.3 Derivation of the conservation law

The definition of a symmetry as an operator that commutes with the Hamiltonian may seem abstract. But it has a less abstract consequence. It implies that the eigenfunctions of the symmetry operation can be taken to be also eigenfunctions of the Hamiltonian, {D.18}. And, as chapter 7.1.4 discussed, the eigenfunctions of the Hamiltonian are stationary. They change in time by a mere scalar factor $e^{\mathrm{i}Et/\hbar}$ of magnitude 1 that does not change their physical properties.

The fact that the eigenfunctions do not change is responsible for the conservation law. Consider what a conservation law really means. It means that there is some number that does not change in time. For example, conservation of angular momentum in the z-direction means that the net angular momentum of the system in the z-direction, a number, does not change.

And if the system of particles is described by an eigenfunction of the symmetry operator, then there is indeed a number that does not change: the eigenvalue of that eigenfunction. The scalar factor $e^{\mathrm{i}Et/\hbar}$ changes the eigenfunction, but not the eigenvalue that would be produced by applying the symmetry operator at different times. The eigenvalue can therefore be looked upon as a specific value of some conserved quantity. In those terms, if the state of the system is given by a different eigenfunction, with a different eigenvalue, it has a different value for the conserved quantity.

> *The eigenvalues of a symmetry of physics describe the possible values of a conserved quantity.*

Of course, the system of particles might not be described by a single eigenfunction of the symmetry operator. It might be a mixture of eigenfunctions, with different eigenvalues. But that merely means that there is quantum mechanical uncertainty in the conserved quantity. That is just like there may be uncertainty in energy. Even if there is uncertainty, still the mixture of eigenvalues does not change with time. Each eigenfunction is still stationary. Therefore the probability of getting a given value for the conserved quantity does not change with time. In particular, neither the expectation value of the conserved quantity, nor the amount of uncertainty in it changes with time.

The eigenvalues of a symmetry operator may require some cleaning up. They may not directly give the conserved quantity in the desired form. Consider for example the eigenvalues of the rotation operator $\mathcal{R}_{z,\gamma}$ discussed in the previous subsections. You would surely expect a conserved quantity of a system to be a real quantity. But the eigenvalues of $\mathcal{R}_{z,\gamma}$ are in general complex numbers.

The one thing that can be said about the eigenvalues is that they are always of magnitude 1. Otherwise an eigenfunction would change in magnitude during the rotation. But a function does not change in magnitude if it is merely viewed under a different angle. And if the eigenvalues are of magnitude 1, then the Euler formula (2.5) implies that they can always be written in the form

$$e^{\mathrm{i}\alpha}$$

where α is some real number. If the eigenvalue does not change with time, then neither does α, which is basically just its logarithm.

But although α is real and conserved, still it is not the desired conserved quantity. Consider the possibility that you perform another rotation of the axis system. Each rotation multiplies the eigenfunction by a factor $e^{\mathrm{i}\alpha}$ for a total of $e^{2\mathrm{i}\alpha}$. In short, if you double the angle of rotation γ, you also double the value of α. But it does not make sense to say that both α and 2α are conserved. If α is conserved, then so is 2α; that is not a second conservation law. Since α is proportional to γ, it can be written in the form

$$\alpha = m\gamma$$

where the constant of proportionality m is independent of the amount of coordinate system rotation.

The constant m is the desired conserved quantity. For historical reasons it is called the "magnetic quantum number." Unfortunately, long before quantum mechanics, classical physics had already figured out that something was preserved. It called that quantity the "angular momentum" L_z. It turns out that what classical physics defines as angular momentum is simply a multiple of the magnetic quantum number:

$$L_z = m\hbar$$

So conservation of angular momentum is the same thing as conservation of magnetic quantum number.

But the magnetic quantum number is more fundamental. Its possible values are pure integers, unlike those of angular momentum. To see why, note that in terms of m, the eigenvalues of $\mathcal{R}_{z,\gamma}$ are of the form

$$e^{\mathrm{i}m\gamma}$$

Now if you rotate the coordinate system over an angle $\gamma = 2\pi$, it gets back to the exact same position as it was in before the rotation. The wave function

should not change in that case, which means that the eigenvalue must be equal to one. And that requires that the value of m is an integer. If m was a fractional number, $e^{\mathrm{i}m2\pi}$ would not be 1.

It may be interesting to see how all this works out for the two examples mentioned in the previous subsection. The first example was the electron in a hydrogen atom where the proton is assumed to be at rest at the origin. Chapter 4.3 found the electron energy eigenfunctions in the form

$$\psi_{nlm}(\vec{r}) = R_{nl}(r)Y_l^m(\theta,\phi) = R_{nl}(r)\Theta_l^m(\theta)e^{\mathrm{i}m\phi}$$

It is the final exponential that changes by the expected factor $e^{\mathrm{i}m\gamma}$ when $\mathcal{R}_{z,\gamma}$ replaces ϕ by $\phi+\gamma$.

The second example was the complete hydrogen atom in empty space. In addendum {A.5}, the energy eigenfunctions were found in the form

$$\psi_{nlm,\mathrm{red}}(\vec{r}-\vec{r}_{\mathrm{p}})\psi_{\mathrm{cg}}(\vec{r}_{\mathrm{cg}})$$

The first term is like before, except that it is computed with a "reduced mass" that is slightly different from the true electron mass. The argument is now the difference in position between the electron and the proton. It still produces a factor $e^{\mathrm{i}m\gamma}$ when $\mathcal{R}_{z,\gamma}$ is applied. The second factor reflects the motion of the center of gravity of the complete atom. If the center of gravity has definite angular momentum around whatever point is used as origin, it will produce an additional factor $e^{\mathrm{i}m_{\mathrm{cg}}\gamma}$. (See addendum {A.6} on how the energy eigenfunctions ψ_{cg} can be written as spherical Bessel functions of the first kind times spherical harmonics that have definite angular momentum. But also see chapter 7.9 about the nasty normalization issues with wave functions in infinite empty space.)

As a final step, it is desirable to formulate a nicer operator for angular momentum. The rotation operators $\mathcal{R}_{z,\gamma}$ are far from perfect. One problem is that there are infinitely many of them, one for every angle γ. And they are all related, a rotation over an angle 2γ being the same as two rotations over an angle γ.

If you define a rotation operator over a very small angle, call it $\mathcal{R}_{z,\varepsilon}$, then you can approximate any other operator $\mathcal{R}_{z,\gamma}$ by just applying $\mathcal{R}_{z,\varepsilon}$ sufficiently many times. To make this approximation exact, you need to make ε infinitesimally small. But when ε becomes zero, $\mathcal{R}_{z,\varepsilon}$ would become just 1. You have lost the nicer operator that you want by going to the extreme. The trick to avoid this is to subtract the limiting operator 1, and in addition, to avoid that the resulting operator then becomes zero, you must also divide by ε. The nicer operator is therefore

$$\lim_{\varepsilon\to 0}\frac{\mathcal{R}_{z,\varepsilon}-1}{\varepsilon}$$

Now consider what this operator really means for a single particle with no spin:

$$\lim_{\varepsilon\to 0}\frac{\mathcal{R}_{z,\varepsilon}-1}{\varepsilon}\Psi(r,\theta,\phi) = \lim_{\varepsilon\to 0}\frac{\Psi(r,\theta,\phi+\varepsilon)-\Psi(r,\theta,\phi)}{\varepsilon}$$

By definition, the final term is the partial derivative of Ψ with respect to ϕ. So the new operator is just the operator $\partial/\partial\phi$!

You can go one better still, because the eigenvalues of the operator just defined are

$$\lim_{\varepsilon\to 0} \frac{e^{\mathrm{i}m\varepsilon} - 1}{\varepsilon} = \mathrm{i}m$$

If you add a factor $\hbar/\mathrm{i}$ to the operator, the eigenvalues of the operator are going to be $m\hbar$, the quantity defined in classical physics as the angular momentum. So you are led to define the angular momentum operator of a single particle as:

$$\widehat{L}_z \equiv \frac{\hbar}{\mathrm{i}} \frac{\partial}{\partial\phi}$$

This agrees perfectly with what chapter 4.2.2 got from guessing that the relationship between angular and linear momentum is the same in quantum mechanics as in classical mechanics.

The angular momentum operator of a general system can be defined using the same scale factor:

$$\boxed{\widehat{L}_z \equiv \frac{\hbar}{\mathrm{i}} \lim_{\varepsilon\to 0} \frac{\mathcal{R}_{z,\varepsilon} - 1}{\varepsilon}} \tag{A.76}$$

The system has definite angular momentum $m\hbar$ if

$$\widehat{L}_z\Psi = m\hbar\Psi$$

Consider now what happens if the angular operator $\widehat{L}_z$ as defined above is applied to the wave function of a system of multiple particles, still without spin. It produces

$$\widehat{L}_z\Psi = \frac{\hbar}{\mathrm{i}} \lim_{\varepsilon\to 0} \frac{\Psi(r_1,\theta_1,\phi_1+\varepsilon,r_2,\theta_2,\phi_2+\varepsilon,\ldots) - \Psi(r_1,\theta_1,\phi_1,r_2,\theta_2,\phi_2,\ldots)}{\varepsilon}$$

The limit in the right hand side is a total derivative. According to calculus, it can be rewritten in terms of partial derivatives to give

$$\widehat{L}_z\Psi = \frac{\hbar}{\mathrm{i}} \left[\frac{\partial}{\partial\phi_1} + \frac{\partial}{\partial\phi_2} + \ldots \right] \Psi$$

The scaled derivatives in the new right hand side are the orbital angular momenta of the individual particles as defined above, so

$$\widehat{L}_z\Psi = \left[\widehat{L}_{z,1} + \widehat{L}_{z,2} + \ldots \right] \Psi$$

It follows that the angular momenta of the individual particles just add, like they do in classical physics.

Of course, even if the complete system has definite angular momentum, the individual particles may not. A particle numbered i has definite angular momentum $m_i\hbar$ if

$$\widehat{L}_{z,i}\Psi \equiv \frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial\phi_i}\Psi = m_i\hbar\Psi$$

If every particle has definite momentum like that, then these momenta directly add up to the total system momentum. At the other extreme, if both the system and the particles have uncertain angular momentum, then the expectation values of the momenta of the particles still add up to that of the system.

Now that the angular momentum operator has been defined, the generator of rotations $\mathcal{R}_{z,\gamma}$ can be identified in terms of it. It turns out to be

$$\boxed{\mathcal{R}_{z,\gamma} = \exp\left(\frac{\mathrm{i}}{\hbar}\widehat{L}_z\gamma\right)} \tag{A.77}$$

To check that it does indeed take the form above, expand the exponential in a Taylor series. Then apply it on an eigenfunction with angular momentum $L_z = m\hbar$. The effect is seen to be to multiply the eigenfunction by the Taylor series of $e^{\mathrm{i}m\gamma}$ as it should. So $\mathcal{R}_{z,\gamma}$ as given above gets all eigenfunctions right. It must therefore be correct since the eigenfunctions are complete.

Now consider the generator of rotations in terms of the individual particles. Since $\widehat{L}_z$ is the sum of the angular momenta of the individual particles,

$$\mathcal{R}_{z,\gamma} = \exp\left(\frac{\mathrm{i}}{\hbar}\widehat{L}_{z,1}\gamma\right)\exp\left(\frac{\mathrm{i}}{\hbar}\widehat{L}_{z,2}\gamma\right)\ldots$$

So, while the contributions of the individual particles to total angular momentum *add* together, their contributions to the generator of rotations *multiply* together. In particular, if a particle i has definite angular momentum $m_i\hbar$, then it contributes a factor $e^{\mathrm{i}m_i\gamma}$ to $\mathcal{R}_{z,\gamma}$.

How about spin? The normal angular momentum discussed so far suggests its true meaning. If a particle i has definite spin angular momentum in the z-direction $m_{s,i}\hbar$, then presumably the wave function changes by an additional factor $e^{\mathrm{i}m_{s,i}\gamma}$ when you rotate the axis system over an angle γ.

But there is something curious here. If the axis system is rotated over an angle 2π, it is back in its original position. So you would expect that the wave function is also again the same as before the rotation. And if there is just orbital angular momentum, then that is indeed the case, because $e^{\mathrm{i}m2\pi} = 1$ as long as m is an integer, (2.5). But for fermions the spin angular momentum m_s in a given direction is half-integer, and $e^{\mathrm{i}\pi} = -1$. Therefore the wave function of a fermion changes sign when the coordinate system is rotated over 2π and is back in its original position. That is true even if there is uncertainty in the spin angular momentum. For example, the wave function of a fermion with spin $\frac{1}{2}$

can be written as, chapter 5.5.1,

$$\Psi_+\uparrow + \Psi_-\downarrow$$

where the first term has $\frac{1}{2}\hbar$ angular momentum in the z-direction and the second term $-\frac{1}{2}\hbar$. Each term changes sign under a turn of the coordinate system by 2π. So the complete wave function changes sign. More generally, for a system with an odd number of fermions the wave function changes sign when the coordinate system is rotated over 2π. For a system with an even number of fermions, the wave function returns to the original value.

Now the sign of the wave function does not make a difference for the observed physics. But it is still somewhat unsettling to see that on the level of the wave function, nature is only the same when the coordinate system is rotated over 4π instead of 2π. (However, it may be only a mathematical artifact. The anti-symmetrization requirement implies that the true system includes all electrons in the universe. Presumably, the number of fermions in the universe is infinite. That makes the question whether the number is odd or even unanswerable. If the number of fermions does turn out to be finite, this book will reconsider the question when people finish counting.)

(Some books now raise the question why the orbital angular momentum functions could not do the same thing. Why could the quantum number of orbital angular momentum not be half-integer too? But of course, it is easy to see why not. If the spatial wave function would be multiple valued, then the momentum operators would produce infinite momentum. You would have to postulate arbitrarily that the derivatives of the wave function at a point only involve wave function values of a single branch. Half-integer spin does not have the same problem; for a given orientation of the coordinate system, the opposite wave function is not accessible by merely changing position.)

A.19.4 Other symmetries

The previous subsections derived conservation of angular momentum from the symmetry of physics with respect to rotations. Similar arguments may be used to derive other conservation laws. This subsection briefly outlines how.

Conservation of linear momentum can be derived from the symmetry of physics with respect to translations. The derivation is completely analogous to the angular momentum case. The translation operator $\mathcal{T}_{z,d}$ shifts the coordinate system over a distance d in the z-direction. Its eigenvalues are of the form

$$e^{\mathrm{i}k_z d}$$

where k_z is a real number, independent of the amount of translation d, that is called the wave number. Following the same arguments as for angular momentum, k_z is a preserved quantity. In classical physics not k_z, but $p_z = \hbar k_z$ is

defined as the conserved quantity. To get the operator for this quantity, form the operator

$$\boxed{\widehat{p}_z = \frac{\hbar}{\mathrm{i}} \lim_{\varepsilon \to 0} \frac{\mathcal{T}_{z,\varepsilon} - 1}{\varepsilon}} \tag{A.78}$$

For a single particle, this becomes the usual linear momentum operator $\hbar\partial/\mathrm{i}\partial z$. For multiple particles, the linear momenta add up.

It may again be interesting to see how that works out for the two example systems introduced earlier. The first example was the electron in a hydrogen atom. In that example it is assumed that the proton is fixed at the origin. The energy eigenfunctions for the electron then were of the form

$$\psi_{nlm}(\vec{r})$$

with $\vec{r}$ the position of the electron. Shifting the coordinate system for this solution means replacing $\vec{r}$ by $\vec{r} + d\hat{k}$. That shifts the position of the electron without changing the position of the proton. The physics is not the same after such a shift. Correspondingly, the eigenfunctions do not change by a factor of the form $e^{\mathrm{i}k_z d}$ under the shift. Just looking at the ground state,

$$\psi_{100}(\vec{r}) = \frac{1}{\sqrt{\pi a_0^3}} e^{-|\vec{r}|/a_0}$$

is enough to see that. An electron around a stationary proton does not have definite linear momentum. In other words, the linear momentum of the electron is not conserved.

However, the physics of the complete hydrogen atom as described in addendum {A.5} is independent of coordinate shifts. A suitable choice of energy eigenfunctions in this context is

$$\psi_{nlm,\mathrm{red}}(\vec{r} - \vec{r}_\mathrm{p}) e^{\mathrm{i}\vec{k}\cdot\vec{r}_\mathrm{cg}}$$

where $\vec{k}$ is a constant wave number vector. The first factor does not change under coordinate shifts because the vector $\vec{r} - \vec{r}_\mathrm{p}$ from proton to electron does not. The exponential changes by the expected factor $e^{\mathrm{i}k_z d}$ because the position $\vec{r}_\mathrm{cg}$ of the center of gravity of the atom changes by an amount d in the z-direction.

The derivation of linear momentum can be extended to conduction electrons in crystalline solids. In that case, the physics of the conduction electrons is unchanged if the coordinate system is translated over a crystal period d. (This assumes that the z-axis is chosen along one of the primitive vectors of the crystal structure.) The eigenvalues are still of the form $e^{\mathrm{i}k_z d}$. However, unlike for linear momentum, the translation d must be the crystal period, or an integer multiple of it. Therefore, the operator $\widehat{p}_z$ is not useful; the symmetry does not continue to apply in the limit $d \to 0$.

The conserved quantity in this case is just the $e^{\mathrm{i}k_z d}$ eigenvalue of $\mathcal{T}_{z,d}$. It is not possible from that eigenvalue to uniquely determine a value of k_z and the corresponding crystal momentum $\hbar k_z$. Values of k_z that differ by a whole multiple of $2\pi/d$ produce the same eigenvalue. But Bloch waves have the same indeterminacy in their value of k_z anyway. In fact, Bloch waves are eigenfunctions of $\mathcal{T}_{z,d}$ as well as energy eigenfunctions.

One consequence of the indeterminacy in k_z is an increased number of possible electromagnetic transitions. Typical electromagnetic radiation has a wave length that is large compared to the atomic spacing. Essentially the electromagnetic field is the same from one atom to the next. That means that it has negligible crystal momentum, using the smallest of the possible values of k_x as measure. Therefore the radiation cannot change the conserved eigenvalue $e^{\mathrm{i}k_z d}$. But it can still produce electron transitions between two Bloch waves that have been assigned different k_z values in some "extended zone scheme," chapter 6.22.4. As long as the two k_z values differ by a whole multiple of $2\pi/d$, the actual eigenvalue $e^{\mathrm{i}k_z d}$ does not change. In that case there is no violation of the conservation law in the transition. The ambiguity in k_z values may be eliminated by switching to a "reduced zone scheme" description, chapter 6.22.4.

The time shift operator $\mathcal{U}_\tau$ shifts the time coordinate over an interval τ. In empty space, its eigenfunctions are exactly the energy eigenfunctions. Its eigenvalues are of the form

$$e^{-\mathrm{i}\omega\tau}$$

where classical physics defines $\hbar\omega$ as the energy E. The energy operator can be defined correspondingly, and is simply the Hamiltonian:

$$H = \mathrm{i}\hbar \lim_{\varepsilon \to 0} \frac{\mathcal{U}_\varepsilon - 1}{\varepsilon} = \mathrm{i}\hbar \frac{\partial}{\partial t} \tag{A.79}$$

In other words, we have reasoned in a circle and rederived the Schrödinger equation from time shift symmetry. But you could generalize the reasoning to the motion of particles in an external field that varies periodically in time.

Usually, nature is not just symmetric under rotating or translating it, but also under mirroring it. A transformation that creates a mirror image of a given system is called a parity transformation. The mathematically cleanest way to do it is to invert the direction of each of the three Cartesian axes. That is called spatial inversion. Physically it is equivalent to mirroring the system using some mirror passing through the origin, and then rotating the system 180° around the axis normal to the mirror.

(In a strictly two-dimensional system, spatial inversion does not work, since the rotation would take the system into the third dimension. In that case, mirroring can be achieved by replacing just x by $-x$ in some suitably chosen xy-coordinate system. Subsequently replacing y by $-y$ would amount to a second mirroring that would restore a nonmirror image. In those terms, in

three dimensions it is replacing z by $-z$ that produces the final mirror image in spatial inversion.)

The analysis of the conservation law corresponding to spatial inversion proceeds much like the one for angular momentum. One difference is that applying the spatial inversion operator a second time turns $-\vec{r}$ back into the original $\vec{r}$. Then the wave function is again the same. In other words, applying spatial inversion twice multiplies wave functions by 1. It follows that the square of every eigenvalue is 1. And if the square of an eigenvalues is 1, then the eigenvalue itself must be either 1 or -1. In the same notation as used for angular momentum, the eigenvalues of the spatial inversion operator can therefore be written as

$$e^{\mathrm{i}m'\pi} = (-1)^{m'} \tag{A.80}$$

where m' must be integer. However, it is pointless to give an actual value for m'; the only thing that makes a difference for the eigenvalue is whether m' is even or odd. Therefore, parity is simply called "odd" or "minus one" or "negative" if the eigenvalue is -1, and "even" or "one" or "positive" if the eigenvalue is 1.

In a system, the ± 1 parity eigenvalues of the individual particles multiply together. That is just like how the eigenvalues of the generator of rotation $\mathcal{R}_{z,\gamma}$ multiply together for angular momentum. Any particle with even parity has no effect on the system parity; it multiples the total eigenvalue by 1. On the other hand, each particle with odd parity flips over the total parity from odd to even or vice-versa; it multiplies the total eigenvalue by -1. Particles can also have intrinsic parity. However, there is no half-integer parity like there is half-integer spin.

A.19.5 A gauge symmetry and conservation of charge

Modern quantum theories are build upon so-called "gauge symmetries." This subsection gives a simple introduction to some of the ideas.

Consider classical electrostatics. The force on charged particles is the product of the charge of the particle times the so-called electric field $\vec{\mathcal{E}}$. Basic physics says that the electric field is minus the derivative of a potential φ. The potential φ is commonly known as the "voltage" in electrical applications. Now it too has a symmetry: adding some arbitrary constant, call it C, to φ does not make a difference. Only *differences* in voltage can be observed physically. That is a very simple example of a gauge symmetry, a symmetry in an unobservable field, here the potential φ.

Note that this symmetry does not involve the gauges used to measure voltages in any way. Instead it is a reference point symmetry; it does not make a difference what voltage you want to declare to be zero. It is conventional to take the earth as the reference voltage, but that is a completely arbitrary choice. So the term "gauge symmetry" is misleading, like many other terms in physics. A

symmetry in a unobservable quantity should of course simply have been called an unobservable symmetry.

There is a relationship between this gauge symmetry in φ and charge conservation. Suppose that, say, a few photons create an electron and an antineutrino. That can satisfy conservation of angular momentum and of lepton number, but it would violate charge conservation. Photons have no charge, and neither have neutrinos. So the negative charge $-e$ of the electron would appear out of nothing. But so what? Photons can create electron-positron pairs, so why not electron-antineutrino pairs?

The problem is that in electrostatics an electron has an electrostatic energy $-e\varphi$. Therefore the photons would need to provide not just the rest mass and kinetic energy for the electron and antineutrino, but also an additional electrostatic energy $-e\varphi$. That additional energy could be determined from comparing the energy of the photons against that of the electron-antineutrino pair. And that would mean that the value of φ at the point of pair creation has been determined. Not just a difference in φ values between different points. And that would mean that the value of the constant C would be fixed. So nature would not really have the gauge symmetry that a constant in the potential is arbitrary.

Conversely, if the gauge symmetry of the potential is fundamental to nature, creation of lone charges must be impossible. Each negatively charged electron that is created must be accompanied by a positively charged particle so that the net charge that is created is zero. In electron-positron pair creation, the positive charge $+e$ of the positron makes the net charge that is created zero. Similarly, in beta decay, an uncharged neutron creates an electron-antineutrino pair with charge $-e$, but also a proton with charge $+e$.

You might of course wonder whether an electrostatic energy contribution $-e\varphi$ is really needed to create an electron. It is because of energy conservation. Otherwise there would be a problem if an electron-antineutrino pair was created at a location P and disintegrated again at a different location Q. The electron would pick up a kinetic energy $-e(\varphi_{\rm P}-\varphi_{\rm Q})$ while traveling from P to Q. Without electrostatic contributions to the electron creation and annihilation energies, that kinetic energy would make the photons produced by the pair annihilation more energetic than those destroyed in the pair creation. So the complete process would create additional photon energy out of nothing.

The gauge symmetry takes on a much more profound meaning in quantum mechanics. One reason is that the Hamiltonian is based on the potential, not on the electric field itself. To appreciate the full impact, consider electrodynamics instead of just electrostatics. In electrodynamics, a charged particle does not just experience an electric field $\vec{\mathcal{E}}$ but also a magnetic field $\vec{\mathcal{B}}$. There is a corresponding additional so-called "vector potential" $\vec{A}$ in addition to the scalar potential φ. The relation between these potentials and the electric and magnetic

fields is given by, chapter 13.1:

$$\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial \vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla \times \vec{A}$$

Here ∇, nabla, is the differential operator of vector calculus (calculus III in the U.S. system):

$$\nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

The gauge property now becomes more general. The constant C that can be added to φ in electrostatics no longer needs to be constant. Instead, it can be taken to be the time-derivative of any arbitrary function $\chi(x, y, z, t)$. However, the gradient of this function must also be subtracted from $\vec{A}$. In particular, the potentials

$$\varphi' = \varphi + \frac{\partial \chi}{\partial t} \qquad \vec{A}' = \vec{A} - \nabla\chi$$

produce the exact same electric and magnetic fields as φ and $\vec{A}$. So they are physically equivalent. They produce the same observable motion.

However, the *wave function* computed using the potentials φ' and $\vec{A}'$ is different from the one computed using φ and $\vec{A}$. The reason is that the Hamiltonian uses the potentials rather than the electric and magnetic fields. Ignoring spin, the Hamiltonian of an electron in an electromagnetic field is, chapter 13.1:

$$H = \frac{1}{2m_{\rm e}}\left(\frac{\hbar}{\rm i}\nabla + e\vec{A}\right)^2 - e\varphi$$

It can be seen by crunching it out that if Ψ satisfies the Schrödinger equation in which the Hamiltonian is formed with φ and $\vec{A}$, then

$$\Psi' = e^{{\rm i}e\chi/\hbar}\Psi \tag{A.81}$$

satisfies the one in which H is formed with φ' and $\vec{A}'$.

To understand what a stunning result that is, recall the physical interpretation of the wave function. According to Born, the square magnitude of the wave function $|\Psi|^2$ determines the probability per unit volume of finding the electron at a given location. But the wave function is a complex number; it can always be written in the form

$$\Psi = e^{{\rm i}\alpha}|\Psi|$$

where α is a real quantity corresponding to a phase angle. This angle is not directly observable; it drops out of the magnitude of the wave function. And the gauge property above shows that not only is α not observable, it can be anything. For, the function χ can change α by a *completely arbitrary* amount $e\chi/\hbar$ and it remains a solution of the Schrödinger equation. The only variables that change are the equally unobservable potentials φ and $\vec{A}$.

As noted earlier, a symmetry means that you can do something and it does not make a difference. Since α can be chosen completely arbitrary, varying with both location and time, this is a very strong symmetry. Zee writes, (Quantum Field Theory in a Nutshell, 2003, p. 135): "The modern philosophy is to look at [the equations of quantum electrodynamics] as a result of [the gauge symmetry above]. If we want to construct a gauge-invariant relativistic field theory involving a spin $\frac{1}{2}$ and a spin 1 field, then we are forced to quantum electrodynamics."

Geometrically, a complex number like the wave function can be shown in a two-dimensional complex plane in which the real and imaginary parts of the number form the axes. Multiplying the number by a factor $e^{ie\chi/\hbar}$ corresponds to rotating it over an angle $e\chi/\hbar$ around the origin in that plane. In those terms, the wave function can be rotated over an arbitrary, varying, angle in the complex plane and it still satisfies the Schrödinger equation.

For a relatively accessible derivation how the gauge invariance produces quantum electrodynamics, see [24, pp. 358ff]. To make some sense out of it, chapter 1.2.5 gives a brief inroduction to relativistic index notation, chapter 12.12 to the Dirac equation and its matrices, addendum {A.1} to Lagrangians, and {A.21} to photon wave functions. The $F^{\mu\nu}$ are derivatives of this wave function, [24, p. 239].

A.19.6 Reservations about time shift symmetry

It is not quite obvious that the evolution of a physical system in empty space is the same regardless of the time that it is started. It is certainly not as obvious as the assumption that changes in spatial position do not make a difference. Cosmology does not show any evidence for a fundamental difference between different locations in space. For each spatial location, others just like it seem to exist elsewhere. But different cosmological times do show a major physical distinction. They differ in how much later they are than the time of the creation of the universe as we know it. The universe is expanding. Spatial distances between galaxies are increasing. It is believed with quite a lot of confidence that the universe started out extremely concentrated and hot at a "Big Bang" about 15 billion years ago.

Consider the cosmic background radiation. It has cooled down greatly since the universe became transparent to it. The expansion stretched the wave length of the photons of the radiation. That made them less energetic. You can look upon that as a violation of energy conservation due to the expansion of the universe.

Alternatively, you could explain the discrepancy away by assuming that the missing energy goes into potential energy of expansion of the universe. However, whether this "potential energy" is anything better than a different name for "energy that got lost" is another question. Potential energy is normally energy

that is lost but can be fully recovered. The potential energy of expansion of the universe cannot be recovered. At least not on a global scale. You cannot stop the expansion of the universe.

And a lack of exact energy conservation may not be such a bad thing for physical theories. Failure of energy conservation in the early universe could provide a possible way of explaining how the universe got all that energy in the first place.

In any case, for practical purposes nontrivial effects of time shifts seem to be negligible in the current universe. When astronomy looks at far-away clusters of galaxies, it sees them as they were billions of years ago. That is because the light that they emit takes billions of years to reach us. And while these galaxies look different from the current ones nearby, there is no evident difference in their basic laws of physics. Also, gravity is an extremely small effect in most other physics. And normal time variations are negligible compared to the age of the universe. Despite the Big Bang, conservation of energy remains one of the pillars on which physics is build.

A.20 Angular momentum of vector particles

This addendum is concerned with vector particles, particles whose wave functions are vectors. To be sure, the wave function of an electron can also be written as a vector, chapters 3.1 and 5.5.1:

$$\text{electron:} \quad \vec{\Psi}(\vec{r};t) \equiv \begin{pmatrix} \Psi_+(\vec{r};t) \\ \Psi_-(\vec{r};t) \end{pmatrix}$$

But that is not a normal vector. It is a two-dimensional vector in three-dimensional space, and is known as a spinor. This addendum is concerned with wave functions that are normal three-dimensional vectors. That is of importance for understanding, for example, the spin angular momentum of photons. A photon is a vector particle, though a special one. It will be shown in this addendum that the spin of a vector particle is 1. The parity of such a particle will also be discussed.

To really appreciate this addendum, you may want to read the previous addendum {A.19} first. In any case, according to that addendum angular momentum is related to what happens to the wave function under rotation of the coordinate system. In particular, the angular momentum in the z-direction is related to what happens if the coordinate system is rotated around the z-axis.

Consider first the simplest possible vector wave function:

$$\vec{\Psi} = \vec{A}' f(r) \qquad \vec{A}' = \begin{pmatrix} A'_x \\ A'_y \\ A'_z \end{pmatrix}$$

Here $\vec{A}'$ is a constant vector. Also r is the distance from the origin around which the angular momentum is defined. Finally, $f(r)$ is any arbitrary function of r. The big question is now what happens to this wave function if the coordinate system is rotated over some angle γ around the z-axis.

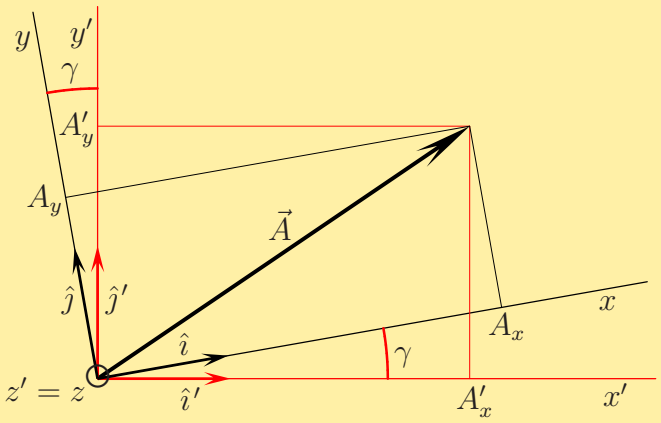


Figure A.8: Effect of rotation of the coordinate system on a vector. The vector is physically the same, but it has a different mathematical representation, different components, in the two coordinate systems.

The factor $f(r)$ does not change under rotations because the distance from the origin does not. But the vector $\vec{A}'$ *does* change. Figure A.8 shows what happens. The vector in the rotated coordinate system x, y, z has components

$$A_x = \cos(\gamma)A'_x + \sin(\gamma)A'_y \qquad A_y = -\sin(\gamma)A'_x + \cos(\gamma)A'_y \qquad A_z = A'_z \quad \text{(A.82)}$$

For example, the first relation expresses that $A'_x\hat{\imath}'$ has a component $\cos(\gamma)A'_x$ in the direction of $\hat{\imath}$, while $A'_y\hat{\jmath}'$ has a component $\sin(\gamma)A'_y$ in that direction. The second expression follows similarly. The z-component does not change under the rotation.

Now consider three very special vectors:

$$\boxed{\vec{Y}_1^1 = \begin{pmatrix} 1/\sqrt{2} \\ \mathrm{i}/\sqrt{2} \\ 0 \end{pmatrix} \qquad \vec{Y}_1^0 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \qquad \vec{Y}_1^{-1} = \begin{pmatrix} -1/\sqrt{2} \\ \mathrm{i}/\sqrt{2} \\ 0 \end{pmatrix}} \qquad \text{(A.83)}$$

If you plug $\vec{A}' = \vec{Y}_1^1$ into the relations (A.82) given above and use the Euler identity (2.5), you get $\vec{A} = e^{\mathrm{i}\gamma}\vec{Y}_1^1$. So the vector $\vec{Y}_1^1$ changes by a mere scalar factor, the exponential $e^{\mathrm{i}\gamma}$, under the rotation of the coordinate system. According to the relationship between rotations and angular momentum, {A.19}, that makes $\vec{Y}_1^1$ a state of definite angular momentum in the z-direction. Also, the magnetic quantum number m_s of the momentum in the z-direction is by definition the coefficient of $\mathrm{i}\gamma$ in the exponential factor $e^{\mathrm{i}\gamma} = e^{1\mathrm{i}\gamma}$. Here that

is 1, so $m_s = 1$. This value is shown as the superscript of the state $\vec{Y}_1^1$. The actual angular momentum in the z-direction is $m_s\hbar$, so it is $\hbar$ for this state. This angular momentum should be called spin, in analogy to the case for the electron. It is due to the fact that the wave function is a vector.

The vector $\vec{Y}_1^0$ has only a z-component, so it does not change under the rotation. Phrased differently, it changes by a unit factor $e^{0\mathrm{i}\gamma}$. That makes its magnetic quantum number m_s zero, as the superscript in $\vec{Y}_1^0$ says. Then the angular momentum in the z-direction of this state is zero too. Finally, the vector $\vec{Y}_1^{-1}$ changes by a factor $e^{-1\mathrm{i}\gamma}$ under rotation, so it has $m_s = -1$, as its superscript says, and the angular momentum in the z-direction is $-\hbar$.

To get at square angular momentum, first the operator $\widehat{S}_z$ of spin angular momentum around the z-axis is needed. The relation between angular momentum and rotations shows that this operator takes the general form, {A.19} (A.76),

$$\widehat{S}_z = \frac{\hbar}{\mathrm{i}} \lim_{\gamma\to 0} \frac{\mathcal{R}_{z,\gamma} - 1}{\gamma}$$

Here $\mathcal{R}_{z,\gamma}$ is the operator that describes the effect of the rotation of the coordinate system on the wave function. Applied on the vector $\vec{A}'$ in the unrotated coordinate system, that means

$$\widehat{S}_z \vec{A}' = \frac{\hbar}{\mathrm{i}} \lim_{\gamma\to 0} \frac{\vec{A} - \vec{A}'}{\gamma}$$

Plugging in the components of $\vec{A}$ as given earlier, (A.82), and taking the limits using l'Hôpital, that produces

$$\widehat{S}_z \begin{pmatrix} A'_x \\ A'_y \\ A'_z \end{pmatrix} = \frac{\hbar}{\mathrm{i}} \begin{pmatrix} A'_y \\ -A'_x \\ 0 \end{pmatrix}$$

So the operator $\widehat{S}_z$ drops the z-component and swaps the other two components, changing the sign of the first, and then adds a factor $\hbar/\mathrm{i}$. If the same operations are performed another time, the net result is:

$$\widehat{S}_z^2 \begin{pmatrix} A'_x \\ A'_y \\ A'_z \end{pmatrix} = \hbar^2 \begin{pmatrix} A'_x \\ A'_y \\ 0 \end{pmatrix}$$

So the square operator just drops the z-component and adds a factor $\hbar^2$.

Of course, the operators $\widehat{S}_x^2$ and $\widehat{S}_y^2$ are defined similarly. There is nothing special about the z-axis. The operator of square spin angular momentum is defined as

$$\widehat{S}^2 \equiv \widehat{S}_x^2 + \widehat{S}_y^2 + \widehat{S}_z^2$$

Since each of the operators in the right hand side drops a different component and adds a factor $\hbar^2$ to the other two, the total for any vector $\vec{A}'$ is,

$$\widehat{S}^2\vec{A}' = 2\hbar^2\vec{A}'$$

So the square spin operator always produces a simple multiple of the original vector. That makes *any* vector an eigenvector of square spin angular momentum. Also, the azimuthal quantum number s, the spin, can by definition be found from equating the coefficient $2\hbar^2$ of $\vec{A}'$ in the right hand side above to $s(s+1)\hbar^2$. The only nonnegative value s that can satisfy this condition is $s = 1$.

That then means that the spin s of vector particles is equal to 1. So a vector particle is a boson of spin 1. The subscript on the special vectors $\vec{Y}_1^{m_s}$ indicates their spin $s = 1$.

You can write the most general vector wave function in the form

$$\vec{\Psi}(\vec{r};t) = \Psi_1(\vec{r};t)\vec{Y}_1^1 + \Psi_0(\vec{r};t)\vec{Y}_1^1 + \Psi_{-1}(\vec{r};t)\vec{Y}_1^{-1}$$

Then you can put the coefficients in a vector much like the wave function of the electron, but now three-dimensional:

$$\text{vector boson:} \quad \vec{\Psi}(\vec{r};t) \equiv \begin{pmatrix} \Psi_1(\vec{r};t) \\ \Psi_0(\vec{r};t) \\ \Psi_{-1}(\vec{r};t) \end{pmatrix}$$

Like the electron, the vector particle can of course also have orbital angular momentum. That is due to the coefficients Ψ_{m_s} in the wave function above. So far it has been assumed that these coefficients only depended on the distance r from the origin. However, consider the following more general component of a vector wave function:

$$\vec{Y}_1^{m_s} f(r) Y_l^{m_l}(\theta,\phi) \tag{A.84}$$

Here θ and ϕ are the position angles in spherical coordinates, and $Y_l^{m_l}$ is a so-called spherical harmonic of orbital angular momentum, chapter 4.2.3. For the above wave function to be properly normalized,

$$\int_{r=0}^{\infty} r^2 f(r)\,\mathrm{d}r = 1$$

(To check this, take a dot, or rather inner, product of the wave function with itself. Then integrate over all space using spherical coordinates and the orthonormality of the spherical harmonics.)

The wave function (A.84) above has orbital angular momentum in the z-direction equal to $m_l\hbar$ in addition to the spin angular momentum $m_s\hbar$. So the total angular momentum in the z-direction is $(m_s + m_l)\hbar$. To check that, note that under rotations of the coordinate system, the vector changes by a factor

$e^{m_s \mathrm{i}\gamma}$ while the spherical harmonic changes by an additional factor $e^{m_l \mathrm{i}\gamma}$. That makes the total change a factor $e^{(m_s+m_l)\mathrm{i}\gamma}$.

In general, then, the magnetic quantum number m_j of the net angular momentum is simply the sum of the spin and orbital ones,

$$\boxed{m_j = m_s + m_l} \tag{A.85}$$

However, the situation for the azimuthal quantum number j of the net angular momentum is not so simple. In general the wave function (A.84) above will have uncertainty in the value of j. Combinations of wave functions of the form (A.84) are usually needed to get states of definite j.

That is a complicated issue best left to chapter 12. But a couple of special cases are worth mentioning already. First, if $m_s = 1$ and $m_l = l$, or alternatively, if $m_s = -1$ and $m_l = -l$, then j is simply the sum of the spin and orbital azimuthal quantum numbers $1 + l$.

The other special case is that there is zero net angular momentum. Zero net angular momentum means that the wave function is exactly the same regardless how the coordinate system is rotated. And that only happens for a vector wave function if it is purely radial:

$$\hat{\imath}_r f(r) \frac{1}{\sqrt{4\pi}}$$

Here $\hat{\imath}_r$ is the unit vector sticking radially outward away from the origin. The final constant is the spherical harmonic Y_0^0. It is needed the satisfy the normalization requirement unless you change the one on f.

The above state has zero net angular momentum. The question of interest is what can be said about its spin and orbital angular momentum. To answer that, it must be rewritten in terms of Cartesian components. Now the unit vector $\hat{\imath}_r$ has Cartesian components

$$\hat{\imath}_r = \frac{x}{r}\hat{\imath} + \frac{y}{r}\hat{\jmath} + \frac{z}{r}\hat{k}$$

The spatial factors in this expression can be written in terms of the spherical harmonics $Y_1^{m_l}$, chapter 4.2.3. That gives the state of zero net angular momentum as

$$\hat{\imath}_r f(r) Y_0^0 = \left(\sqrt{1\!/\!_3}\vec{Y}_1^1 Y_1^{-1} - \sqrt{1\!/\!_3}\vec{Y}_1^0 Y_1^0 + \sqrt{1\!/\!_3}\vec{Y}_1^{-1} Y_1^1 \right) f(r)$$

To check this, just plug in the expressions for the $\vec{Y}_1^{m_s}$ of (A.83), and for the $Y_1^{m_l}$ of table 4.3.

The bottom line is that by combining states of unit spin $s = 1$, and unit orbital angular momentum $l = 1$, you can create a state of zero net angular momentum, $j = 0$. Note also that in each of the three terms in the right hand side above, m_s and m_l add up to zero. A state of zero angular momentum $j =$

0 must have $m_j = 0$ without uncertainty. Further note that the values of both the spin and orbital angular momentum in the z-direction are uncertain. Each of the two has measurable values $\hbar$, 0, or $-\hbar$ with equal probability $\frac{1}{3}$.

The above relation may be written more neatly in terms of "ket notation." In ket notation, an angular momentum state with azimuthal quantum number j and magnetic quantum number m_j is indicated as $|j\ m_j\rangle$. Using this notation, and dropping the common factor $f(r)$, the above relation can be written as

$$|0\ 0\rangle_j = \sqrt{1/3}|1\ 1\rangle_s|1\ {-1}\rangle_l - \sqrt{1/3}|1\ 0\rangle_s|1\ 0\rangle_l + \sqrt{1/3}|1\ {-1}\rangle_s|1\ 1\rangle_l$$

Here the subscripts j, s, and l indicate net, spin, and orbital angular momentum, respectively.

There is a quicker way to get this result than going through the above algebraic mess. You can simply read off the coefficients in the appropriate column of the bottom-right tabulation in figure 12.6. (In this figure take a to stand for spin, b for orbital, and ab for net angular momentum.) Figure 12.6 also has the coefficients for many other net spin states that you might need. A derivation of the figure must wait until chapter 12.

The parity of vector wave functions is also important. Parity is what happens to a wave function if you invert the positive direction of all three Cartesian axes. What happens to a vector wave function under such an inversion can vary. A normal, or "polar," vector changes sign when you invert the axes. For example, a position vector $\vec{r}$ in classical physics is a polar vector. Each position coordinate x, y, and z changes sign, and therefore so does the entire vector. Similarly, a velocity vector $\vec{v}$ is a polar vector; it is just the time derivative of position. A particle with a vector wave function that behaves like a normal vector has negative intrinsic parity. The sign of the wave function flips over under axes inversion. Particles of this type turn out to include the photon.

But now consider an example like a classical angular momentum vector, $\vec{r} \times m\vec{v}$. Since both the position and the velocity change sign under spatial inversion, a classical angular momentum vector stays the same. A vector that does not change under axes inversion is called a "pseudovector" or "axial" vector. A particle whose wave function behaves like a pseudovector has positive intrinsic parity.

Note however that the orbital angular momentum of the particle also has an effect on the net parity. In particular, if the quantum number of orbital angular momentum l is odd, then the net parity is the opposite of the intrinsic one. If the quantum number l is even, then the net parity is the intrinsic one. The reason is that spherical harmonics change sign under spatial inversion if l is odd, but not when l is even, {D.14}.

Particles of all types often have definite parity. Such a particle may still have uncertainty in l. But if parity is definite, the measurable values of l will need to be all even or all odd.

A.21 Photon type 2 wave function

In quantum mechanics, photons are the particles of the electromagnetic field. To actually use photons, something like a wave function for them is needed. But that is not quite trivial for a purely relativistic particle with zero rest mass like the photon. That is the primary topic of this addendum. It will be assumed throughout that the photon is in empty space.

A.21.1 The wave function

To see the problem with a photon wave function, a review of the wave function of the nonrelativistic electron is useful, chapters 3.1 and 5.5.1. The electron wave function can be written as a vector with two components:

$$\text{electron:} \quad \vec{\Psi}(\vec{r};t) \equiv \begin{pmatrix} \Psi_+(\vec{r};t) \\ \Psi_-(\vec{r};t) \end{pmatrix}$$

This wave function takes on two different meanings

1. It gives the probability per unit volume of finding the electron at a given position with a given spin. For example, $|\Psi_+(\vec{r};t)|^2\,\mathrm{d}^3\vec{r}$ gives the probability of finding the electron with spin-up in an vicinity of infinitesimal volume $\mathrm{d}^3\vec{r}$ around position $\vec{r}$. That is the Born statistical interpretation.
2. It is the unobservable function that nature seems to use to do its quantum "computations" of how physics behaves.

Now a wave function of type 1 is not really meaningful for a photon. What would it mean, find a photon? Since the photon has no rest mass, you cannot bring them to a halt: there would be nothing left. And anything you do to try to localize the electromagnetic field is likely to just produce new photons. (To be sure, with some effort something can be done towards a meaningful wave function of type 1, e.g. [Sype, J.E. 1995 Phys. Rev. A 52, 1875]. It would have two components like the electron, since the photon has two independent spin states. But wave functions of that type are not widely accepted, nor useful for the purposes here.)

So what? A wave function of type 1 is not that great anyway. For one, it only defines the magnitudes of the components of the wave function. If you only define the magnitude of a complex function, you define only half of it. True, even as a type 2 wave function the classical electron wave function is not quite unique. You can still multiply either component by a factor $e^{\mathrm{i}\alpha}$, with α a real constant, without changing any of the physics. But that is not by far as bad as completely ignoring everything else besides the magnitude.

Furthermore, relativistic quantum mechanics has discovered that what we call an electron is something cloaked in a cloud of virtual particles. It is anybody's guess what is inside that cloak, but it will not be anything resembling

what we would call an electron. So what does it really mean, finding an electron within an infinitesimal volume around a point? What happens to that cloak? And to really locate an electron in an infinitesimal volume requires infinite energy. If you try to locate the electron in a region that is small enough, you are likely to just create additional electron-positron pairs much like for photons.

For most practical purposes, classical physics understands the particle behavior of electrons very well, but not their wave behavior. Conversely, it understands the wave behavior of photons very well, but not their particle behavior. But when you go to high enough energies, that distinction becomes much less obvious.

The photon most definitely has a wave function of type 2 above. In quantum electrodynamics, it may simply be called the photon wave function, [24, p. 240]. However, since the term already seems to be used for type 1 wave functions, this book will use the term "photon type 2 wave function." It may not tell you where to find that elusive photon, but you will definitely need it to figure out how that photon interacts with, say, an electron.

What the type 2 wave function of the photon is can be guessed readily from classical electromagnetics. After all, the photon is supposed to be the particle of the electromagnetic field. So, consider first electrostatics. In classical electrostatics the forces on charged particles are described by an electric force per unit charge $\vec{\mathcal{E}}$. That is called the electric field.

But quantum mechanics uses potentials, not forces. For example, the solution of the hydrogen atom of chapter 4.3 used a potential energy of the electron V. In electrostatics, this potential energy is written as $V = -e\varphi$ where $-e$ is the charge of the electron and φ is called the electrostatic potential. This potential is not directly observable nor unique; you can add any constant to it without changing the observed physics.

Clearly, an unobservable function φ for the electromagnetic field sounds much like a wave function for the particle of that field, the photon. But actually, the electrostatic potential φ is only part of it. In classical electromagnetics, there is not just an electric field $\vec{\mathcal{E}}$, there is also a magnetic field $\vec{\mathcal{B}}$. It is known that this magnetic field can be represented by a so-called "vector potential" $\vec{A}$.

The following relationships give the electric and magnetic fields in terms of these potentials:

$$\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial \vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla \times \vec{A} \tag{A.86}$$

Here the operator

$$\nabla = \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

is called nabla or del. As an example, for the z components of the fields:

$$\mathcal{E}_z = -\frac{\partial \varphi}{\partial z} - \frac{\partial A_z}{\partial t} \qquad \mathcal{B}_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}$$

When both potentials are allowed for, the nonuniqueness becomes much larger. In particular, for any arbitrary function χ of position and time, you can find two different potentials φ' and $\vec{A}'$ that produce the exact same electric and magnetic fields as φ and $\vec{A}$. These potentials are given by

$$\varphi' = \varphi - \frac{\partial \chi}{\partial t} \qquad \vec{A}' = \vec{A} + \nabla \chi \tag{A.87}$$

This indeterminacy in potentials is the famous "gauge property" of the electromagnetic field.

Finally, it turns out that classical relativistic mechanics likes to combine the four scalar potentials in a four-dimensional vector, or four-vector, chapter 1.3.2:

$$\overrightarrow{A} = \begin{pmatrix} \varphi/c \\ \vec{A} \end{pmatrix} = \begin{pmatrix} \varphi/c \\ A_x \\ A_y \\ A_z \end{pmatrix}$$

That is the one. Quantum mechanics takes a four-vector potential of this form to be the type 2 wave function of the photon $\overrightarrow{A}_\gamma$. It keeps the gauge property (A.87) for this wave function. However, note the following important caveat:

> *The photon wave function $\overrightarrow{A}_\gamma$ should not be confused with the classical four-potential $\overrightarrow{A}$.*

Wave functions are in general complex. The classical four-potential, and especially its physically observable derivatives, the electric and magnetic fields, must be real. Indeed, according to quantum mechanics, observable quantities correspond to eigenvalues of Hermitian operators, not to wave functions. What the operators of the observable electric and magnetic fields are will be discussed in addendum {A.23}.

A.21.2 Simplifying the wave function

To use the photon wave function in practical applications, it is essential to simplify it. That can be done by choosing a clever gauge function χ in the gauge property (A.87).

One very helpful simplification is to choose χ so that

$$\frac{1}{c}\frac{\partial \varphi_\gamma/c}{\partial t} + \nabla \cdot \vec{A}_\gamma = 0 \tag{A.88}$$

where c is the speed of light. This is called the "Lorenz condition." A corresponding gauge function is a "Lorenz gauge." The reason why the Lorenz condition is a good one is because all observers in inertial motion will agree

it is true. (You can crunch that out using the Lorentz transform as given in chapter 1.2.1 (1.6). The four-vector $\overrightarrow{A}_\gamma$ transforms the same way as the four-vector $\overrightarrow{r}$. However, you will need to use the inverse transform for one of the two four-vectors. Alternatively, those familiar with index notation as briefly described in chapter 1.2.5 recognize the Lorenz condition as being simply $\partial_\mu A^\mu_\gamma = 0$. That is unchanged going from one observer to the next, because the upper index transforms under the Lorentz transform and the lower index under the inverse Lorentz transform.)

To achieve the Lorenz condition, assume an initial wave function $(\varphi'_\gamma, \vec{A}'_\gamma)$ that does not satisfy it. Then plug the gauge property (A.87) into the Lorenz condition above. That shows that the needed gauge function χ must satisfy

$$-\frac{1}{c^2}\frac{\partial^2 \chi}{\partial t^2} + \nabla^2 \chi = \frac{1}{c}\frac{\partial \varphi'_\gamma/c}{\partial t} + \nabla \cdot \vec{A}'_\gamma$$

This equation for χ is called an inhomogeneous Klein-Gordon equation. (More generically, it is called an inhomogeneous wave equation.)

There is another reason why you want to satisfy the Lorenz condition. The photon is a purely relativistic particle with zero rest mass. Then following the usual ideas of quantum mechanics, in empty space its wave function should satisfy the homogeneous Klein-Gordon equation, {A.14} (A.43):

$$\boxed{-\frac{1}{c^2}\frac{\partial^2 \overrightarrow{A}_\gamma}{\partial t^2} + \nabla^2 \overrightarrow{A}_\gamma = 0} \tag{A.89}$$

Unfortunately, that is not automatic. In general, gauge transforms mess up this equation. However, as long as gauge transforms respect the Lorenz condition, they also respect the Klein-Gordon equation. So reasonably speaking, "normal" photon wave functions, the ones that do satisfy the Klein-Gordon equation, should be exactly the ones that also satisfy the Lorenz condition.

Maxwell's classical electromagnetics provides additional support for that idea. There the Klein-Gordon equation for the potentials also requires that the Lorenz condition is satisfied, {A.37}.

Since the inhomogeneous Klein-Gordon equation for the gauge function χ is second order in time, it still leaves two initial conditions to be chosen. These can be chosen such as to make the initial values for φ_γ and its time-derivative zero. That then makes φ_γ completely zero, because it satisfies the homogeneous Klein-Gordon equation.

And so the fully simplified photon wave function becomes:

$$\boxed{\text{Coulomb-Lorenz gauge:} \qquad \overrightarrow{A}_\gamma = \begin{pmatrix} 0 \\ \vec{A}_\gamma \end{pmatrix} \qquad \nabla \cdot \vec{A}_\gamma = 0} \tag{A.90}$$

The final condition applies because of the Lorenz condition (A.88). Using an expensive word, the final condition says that $\vec{A}_\gamma$ must be "solenoidal." A gauge function that makes $\vec{A}_\gamma$ solenoidal is called a "Coulomb gauge."

It should be noted that the Coulomb gauge is not Lorentz invariant. A moving observer will not agree that the potential φ_γ is zero and that $\vec{A}_\gamma$ is solenoidal. In real life that means that if you want to study a process in say a center-of-mass system, first switch to that system and then assume the Coulomb gauge. Not the other way around. The Coulomb-Lorenz gauge is too helpful not to use, although that is possible, [24, p. 241].

A.21.3 Photon spin

Now that the photon wave function has been simplified the photon spin can be determined. Recall that for the electron, the two components of the wave function correspond to its two possible values of the spin angular momentum $\widehat{S}_z$ in the chosen z-direction. In particular, Ψ_+ corresponds to $\widehat{S}_z = \frac{1}{2}\hbar$, and Ψ_- to $\widehat{S}_z = -\frac{1}{2}\hbar$. Since the wave function of the photon is a four-dimensional vector, at first it might therefore look like the photon should have spin $^3\!/_2$. That would make $\widehat{S}_z$ one of $\frac{3}{2}\hbar$, $\frac{1}{2}\hbar$, $-\frac{1}{2}\hbar$, or $-\frac{3}{2}\hbar$. But that is not true.

The simplified wave function (A.90) has only three nontrivial components. And the gauge property requires that this simplified wave function still describes all the physics. Since the only nontrivial part left is the three-dimensional vector $\vec{A}_\gamma$, the spin of the photon must be 1. The possible values of the spin in the z-direction $\widehat{S}_z$ are $\hbar$, 0, and $-\hbar$. The photon is a vector boson like discussed in addendum {A.20}.

However, that is not quite the end of the story. There is still that additional condition $\nabla \cdot \vec{A}_\gamma = 0$ to satisfy. In principle this constraint allows another component of the wave function to be eliminated. However, all three remaining components are spatial ones. So it does not make much sense to eliminate one and not the other. More importantly, it is known from relativity that $\vec{A}$ behaves like a normal three-dimensional vector under rotations of the coordinate system, not like a two-dimensional spinor like the electron wave function. That is implicit in the fact that the complete four-vector transforms according to the Lorentz transform, chapter 1.3.2. The spin is really 1.

Still, the additional constraint does limit the angular momentum of the photon. In particular, a photon does not have independent spin and orbital angular momentum. The two are intrinsically linked. What that means for the net angular momentum of photons is worked out in subsection A.21.7.

For now it may already be noted that the photon has no state of zero net angular momentum. A state of zero angular momentum needs to look the same from all directions. That is a consequence of the relationship between angular momentum and symmetry, chapter 7.3. Now the only vector wave functions that look the same from all directions are of the form $\hat{\imath}_r f(r)$. Here r is the distance from the origin around which the angular momentum is measured and $\hat{\imath}_r$ the unit vector pointing away from the origin. Such a wave function cannot

satisfy the condition $\nabla \cdot \hat{\imath}_r f(r) = 0$. That follows from applying the divergence theorem for a sphere around the origin.

A.21.4 Energy eigenstates

Following the rules of quantum mechanics, {A.14}, photon states of definite energy E take the form

$$\vec{A}_\gamma = c_0 \vec{A}^{\rm e}_\gamma e^{-{\rm i}Et/\hbar}$$

Here c_0 is an arbitrary constant. More importantly $\vec{A}^{\rm e}_\gamma$ is the energy eigenfunction, which is independent of time.

Substitution in the Klein-Gordon equation and cleaning up shows that this eigenfunction needs to satisfy the eigenvalue problem, {A.14},

$$\boxed{-\nabla^2 \vec{A}^{\rm e}_\gamma = k^2 \vec{A}^{\rm e}_\gamma \qquad k \equiv \frac{E}{\hbar c} = \frac{p}{\hbar} \qquad E = \hbar\omega \quad p = \hbar k} \tag{A.91}$$

Here p is the magnitude of the linear momentum of the photon. The so-called Planck-Einstein relation gives the energy E in terms of the photon frequency ω, while the de Broglie relation gives the momentum p in terms of the photon wave number k.

A.21.5 Normalization of the wave function

A classical wave function for a particle is normalized by demanding that the square integral of the wave function is 1. That does not work for a relativistic particle like the photon, since the Klein-Gordon equation does not preserve the square integral of the wave function, {A.14}.

However, the Klein-Gordon equation does preserve the following integral, {D.36.1},

$$\int_{\rm all} \left(\left| \frac{\partial \vec{A}_\gamma}{\partial t} \right|^2 + c^2 \left| \nabla \vec{A}_\gamma \right|^2 \right) {\rm d}^3\vec{r} = \text{the same for all time}$$

Reasonably speaking, you would expect this integral to be related to the energy in the electromagnetic field. After all, what other scalar physical quantity is there to be preserved?

Consider for a second the case that $\vec{A}_\gamma$ was a classical potential $\vec{A}$ instead of a photon wave function. Then the above integral can be rewritten in terms of the electric and magnetic fields $\vec{\mathcal{E}}$ and $\vec{\mathcal{B}}$ as, {D.36.1},

$$\int_{\rm all} \left(\left| \vec{\mathcal{E}} \right|^2 + c^2 \left| \vec{\mathcal{B}} \right|^2 \right) {\rm d}^3\vec{r} = \text{the same for all time}$$

Now classical physics does not have photons of energy $\hbar\omega$. All it has are electric and magnetic fields. Then surely the integral above must be a measure for the energy in the electromagnetic field? What is more logical than that the energy in the electromagnetic field per unit volume would be given by the square magnitudes of the electric and magnetic fields? No fields, no energy.

Of course, there needs to be an additional constant; the integral above does not have units of energy. If you check, you find that the "permittivity of space" $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m has the right units to be the constant. Actually, it turns out that the correct constant is $\frac{1}{2}\epsilon_0$. But that is not a fundamental issue; classical physics could just as well have defined ϵ_0 as half of what it did.

Now the photon wave function is not physically observable and does not have to conform to the rules of classical physics. But if you have to choose a normalization constant anway? Why not choose it so that what classical physics would take to be the energy is in fact the correct energy $\hbar\omega$? It is likely to simplify your life a lot.

So, the photon wave function normalization that will be used in this book is:

$$\boxed{\tfrac{1}{2}\epsilon_0 \int_{\text{all}} \left(\left| \vec{\mathcal{E}}^{\,\text{n}}_\gamma \right|^2 + c^2 \left| \vec{\mathcal{B}}^{\,\text{n}}_\gamma \right|^2 \right) \mathrm{d}^3\vec{r} = \hbar\omega} \tag{A.92}$$

Here $\vec{\mathcal{E}}^{\,\text{n}}_\gamma$ and $\vec{\mathcal{B}}^{\,\text{n}}_\gamma$ are what classical physics would take to be the electric and magnetic fields for the normalized photon energy eigenfunction $\vec{A}^{\,\text{n}}_\gamma$. Specifically,

$$\vec{\mathcal{E}}^{\,\text{n}}_\gamma = \mathrm{i}kc\vec{A}^{\,\text{n}}_\gamma \qquad \vec{\mathcal{B}}^{\,\text{n}}_\gamma = \nabla \times \vec{A}^{\,\text{n}}_\gamma$$

(To be sure, classical physics would take $\vec{\mathcal{E}}$ to be minus the time derivative of the potential $\vec{A}$. But for an energy eigenstate, the time derivative gives a simple factor $-\mathrm{i}\omega = -\mathrm{i}kc$.) The functions $\vec{\mathcal{E}}^{\,\text{n}}_\gamma$ and $\vec{\mathcal{B}}^{\,\text{n}}_\gamma$ will be referred to as "unobservable fields" to avoid confusion with the observable electric and magnetic fields.

Assume that you start with an unnormalized energy eigenfunction $\vec{A}^{\,\text{e}}_\gamma$. Then the normalized functions are usually most conveniently written as

$$\boxed{\vec{A}^{\,\text{n}}_\gamma = \frac{\varepsilon_k}{\mathrm{i}kc}\vec{A}^{\,\text{e}}_\gamma \qquad \vec{\mathcal{E}}^{\,\text{n}}_\gamma = \varepsilon_k \vec{A}^{\,\text{e}}_\gamma \qquad c\vec{\mathcal{B}}^{\,\text{n}}_\gamma = \frac{\varepsilon_k}{\mathrm{i}k}\nabla \times \vec{A}^{\,\text{e}}_\gamma} \tag{A.93}$$

Here the constant ε_k is to be found by substitution into the normalization condition (A.92).

A.21.6 States of definite linear momentum

The simplest quantum states for photons are states of definite linear momentum $\vec{p}$. And to make it even simpler, it will be assumed that the z-axis is chosen in the direction of the linear momentum.

In that case, the photon wave function takes the form

$$\vec{A}^{\rm e}_\gamma = \vec{A}^0 e^{{\rm i}kz} \qquad \vec{p} = \hat{k}\,\hbar k$$

Here $\vec{A}^0$ is a constant vector. That this wave function has definite linear momentum $\vec{p}$ may be verified by applying the linear momentum operator $\widehat{\vec{p}} = \hbar\nabla/{\rm i}$ on it. And substitution into the eigenvalue problem (A.91) verifies that it is an energy eigenfunction.

The vector $\vec{A}^0$ is not completely arbitrary; its z-component must be zero. That is in order that $\nabla \cdot \vec{A}^{\rm e}_\gamma$ is zero as the Coulomb-Lorenz gauge requires. So the wave function can be written as

$$\vec{A}^{\rm e}_\gamma = A^0_x\,\hat{\imath} e^{{\rm i}kz} + A^0_y\,\hat{\jmath} e^{{\rm i}kz}$$

The bottom line is that there are only two independent states, even though the wave function is a three-dimensional vector. The wave function cannot have a component in the direction of motion. It may be noted that the first term in the right hand side above is called a wave that is "linearly polarized" in the x-direction. Similarly, the second term is a wave that is linearly polarized in the y-direction. There is no longitudinal polarization of photons possible.

There is another useful way to write the wave function:

$$\vec{A}^{\rm e}_\gamma = c_1(\hat{\imath} + {\rm i}\hat{\jmath})e^{{\rm i}kz} + c_2(-\hat{\imath} + {\rm i}\hat{\jmath})e^{{\rm i}kz}$$

where c_1 and c_2 are constants. The first term in this expression is called "right-circularly polarized." It has angular momentum $\hbar$ in the z-direction. (To see why is a matter of rotating the coordinate system around the z-axis, {A.20}. The exponential does not change in such a rotation.) Similarly, the second state has angular momentum $-\hbar$ in the z-direction and is called left-circularly polarized. There is no state with angular momentum zero in the z-direction. In fact, it is exactly the missing z-component of $\vec{A}^0$ that would provide such a state, {A.20}.

There are still only two independent states. But another way of thinking about that is that the spin angular momentum in the direction of motion cannot be zero. The relative spin in the direction of motion, m_s/s is called the "helicity." It turns out that for a particle with zero rest mass like the photon, the helicity can only be 1 (right handed) or -1 (left handed), [24, p. 65].

Note further that the angular momenta in the x and y directions are uncertain. It so happens that the angular momentum in the direction of motion commutes with all three components of linear momentum, chapter 4.5.4. So it can have definite values. But the x and y angular momenta do not commute.

For later use, it is necessary to normalize the wave function using the procedure described in the previous subsection. To do so, it must be assumed that the photon is in a periodic box of volume $\mathcal{V}$, like in chapter 6.17. In infinite

space the wave function cannot be normalized, as it does not become zero at infinity. For the right-circularly polarized wave function as given above,

$$\boxed{\vec{A}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}kc}\frac{\hat{\imath}+{\rm i}\hat{\jmath}}{\sqrt{2}}e^{{\rm i}kz} \quad \vec{\mathcal{E}}^{\rm n}_\gamma = \varepsilon_k\frac{\hat{\imath}+{\rm i}\hat{\jmath}}{\sqrt{2}}e^{{\rm i}kz} \quad c\vec{\mathcal{B}}^{\rm n}_\gamma = \varepsilon_k\frac{-{\rm i}\hat{\imath}+\hat{\jmath}}{\sqrt{2}}e^{{\rm i}kz} \quad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}}} \tag{A.94}$$

In order to compare to the classical electromagnetic wave in chapter 7.7.1, another example is needed. This photon wave function has its linear momentum in the y-direction, and it is linearly polarized in the z-direction. Then an unnormalized energy eigenfunction is

$$\vec{A}^{\rm e}_\gamma = \hat{k}e^{{\rm i}ky}$$

The normalized eigenfunction and unobservable fields are in that case

$$\boxed{\vec{A}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}kc}\hat{k}e^{{\rm i}ky} \qquad \vec{\mathcal{E}}^{\rm n}_\gamma = \varepsilon_k\hat{k}e^{{\rm i}ky} \qquad c\vec{\mathcal{B}}^{\rm n}_\gamma = \varepsilon_k\hat{\imath}e^{{\rm i}ky} \qquad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}}} \tag{A.95}$$

Note that $\vec{\mathcal{E}}^{\rm n}_\gamma$, $\vec{\mathcal{B}}^{\rm n}_\gamma$, and the linear momentum are all orthogonal. That will reflect in the observable fields associated with the photon state. For the circularly polarized state, the electric and magnetic fields are not orthogonal. However, the observable fields will be.

For a general direction of the wave motion and its linear polarization, the above expession becomes

$$\boxed{\vec{A}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}kc}\hat{\imath}_{\mathcal{E}}e^{{\rm i}\vec{k}\cdot\vec{r}} \qquad \vec{\mathcal{E}}^{\rm n}_\gamma = \varepsilon_k\hat{\imath}_{\mathcal{E}}e^{{\rm i}\vec{k}\cdot\vec{r}} \qquad c\vec{\mathcal{B}}^{\rm n}_\gamma = \varepsilon_k\hat{\imath}_{\mathcal{B}}e^{{\rm i}\vec{k}\cdot\vec{r}} \qquad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}}} \tag{A.96}$$

Here $\vec{k}$ and the unit vectors $\hat{\imath}_{\mathcal{E}}$ and $\hat{\imath}_{\mathcal{B}} = \vec{k}\times\hat{\imath}_{\mathcal{E}}/k$ are all orthogonal

For convenience, the density of states as needed for Fermi's golden rule will be listed here. It was given earlier in chapter 6.3 (6.7) and 6.19:

$$\frac{{\rm d}N}{{\rm d}E} = \frac{\omega^2}{\hbar\pi^2c^3}\mathcal{V}$$

A.21.7 States of definite angular momentum

It is often convenient to describe photons in terms of states of definite net angular momentum. That makes it much easier to apply angular momentum conservation in the emission of radiation by atoms or atomic nuclei. Unfortunately, angular momentum states are a bit of a mess compared to the linear momentum states of the previous subsection. Fortunately, engineers are brave.

Before diving in, it is a good idea to look first at a *spinless* particle. Assume that this hypothetical particle is in an energy eigenstate. Also assume that this

state has square orbital angular momentum $l(l+1)\hbar^2$ where l is called the azimuthal quantum number. And that the state has orbital angular momentum in the z-direction $m_l\hbar$ where m_l is called the magnetic quantum number. Then according to quantum mechanics, chapter 4.2.3, l must be a nonnegative integer and m_l must be an integer no larger in magnitude than l. Also, in spherical coordinates (r, θ, ϕ), figure 4.7, the angular dependence of the energy eigenfunction must be given by the so-called spherical harmonic $Y_l^{m_l}(\theta, \phi)$. If in addition the particle is in empty space, the energy eigenfunction takes the general form, {A.6},

$$\psi = j_l(kr) Y_l^{m_l}(\theta, \phi) \qquad \text{with} \qquad -\nabla^2 \psi = k^2 \psi$$

Here k is a constant related to the energy of the particle and whether it is relativistic or not, {A.14} (A.44). Further j_l is the so-called "spherical Bessel function of the first kind of order l," {A.6}. The parity of the eigenfunction is positive if l is even and negative if l is odd, {D.14}. The eigenfunction is of order r^l near the origin. That is only nonzero at the origin $r = 0$ if $l = 0$. That is important if there is, say, a vanishingly small atom is located at the origin. All states except $l = 0$ are virtually zero at such an atom. So the atom only has a decent chance to interact with the particle if the particle is in a state $l = 0$. End discussion of the hypothetical spinless particle.

Now the photon is a particle with spin 1. Its wave function is essentially a vector $\vec{A}_\gamma$. The angular momentum states and parity for such particles were discussed in {A.20}. But the photon is a special case because it must be solenoidal, it must satisfy $\nabla \cdot \vec{A}_\gamma = 0$. Normally, for three-dimensional vectors you expect three types of angular momentum states, like in {A.20}. But for the photon there are only two types.

The two types of photon energy eigenfunctions with definite net angular momentum are, {D.36.2} and with drums please,

$$\boxed{\vec{A}_\gamma^{\rm E} = \nabla \times \vec{r} \times \nabla j_\ell(kr) Y_\ell^{m_\ell}(\theta, \phi) \qquad \vec{A}_\gamma^{\rm M} = \vec{r} \times \nabla j_\ell(kr) Y_\ell^{m_\ell}(\theta, \phi)} \tag{A.97}$$

Here ℓ is the azimuthal quantum number of the *net* photon angular momentum, orbital plus spin. And m_ℓ is the corresponding net magnetic quantum number.

The azimuthal quantum number ℓ is at least 1; the expressions above produce zero for $\ell = 0$. (Y_0^0 is just a constant and the gradient of a radial function is in the direction of $\vec{r}$.) The photon energy is related to the wave number k as $\hbar kc$ with c the speed of light, (A.91). That is really the Planck-Einstein relation, because kc is the photon frequency ω.

The parity of the electric multipole wave functions is negative if ℓ is odd and positive if ℓ is even, {D.36.2.7}. The parity of the magnetic multipole wave functions is exactly the other way around. From that it can be seen, {D.36.2.8}, that magnetic multipole wave functions have orbital angular momentum $l = \ell$. The electric ones have uncertainty in orbital angular momentum, with nonzero probabilities for both $l = \ell - 1$ and $l = \ell + 1$.

Atomic or nuclear transitions in which a photon in a state $\vec{A}^{\rm E}_\gamma$ is emitted or absorbed are called "electric multipole" transitions. They are indicated as Eℓ transitions.

In particular, for net angular momentum $\ell = 1$, they are called E1 or electric dipole transitions. That is the normal kind. However, as discussed in chapter 7.4, such transitions may not be able to satisfy conservation of angular momentum and parity. Since the photon in the state has $\ell = 1$, transitions in which the atomic angular momentum changes by more than one unit cannot be accomodated. Neither can transitions in which the atomic or nuclear momentum parity does not change, because the E1 photon has odd parity.

Such transitions may be accomodated by transitions in which photons in different states are emitted or absorbed, using the photon angular momenta and parities as noted above. Electric multipole transitions with $\ell = 2$ are called E2 or electric quadrupole transitions. Those with $\ell = 3$ are E3 or electric octupole ones, with $\ell = 4$ E4 or electric hexadecapole ones, with $\ell = 5$ E5 or electric triakontadipole ones, for $\ell = 6$ E6 or electric hexacontatetrapole ones and so on until your knowledge of latin and greek powers of 2 runs out.

Similarly, transitions in which photons in a state $\vec{A}^{\rm M}_\gamma$ are emitted or absorbed are called "magnetic multipole transitions." The same latin applies.

Like the states of definite linear momentum in the previous subsection, the states of definite angular momentum cannot be normalized in infinite space. To deal with that, it will be assumed that the photon is confined inside a sphere of a very large radius $r_{\max}$. As a "boundary condition" on the sphere, it will be assumed that the Bessel function is zero. In terms of the wave functions, that works out to mean that the magnetic ones are zero on the sphere, but only the radial component of the electric ones is.

The normalized wave function and unobservable fields for electric multipole photons are then, subsection A.21.5 and {D.36},

$$\boxed{\vec{A}^{\rm En}_\gamma = \frac{\varepsilon^{\rm E}_k}{{\rm i}kc}\vec{A}^{\rm E}_\gamma \qquad \vec{\mathcal{E}}^{\rm En}_\gamma = \varepsilon^{\rm E}_k \vec{A}^{\rm E}_\gamma \qquad c\vec{\mathcal{B}}^{\rm En}_\gamma = \frac{\varepsilon^{\rm E}_k k}{\rm i}\vec{A}^{\rm M}_\gamma \qquad \varepsilon^{\rm E}_k \sim \sqrt{\frac{2\hbar\omega}{\ell(\ell+1)\epsilon_0 r_{\max}}}} \tag{A.98}$$

(The expression for the magnetic field arises because for a solenoidal vector $\nabla \times \nabla \times = -\nabla^2$, and that produces a factor k^2 according to the energy eigenvalue problem.)

The normalized wave function and unobservable fields for magnetic multipole photons are

$$\boxed{\vec{A}^{\rm Mn}_\gamma = \frac{\varepsilon^{\rm E}_k}{{\rm i}c}\vec{A}^{\rm M}_\gamma \qquad \vec{\mathcal{E}}^{\rm Mn}_\gamma = k\varepsilon^{\rm E}_k \vec{A}^{\rm M}_\gamma \qquad c\vec{\mathcal{B}}^{\rm Mn}_\gamma = \frac{\varepsilon^{\rm E}_k}{\rm i}\vec{A}^{\rm E}_\gamma \qquad \varepsilon^{\rm E}_k \sim \sqrt{\frac{2\hbar\omega}{\ell(\ell+1)\epsilon_0 r_{\max}}}} \tag{A.99}$$

Assume now that there is an atom or atomic nucleus at the origin that interacts with the photon. An atom or nucleus is typically very small compared to the wave length of the photon that it interacts with. Phrased more appropriately, if R is the typical size of the atom or nucleus, then kR is typically small. The atom or nucleus is just a tiny speck at the origin.

Now the wave functions $\vec{A}^{\rm E}_\gamma$ are larger at small radii than the $\vec{A}^{\rm M}_\gamma$. In particular, the $\vec{A}^{\rm E}_\gamma$ are of order $r^{\ell-1}$ while the $\vec{A}^{\rm M}_\gamma$ are of order r^ℓ, one power of r smaller. These powers of r reflect the lowest measurable orbital angular momentum of the states.

A glance at the unobservable fields of electric multipole photons above then shows that for these photons, the field is primarily electric at the atom or nucleus. And even the electric field will be small unless $\ell = 1$, in other words, unless it is an electric dipole photon.

For the magnetic multipole photons, it is the magnetic field that dominates at the atom or nucleus. And even that will be small unless $\ell = 1$, which means a magnetic dipole photon. Note that the magnetic field acts as if it had one unit or orbital angular momentum less than the photon; the magnetic field is essentially the wave function of an electric multipole photon.

For later reference, the density of states as needed for Fermi's golden rule will be listed here, {D.36.2.6}:

$$\boxed{\frac{{\rm d}N}{{\rm d}E} \approx \frac{1}{\hbar\pi c} r_{\max}} \tag{A.100}$$

This approximation applies for large cut-off radius $r_{\max}$, which should always be valid.

A.22 Forces by particle exchange

As noted in chapter 7.5.2, the fundamental forces of nature arise from the exchange of particles. This addendum will illustrate the general idea. It will first derive the hypothetical "Koulomb" force due to the exchange of equally hypothetical particles called "fotons."

The Koulomb potential provides a fairly simple model of a quantum field. It also provides a simple context to introduce some key concepts in quantum field theories, such as Green's functions, variational calculus, Lagrangians, the limitation of the speed of light, description in terms of momentum modes, Fock space kets, annihilation and creation operators, antiparticles, special relativity, the imperfections of physicists, and Lorentz invariance. The Koulomb potential can also readily be modified to explain nuclear forces. However, that will have to wait until a later addendum, {A.42}.

In the current addendum, the Koulomb potential provides the starting point for a discussion of the electromagnetic field. The classical Maxwell equations for

the electromagnetic field will be derived in a slightly unconventional way. Who needs to know classical electromagnetics when all it takes is quantum mechanics, relativity, and a few plausible guesses to derive electromagnetics from scratch?

To quantize the electromagnetic field is not that straightforward; it has unexpected features that do not occur for the Koulomb field. This book follows the derivation as formulated by Fermi in 1932. This derivation is the basis for more advanced modern quantum field approaches. These advanced theories will not be covered, however.

Essentially, the Fermi derivation splits off the Coulomb potential from the electromagnetic field. What is left is then readily described by a simple quantum field theory much like for the Koulomb potential. This is sufficient to handle important applications such as the emission or absorption of radiation by atoms and atomic nuclei. That, however, will again be done in subsequent addenda.

A word to the wise. While this addendum is on the calculus level like virtually everything else in this book, there is just quite a lot of mathematics. Some mathematical maturity may be needed not to get lost. Note that this addendum is *not* needed to understand the discussion of the emission and absorption of radiation in the subsequent addenda.

A.22.1 Classical selectostatics

The Koulomb force holds the sarged spotons and selectons together in satoms. The force is due to the exchange of massless particles called fotons between the sarged particles. (It will be assumed that the spoton is an elementary particle, though really it consists of three skarks.)

This subsection will derive the selectostatic Koulomb force by representing the fotons by a classical field, not a quantum field. The next subsection will explain classical selectodynamics, and how it obeys the speed of light. Subsection A.22.3 will eventually fully quantize the selectic field. It will show how quantum effects modify some of the physics expected from the classical analysis.

Physicists have some trouble measuring the precise properties of the selectic field. However, a few basic quantum ideas and some reasonable guesses readily substitute for the lack of empirical data. And guessing is good. If you can guess a self-consistent Koulomb field, you have a lot of insight into its nature.

Consider first the wave function for the exchanged foton in isolation. A foton is a boson without spin. That means that its wave function is a simple function, not some vector. But since the foton is massless, the Schrödinger equation does not apply to it. The appropriate equation follows from the relativistic expression for the energy of a massless particle as given by Einstein, chapter 1.1.2 (1.2):

$$E^2 = \vec{p}^{\,2}c^2$$

Here E is the foton energy, $\vec{p}$ its linear momentum, and c the speed of light. The squares are used because momentum is really a vector, not a number like energy.

Quantum mechanics replaces the momentum vector by the operator

$$\widehat{\vec{p}} = \frac{\hbar}{\mathrm{i}}\nabla \qquad \nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

Note the vector operator ∇, called "nabla" or "del." This operator is treated much like an ordinary vector in various computations. Its properties are covered in Calculus III in the U.S. system. (Brief summaries of properties of relevance here can be found in the notations section.)

The Hamiltonian eigenvalue problem for a foton wave function φ_{f} then takes the form

$$\widehat{\vec{p}}^{\,2}c^2\varphi_{\mathrm{f}} = E^2\varphi_{\mathrm{f}}$$

A solution φ_{f} to this equation is an energy eigenstate. The corresponding value of E is the energy of the state. (To be picky, the above is an eigenvalue problem for the square Hamiltonian. But eigenfunctions of an operator are also eigenfunctions of the square operator. The reverse is not always true, but that is not a concern here.)

Using the momentum operator as given above and some rearranging, the eigenvalue problem becomes

$$-\nabla^2\varphi_{\mathrm{f}} = \frac{E^2}{\hbar^2c^2}\varphi_{\mathrm{f}} \qquad \nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{A.101}$$

This is called the "time-independent Klein-Gordon equation" for a massless particle.

For foton wave functions that are not necessarily energy eigenstates, quantum mechanics replaces the energy E by the operator $\mathrm{i}\hbar\partial/\partial t$. That gives the time-dependent Klein-Gordon equation as:

$$-\nabla^2\varphi_{\mathrm{f}} = -\frac{1}{c^2}\frac{\partial^2\varphi_{\mathrm{f}}}{\partial t^2} \tag{A.102}$$

Now consider solutions of this equation of the form

$$\varphi_{\mathrm{f}}(\vec{r};t) = e^{-\mathrm{i}\omega t}\varphi_{\mathrm{fs}}(\vec{r})$$

Here ω is a positive constant called the angular frequency. Substitution in the time-dependent Klein-Gordon equation shows that this solution also satisfies the time-independent Klein-Gordon equation, with energy

$$E = \hbar\omega$$

That is the famous "Planck-Einstein" relation. It is implicit in the association of E with $\mathrm{i}\hbar\partial/\partial t$.

Note however that there will also be a solution of the form

$$\varphi_{\rm f}(\vec{r};t) = e^{{\rm i}\omega t}\varphi_{\rm fs}(\vec{r})$$

This solution too has energy $\hbar\omega$. The difference in sign in the exponential is taken to mean that the particle moves backwards in time. Note that changing the sign in the exponential is equivalent to changing the sign of the time t. At least it is if you require that $\omega = E/\hbar$ cannot be negative. If a particle moves backwards in time, it is called an "antiparticle." So the wave function above describes an "antifoton."

There is really no physical difference between a foton and an antifoton. That is not necessarily true for other types of particles. Quantities such as electric charge, lepton number, baryon number, strangeness, etcetera take opposite values for a particle and its antiparticle.

There is a very important difference between the Klein-Gordon equation and the Schrödinger equation. The Schrödinger equation describes nonrelativistic physics where particles can neither be destroyed nor created. Mass must be conserved. But the Klein-Gordon equation applies to relativistic physics. In relativistic physics particles can be created out of pure energy or destroyed following Einstein's famous relationship $E = mc^2$, chapter 1.

There is a mathematical consequence to this. It concerns the integral

$$\int |\varphi_{\rm f}|^2 {\rm d}^3\vec{r}$$

(In this addendum, integrals like this are over all space unless explicitly stated otherwise. It is also assumed that the fields vanish quickly enough at large distances that such integrals are finite. Alternatively, for particles confined in a large box it is assumed that the box is periodic, chapter 6.17.) Now for solutions of the Schrödinger equation, the integral $\int |\varphi_{\rm f}|^2 {\rm d}^3\vec{r}$ keeps the same value, 1, for all time. Physically, the integral represent the probability of finding the particle. The probability of finding the particle if you look in all space must be 1.

But fotons are routinely destroyed or created by sarged particles. So the probability of finding a foton is not a preserved quantity. (It is not even clear what finding a foton would mean in the first place.) The Klein-Gordon equation reflects that. It does not preserve the integral $\int |\varphi_{\rm f}|^2 {\rm d}^3\vec{r}$. (There is one exception: if the wave function is purely described by particle states or purely described by antiparticle states, the integral is still preserved.)

But the Klein-Gordon equation does preserve an other integral, {D.32}. That is

$$\int \left| \frac{1}{c}\frac{\partial \varphi_{\rm f}}{\partial t} \right|^2 + |\nabla \varphi_{\rm f}|^2 \, {\rm d}^3\vec{r}$$

Now if the number of fotons is not a preserved quantity, what can this preserved integral stand for? Not momentum or angular momentum, which

are vectors. The integral must obviously stand for the energy. Energy is still preserved in relativity, even if the number of particles of a given type is not.

Of course, the energy of a foton wave function $\varphi_{\rm f}$ is also given by the Planck-Einstein relation. But wave functions are not observable. Still, fotons do affect spotons and selectons. That is observable. So there must be an observable foton field. This observable field will be called the foton potential. It will be indicated by simply φ, without a subscript f. Quantum uncertainty in the values of the field will be ignored in this subsection. So the field will be modeled as a classical (i.e. nonquantum) field.

And if there is an observable field, there must be an observable energy associated with that field. Now what could the expression for the energy in the field be? Obviously it will have to take the form of the integral above. What other options are there that are plausible? Of course, there will be some additional empirical constant. If the integral is constant, then any multiple of it will be constant too. And the above integral will not have units of energy as it is. The needed empirical constant is indicated by ϵ_1 and is called, um no, the permissivity of space. It is a measure of how efficient the foton field is in generating energy. To be precise, for arcane historical reasons the constant in the energy is actually defined as half the permissivity. The bottom line is that the expression for the energy in the observable foton field is:

$$E_\varphi = \frac{\epsilon_1}{2} \int \left| \frac{1}{c} \frac{\partial \varphi}{\partial t} \right|^2 + |\nabla \varphi|^2 \, {\rm d}^3 \vec{r} \tag{A.103}$$

That is really all that is needed to figure out the properties of classical selectostatics in this subsection. It will also be enough to figure out classical selectodynamics in the next subsection.

The first system that will be considered here is that of a foton field and a single spoton. It will be assumed that the spoton is pretty much located at the origin. Of course, in quantum mechanics a particle must have some uncertainty in position, or its kinetic energy would be infinite. But it will be assumed that the spoton wave function is only nonzero within a small distance ε of the origin. Beyond that distance, the spoton wave function is zero.

However, since this is a classical derivation and not a quantum one, the term "spoton wave function" must not be used. So imagine instead that the spoton sarge $s_{\rm p}$ is smeared out over a small region of radius ε around the origin.

For a smeared out sarge, there will be a "sarge density" $\sigma_{\rm p}$, defined as the local sarge per unit volume. This sarge density can be expressed mathematically as

$$\sigma_{\rm p}(\vec{r}) = s_{\rm p} \delta^3_\varepsilon(\vec{r})$$

Here $\delta^3_\varepsilon(\vec{r})$ is some function that describes the detailed shape of the smeared-out sarge distribution. The integral of this function must be 1, because the sarge

density $\sigma_{\rm p}$ must integrate to the total spoton sarge $s_{\rm p}$. So:

$$\int \delta_\varepsilon^3(\vec{r})\,\mathrm{d}^3\vec{r} = 1$$

To ensure that the sarge density is zero for distances from the origin r greater than the given small value ε, $\delta_\varepsilon^3(\vec{r})$ must be zero at these distances. So:

$$\delta_\varepsilon^3(\vec{r}) = 0 \quad \text{if} \quad r \equiv |\vec{r}| \geqslant \varepsilon$$

In the limit that ε becomes zero, $\delta_\varepsilon^3(\vec{r})$ becomes the so-called three-dimensional "Dirac delta function" $\delta^3(\vec{r})$. This function is totally concentrated at a single point, the origin. But its integral over that single point is still 1. That is only possible if the function value at the point is infinite. Now infinity is not a proper number, and so the Dirac delta function is not a proper function. However, mathematicians have in fact succeeded in generalizing the idea of functions to allow delta functions. That need not concern the discussion here because "Physicists are sloppy about mathematical rigor," as Zee [53, p. 22] very rightly states. Delta functions are named after the physicist Dirac. They are everywhere in quantum field theory. That is not really surprising as Dirac was one of the major founders of the theory. See section 7.9 for more on delta functions.

Here the big question is how the spoton manages to create a foton field around itself. That is not trivial. If there was a nonzero probability of finding an energetic foton well away from the spoton, surely it would violate energy conservation. However, it turns out that the time-independent Klein-Gordon equation (A.101) actually has a very simple solution where the foton energy E appears to be zero away from the origin. In spherical coordinates, it is

$$\varphi_{\rm f} = \frac{C}{r} \quad \text{if} \quad r \neq 0$$

Here C is some constant which is still arbitrary about this stage. To check the above solution, plug it into the energy eigenvalue problem (A.101) with E zero.

This then seems to be a plausible form for the observable potential φ away from the spoton at the origin. However, while the energy of a C/r potential appears to be zero, it is not really. Such a potential is infinite at the origin, and you cannot just ignore that. The correct foton field energy is given by the earlier integral (A.103). For a steady potential, it can be written as

$$E_\varphi = \frac{\epsilon_1}{2}\int (\nabla\varphi)^2\,\mathrm{d}^3\vec{r} = -\frac{\epsilon_1}{2}\int \varphi\left(\nabla^2\varphi\right)\,\mathrm{d}^3\vec{r} \tag{A.104}$$

The final integral comes from an integration by parts. (See {A.2} and {D.32} for examples how to do such integrations by parts.) Note that it looks like the energy could be zero according to this final integral: $\nabla^2\varphi$ is zero if $\varphi = C/r$.

That is true outside the small vicinity around the origin. But if you look at the equivalent first integral, it is obvious that the energy is not zero: its integrand is everywhere positive. So the energy must be positive. It follows that in the final integral, the region around the origin, while small, still produces an energy that is not small. The integrand must be not just nonzero, but large in this region.

All that then raises the question why there is a foton field in the first place. The interest in this subsection is in the selectostatic field. That is supposed to be the stable ground state of lowest energy. According to the above, the state of lowest energy would be when there is no foton field; $\varphi = 0$.

And so it is. The only reasonable way to explain that there is a nontrivial foton field in the ground state of the spoton-foton system is if the foton field energy is compensated for by something else. There must be an energy of interaction between the foton field and the spoton.

Consider the mathematical form that this energy could take in a given volume element $\mathrm{d}^3\vec{r}$. Surely the simplest possibility is that it is proportional to the potential φ at the location times the sarge $\sigma_{\rm p}\mathrm{d}^3\vec{r}$. Therefore the total energy of spoton-foton interaction is presumably

$$E_{\varphi{\rm p}} = -\int \varphi(\vec{r})\sigma_{\rm p}(\vec{r})\,\mathrm{d}^3\vec{r} = -\int \varphi(\vec{r})s_{\rm p}\delta^3_\varepsilon(\vec{r})\,\mathrm{d}^3\vec{r} \tag{A.105}$$

Note that this expression really *defines* the sarge $s_{\rm p}$. Sarge gives the strength of the coupling between spoton and foton field. Its units and sign follow from writing the energy as the expression above.

The question is now, what is the ground state foton field? In other words, for what potential φ is the complete system energy minimal? To answer that requires "variational calculus." Fortunately, variational calculus is just calculus. And you need to understand how it works if you want to make any sense at all out of books on quantum field theory.

Suppose that you wanted an equation for the minimum of some function f depending on a single variable x. The equation would be that $\mathrm{d}f/\mathrm{d}x = 0$ at the position of the minimum $x_{\rm min}$. In terms of differentials, that would mean that the function does not change going from position $x_{\rm min}$ to a slightly different position $x_{\rm min} + \mathrm{d}x$:

$$\mathrm{d}f = \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x = 0 \quad \text{at} \quad x = x_{\rm min}$$

It is the same for the change in net energy $E_\varphi + E_{\varphi{\rm p}}$. Assume that $\varphi_{\rm min}$ is the desired potential at minimum net energy. Then at $\varphi_{\rm min}$ the net energy should not change when you change φ by an infinitesimal amount $\mathrm{d}\varphi$. Or rather, by an infinitesimal amount $\delta\varphi$: the symbol δ is used in variational calculus instead of d. That is to avoid confusion with any symbol d that may already be around.

So the requirement for the ground state potential is

$$\delta(E_\varphi + E_{\varphi{\rm p}}) = 0 \quad \text{when} \quad \varphi = \varphi_{\rm min} \;\rightarrow\; \varphi = \varphi_{\rm min} + \delta\varphi$$

Using the expressions (A.104) and (A.105) for the energies, that means that

$$\delta\left[\frac{\epsilon_1}{2}\int(\nabla\varphi)^2\,\mathrm{d}^3\vec{r} - \int s_\mathrm{p}\delta_\varepsilon^3\varphi\,\mathrm{d}^3\vec{r}\right] = 0 \quad \text{when} \quad \varphi = \varphi_\mathrm{min} \;\rightarrow\; \varphi = \varphi_\mathrm{min} + \delta\varphi$$

The usual rules of calculus can be used, (see {A.2} for more details). The only difference from basic calculus is that the change $\delta\varphi$ may depend on the point that you look at. In other words, it is some arbitrary but small function of the position $\vec{r}$. For example,

$$\delta(\nabla\varphi)^2 = 2(\nabla\varphi)\delta(\nabla\varphi) \qquad \delta(\nabla\varphi) = \nabla(\varphi_\mathrm{min} + \delta\varphi) - \nabla(\varphi_\mathrm{min}) = \nabla\delta\varphi$$

Also, φ by itself is validly approximated as φ_min, but $\delta\varphi$ is a completely separate quantity that can be anything. Working it out gives

$$\frac{\epsilon_1}{2}\int 2(\nabla\varphi_\mathrm{min})\cdot(\nabla\delta\varphi)\,\mathrm{d}^3\vec{r} - \int s_\mathrm{p}\delta_\varepsilon^3\delta\varphi\,\mathrm{d}^3\vec{r} = 0$$

Performing an integration by parts moves the ∇ from $\delta\varphi$ to $\nabla\varphi_\mathrm{min}$ and adds a minus sign. Then the two integrals combine as

$$-\int\left(\epsilon_1\nabla^2\varphi_\mathrm{min} + s_\mathrm{p}\delta_\varepsilon^3\right)\delta\varphi\,\mathrm{d}^3\vec{r} = 0$$

If this is supposed to be zero for *whatever* you take the small change $\delta\varphi$ in field to be, then the parenthetical expression in the integral will have to be zero. If the parenthetical expression is nonzero somewhere, you can easily make up a nonzero change $\delta\varphi$ in that region so that the integral is nonzero.

The parenthetical expression can now be rearranged to give the final result:

$$-\nabla^2\varphi = \frac{s_\mathrm{p}}{\epsilon_1}\delta_\varepsilon^3 \tag{A.106}$$

Here the subscript "min" was left away again as the ground state is the only state of interest here anyway.

The above equation is the famous "Poisson equation" for the selectostatic potential φ. The same equation appears in electrostatics, chapter 13.3.4. So far, this is all quite encouraging. Note also that the left hand side is the steady Klein-Gordon equation. The right hand side is mathematically a "forcing" term; it forces a nonzero solution for φ.

Beyond the small vicinity of radius ε around the origin, the spoton sarge density in the right hand side is zero. That means that away from the spoton, you get the time-independent Klein-Gordon equation (A.101) with $E = 0$. That was a good guess, earlier. Assuming spherical symmetry, away from the spoton the solution to the Poisson equation is then indeed

$$\varphi = \frac{C}{r} \quad \text{if} \quad r \geqslant \varepsilon$$

But now that the complete Poisson equation (A.106) is known, the constant C can be figured out, {D.2}. The precise field turns out to be

$$\varphi = \frac{s_{\rm p}}{4\pi\epsilon_1 r} \quad \text{if} \quad r \geqslant \varepsilon$$

For unit value of $s_{\rm p}/\epsilon_1$ the above solution is called the "fundamental solution" or "Green's function" of the Poisson equation. It is the solution due to a delta function.

If the spoton is not at the origin, but at some position $\vec{r}_{\rm p}$, you simply replace r by the distance from that point:

$$\varphi^{\rm p} = \frac{s_{\rm p}}{4\pi\epsilon_1 |\vec{r} - \vec{r}_{\rm p}|} \tag{A.107}$$

The superscript p indicates that this potential is created by a spoton at a position $\vec{r}_{\rm p}$. This solution of the Poisson equation will become very important in the Fermi derivation.

Now the net energy is of interest. It can be simplified by substituting the Poisson equation (A.106) in the expression (A.104) for the foton field energy and adding the interaction energy (A.105). That gives

$$E_\varphi + E_{\varphi{\rm p}} = \tfrac{1}{2} \int \varphi(\vec{r}) s_{\rm p} \delta^3_\varepsilon(\vec{r}) \, {\rm d}^3\vec{r} - \int \varphi(\vec{r}) s_{\rm p} \delta^3_\varepsilon(\vec{r}) \, {\rm d}^3\vec{r}$$

which simplifies to

$$E_\varphi + E_{\varphi{\rm p}} = -\tfrac{1}{2} \int \varphi(\vec{r}) s_{\rm p} \delta^3_\varepsilon(\vec{r}) \, {\rm d}^3\vec{r} \tag{A.108}$$

Note that the spoton-foton interaction energy is twice the foton field energy, and negative instead of positive. That means that the total energy has been *lowered* by an amount equal to the foton field energy, despite the fact that the field energy itself is positive.

The fact that there is a foton field in the ground state has now been explained. The interaction with the spoton lowers the energy more than the field itself raises it.

Note further from the solution for φ above that φ is large in the vicinity of the spoton. As a result, the energy in the foton field becomes infinite when the spoton sarge contracts to a point. (That is best seen from the original integral for the foton field energy in (A.104).) This blow up is very similar to the fact that the energy in a classical electromagnetic field is infinite for a point charge. For the Koulomb field, the interaction energy blows up too, as it is twice the foton field energy. All these blow ups are a good reason to use a sarge density rather than a point sarge. Then all energies are normal finite numbers.

The final step to derive the classical Koulomb force is to add a selecton. The selecton is also sarged, so it too generates a field. To avoid confusion, from now

on the field generated by the spoton will always be indicated by $\varphi^{\rm p}$, and the one generated by the selecton by $\varphi^{\rm e}$. The variational analysis can now be repeated including the selecton, {D.37.1}. That shows that there are three effects that produce the Koulomb force between the spoton and selecton:

1. the selecton sarge interacts with the potential $\varphi^{\rm p}$ generated by the spoton;
2. the spoton sarge interacts with the potential $\varphi^{\rm e}$ generated by the selecton;
3. the energy in the combined foton field $\varphi^{\rm p}+\varphi^{\rm e}$ is different from the sum of the energies of the separate fields $\varphi^{\rm p}$ and $\varphi^{\rm e}$.

All three effects turn out to produce the same energy, but the first two energies are negative and the third positive. So the net energy change is the same as if there was just item 1, the interaction of the selecton sarge density $\sigma_{\rm e}$ with the potential $\varphi^{\rm p}$ produced by the spoton. That is of course given by a similar expression as before:

$$V_{\rm ep} = -\int \varphi^{\rm p}(\vec{r})\sigma_{\rm e}(\vec{r})\,{\rm d}^3\vec{r}$$

The expression for $\varphi^{\rm p}(\vec{r})$ was given above in (A.107) for any arbitrary position of the spoton $\vec{r}_{\rm p}$. And it will be assumed that the selecton sarge density $\sigma_{\rm e}$ is spread out a bit just like the spoton one, but around a different location $\vec{r}_{\rm e}$. Then the interaction energy becomes

$$V_{\rm ep} = -\int \frac{s_{\rm p}}{4\pi\epsilon_1|\vec{r}-\vec{r}_{\rm p}|} s_{\rm e}\delta^3_\varepsilon(\vec{r}-\vec{r}_{\rm e})\,{\rm d}^3\vec{r}$$

Since ε is assumed small, the selecton sarge density is only nonzero very close to the nominal position $\vec{r}_{\rm e}$. Therefore you can approximate $\vec{r}$ as $\vec{r}_{\rm e}$ in the fraction and take it out of the integral as a constant. Then the delta function integrates to 1, and you get

$$V_{\rm ep} = -\frac{s_{\rm p}s_{\rm e}}{4\pi\epsilon_1|\vec{r}_{\rm e}-\vec{r}_{\rm p}|} \tag{A.109}$$

That then is the final energy of the Koulomb interaction between the two sarged particles. Because the spoton and the selecton both interact with the foton field, in effect it produces a spoton-selecton interaction energy.

Of course, in classical physics you would probably want to know the actual force on say the selecton. To get it, move the origin of the coordinate system to the spoton and rotate it so that the selecton is on the positive x-axis. Now give the selecton a small displacement $\partial x_{\rm e}$ in the x-direction. Slowly of course; this is supposed to be selectostatics. Because of energy conservation, the work done by the force $F_{x{\rm e}}$ during this displacement must cause a corresponding small decrease in energy. So:

$$F_{x_{\rm e}}\partial x_{_{\rm e}} = -\partial V_{\rm ep}$$

But on the positive x-axis, $|\vec{r}_{\rm e} - \vec{r}_{\rm p}|$ is just the x-position of the selecton $x_{\rm e}$, so

$$F_{x_{\rm e}} = -\frac{\partial V_{\rm ep}}{\partial x_{\rm e}} = -\frac{s_{\rm p} s_{\rm e}}{4\pi\epsilon_1 x_{\rm e}^2}$$

It is seen that if the sarges have equal sign, the force is in the negative x-direction, towards the spoton. So sarges of the same sign attract.

More generally, the force on the selecton points towards the spoton if the sarges are of the same sign. It points straight away from the spoton if the sarges are of opposite sign.

The Koulomb energy $V_{\rm ep}$ looks almost exactly the same as the Coulomb energy in electrostatics. Recall that the Coulomb energy was used in chapter 4.3 to describe the attraction between the proton and electron in a hydrogen atom. The difference is that the Coulomb energy has no minus sign. That means that while like sarges attract, like charges repel each other. For example, two spotons attract, but two protons repel.

Now a spoton must necessarily create a foton field that is attractive to spotons. Otherwise there should be no field at all in the ground state. And if spotons create fields that attract spotons, then spotons attract. So the Koulomb force is clearly right.

It is the Coulomb force that does not seem to make any sense. Much more will be said about that in later subsections.

A.22.2 Classical selectodynamics

According to the previous section the Koulomb energy between a spoton and a selecton is given by

$$V_{\rm ep} = -\frac{s_{\rm p} s_{\rm e}}{4\pi\epsilon_1 |\vec{r}_{\rm e} - \vec{r}_{\rm p}|}$$

However, this result can only be correct in a stationary state like a ground state, or maybe some other energy state.

To see the problem, imagine that the spoton is suddenly given a kick. According to the Koulomb potential given above, the selecton notices that instantly. There is no time in the Koulomb potential, so there is no time delay. But Einstein showed that no observable effect can move faster than the speed of light. So there should be a time delay.

Obviously then, to discuss unsteady evolution will require the full governing equations for selectodynamics. The big question is how to find these equations.

The quantum mechanics in this book is normally based on some Hamiltonian H. But there is a more basic quantity for a system than the Hamiltonian. That quantity is called the "Lagrangian" $\mathcal{L}$. If you can guess the correct Lagrangian of a system, its equations of motion follow. That is very important for quantum field theories. In fact, a lot of what advanced quantum field theories really do is guess Lagrangians.

To get at the Lagrangian for selectodynamics, consider first the motion of the spoton for a *given* foton field φ. The Hamiltonian of the spoton by itself is just the energy of the spoton. As discussed in the previous subsection, a spoton has a potential energy of interaction with the given foton field

$$E_{\varphi\mathrm{p}} = -\int \varphi(\vec{r};t)\sigma_{\mathrm{p}}(\vec{r};t)\,\mathrm{d}^3\vec{r} \tag{A.110}$$

Here σ_{p} was the sarge density of the spoton. The foton potential and sarge density can now of course also depend on time.

However, to discuss the dynamics of the spoton, it is easier to consider it a point particle located at a single moving point $\vec{r}_{\mathrm{p}}$. Therefore it will be assumed that the sarge density is completely concentrated at that one point. That means that the only value of the foton field of interest is the value at $\vec{r}_{\mathrm{p}}$. And the sarge distribution integrates to the net spoton sarge s_{p}. So the above energy of interaction becomes approximately

$$E_{\varphi\mathrm{p}} \approx -\varphi_{\mathrm{p}} s_{\mathrm{p}} \qquad \varphi_{\mathrm{p}} \equiv \varphi(\vec{r}_{\mathrm{p}};t) \tag{A.111}$$

In terms of the components of position, this can be written out fully as

$$E_{\varphi\mathrm{p}} \approx -\varphi(r_{\mathrm{p}_1}, r_{\mathrm{p}_2}, r_{\mathrm{p}_3};t)s_{\mathrm{p}}$$

Note that in this addendum the position components are indicated as r_{p_1}, r_{p_2}, and r_{p_3} instead of the more familiar r_{p_x}, r_{p_y}, and r_{p_z} or x_{p}, y_{p}, and z_{p}. That is in order that a generic position component can be indicated by r_{p_i} where i can be 1, 2, or 3.

In addition to the interaction energy above there is the kinetic energy of the spoton,

$$E_{\mathrm{p,kin}} = \tfrac{1}{2}m_{\mathrm{p}}\vec{v}_{\mathrm{p}}^{\,2}$$

Here m_{p} is the mass of the spoton and $\vec{v}_{\mathrm{p}}$ its velocity,

$$\vec{v}_{\mathrm{p}} \equiv \frac{\mathrm{d}\vec{r}_{\mathrm{p}}}{\mathrm{d}t}$$

The kinetic energy can be written out in terms of the velocity components as

$$E_{\mathrm{p,kin}} = \tfrac{1}{2}m_{\mathrm{p}}\left[(v_{\mathrm{p}_1})^2 + (v_{\mathrm{p}_2})^2 + (v_{\mathrm{p}_3})^2\right] \qquad \text{with } v_{\mathrm{p}_i} = \frac{\mathrm{d}r_{\mathrm{p}_i}}{\mathrm{d}t} \text{ for } i = 1,2,3$$

Now the Hamiltonian of the spoton is the sum of the kinetic and potential energies. But the Lagrangian is the *difference* between the kinetic and potential energies:

$$\mathcal{L}_{\mathrm{p}}(\vec{v}_{\mathrm{p}}, \vec{r}_{\mathrm{p}}) = \tfrac{1}{2}m_{\mathrm{p}}\vec{v}_{\mathrm{p}}^{\,2} + \varphi(\vec{r}_{\mathrm{p}};t)s_{\mathrm{p}}$$

This Lagrangian can now be used to find the equation of motion of the spoton. This comes about in a somewhat weird way. Suppose that there is some range of times, from a time t_1 to a time t_2, during which you want to know the motion of the spoton. (Maybe the spoton is at rest at time t_1 and becomes again at rest at time t_2.) Suppose further that you now compute the so-called "action" integral

$$\mathcal{S} \equiv \int_{t_1}^{t_2} \mathcal{L}_{\rm p}(\vec{v}_{\rm p}, \vec{r}_{\rm p})\,{\rm d}t$$

If you use the correct velocity and position of the spoton, you will get some number. But now suppose that you use a slightly different (wrong) spoton path. Suppose it is different by a small amount $\delta\vec{r}_{\rm p}$, which of course depends on time. You would think that the value of the action integral would change by a corresponding small amount. But that is not true. Assuming that the path used in the original integral was indeed the right one, and that the change in path is infinitesimally small, the action integral does not change. Mathematically

$$\delta\mathcal{S} = 0 \quad \text{at the correct path}$$

Yes, this is again variational calculus. The action may not be minimal at the correct spoton path, but it is definitely stationary at it.

Probably this sounds like a stupid mathematical trick. But in the so-called path integral approach to quantum field theory, the action is *central* to the formulation.

For classical physics the action by itself is pretty useless. However, with some manipulations, you can get the evolution equations for your system out of it, {A.1}. They are found as

$$\frac{\rm d}{{\rm d}t}\left(\frac{\partial\mathcal{L}}{\partial v_{{\rm p}_i}}\right) = \left(\frac{\partial\mathcal{L}}{\partial r_{{\rm p}_i}}\right) \tag{A.112}$$

Here $i = 1$, 2, or 3 gives the equation in the x, y, or z direction, respectively.

Note that for the governing equations it does not matter at all what you take the times t_1 and t_2 in the action to be. They are pretty vaguely defined anyway. You might want to let them go to minus and plus infinity to get rid of them.

The next step is to write out the governing equation (A.112) in terms of physical quantities. To do that correctly, the trick is that the Lagrangian must be treated as a function of velocity and position, as independent variables. In reality velocity and position are not independent; velocity is the derivative of position. But when differentiating the Lagrangian you are supposed to forget about that. Consider how this works out for the x-component, $i = 1$,

$$\frac{\partial\mathcal{L}}{\partial v_{{\rm p}_1}} = m_{\rm p} v_{{\rm p}_1} \qquad \frac{\partial\mathcal{L}}{\partial r_{{\rm p}_1}} = \frac{\partial\varphi(r_{{\rm p}_1}, r_{{\rm p}_2}, r_{{\rm p}_3}; t)}{\partial r_{{\rm p}_1}} s_{\rm p}$$

That are simple differentiations taking the given Lagrangian at face value.

However, when you do the remaining time derivative in (A.112) you have to do it properly, treating the velocity as the function of time that it is. That gives the final equation of motion as

$$m_{\rm p}\frac{{\rm d}v_{{\rm p}_1}}{{\rm d}t} = \frac{\partial\varphi(r_{{\rm p}_1}, r_{{\rm p}_2}, r_{{\rm p}_3}; t)}{\partial r_{{\rm p}_1}} s_{\rm p} \tag{A.113}$$

Note that the left hand side is mass times acceleration in the x-direction. So the right hand side must be the selectic force on the spoton. This force is called the "Sorentz force." It is seen that the Sorentz force is proportional to the derivative of the foton potential, evaluated at the position of the spoton. If you compare the Sorentz force with the force in electrostatics, you see that the force in electrostatics has an additional minus sign. That reflects again that equal sarges attract, while equal charges repel.

So far, it was assumed that the foton field was given. But in reality the foton field is not given, it depends on the motion of the spoton. To describe the field, its energies must be added to the Lagrangian too. The total energy in the foton field was given in the previous subsection as (A.103). Using some shorthand notation, this becomes

$$E_\varphi = \frac{\epsilon_1}{2}\int \frac{1}{c^2}\varphi_t^2 + \sum_{i=1}^{3}\varphi_i^2\,{\rm d}^3\vec{r}$$

The shorthand is to indicate derivatives by subscripts, as in

$$\varphi_t \equiv \frac{\partial\varphi}{\partial t} \qquad \varphi_i \equiv \frac{\partial\varphi}{\partial r_i}$$

with $i = 1$, 2, or 3 for the x, y, and z derivatives respectively. For example, φ_1 would be the partial x-derivative of φ.

Actually, even more concise shorthand will be used. If an index like i occurs twice in a term, summation over that index is to be understood. The summation symbol will then not be shown. That is called the Einstein summation convention. So the energy in the foton field will be indicated briefly as

$$E_\varphi = \frac{\epsilon_1}{2}\int \frac{1}{c^2}\varphi_t^2 + \varphi_i^2\,{\rm d}^3\vec{r}$$

(Note that $\varphi_i^2 = \varphi_i\varphi_i$, so i occurs twice in the second term of the integrand.) All this is done as a service to you, the reader. You are no doubt getting tired of having to look at all these mathematical symbols.

Now the first term in the energy above is a time derivative, just like $\vec{v}_{\rm p}$ was the time derivative of the spoton position. So this term has presumably the

same sign in the Lagrangian, while the sign of the other term flips over. That makes the total selectodynamic Lagrangian equal to

$$\mathcal{L}_{\varphi\mathrm{p}} = \frac{\epsilon_1}{2} \int \frac{1}{c^2}\varphi_t^2 - \varphi_i^2 \,\mathrm{d}^3\vec{r} + \tfrac{1}{2} m_\mathrm{p} \vec{v}_\mathrm{p}^2 + \varphi_\mathrm{p} s_\mathrm{p}$$

The last two terms are as before for a given field.

However, for the final term it is now desirable to go back to the representation of the spoton in terms of a sarge density σ_p, as in (A.110). The final term as written would lead to a nasty delta function in the analysis of the field. In the sarge density form the term can be brought inside the integral to give the complete Lagrangian as

$$\mathcal{L}_{\varphi\mathrm{p}} = \int \frac{\epsilon_1}{2} \left(\frac{1}{c^2}\varphi_t^2 - \varphi_i^2 \right) + \varphi\sigma_\mathrm{p} \,\mathrm{d}^3\vec{r} + \tfrac{1}{2} m_\mathrm{p} \vec{v}_\mathrm{p}^2 \tag{A.114}$$

Note that there is no longer a subscript p on φ; it is the integration against the sarge density that picks out the value of φ at the spoton.

An integrand of a spatial integral in a Lagrangian is called a "Lagrangian density" and indicated by the symbol $\pounds$. In this case:

$$\pounds = \frac{\epsilon_1}{2} \left(\frac{1}{c^2}\varphi_t^2 - \varphi_i^2 \right) + \varphi\sigma_\mathrm{p} \tag{A.115}$$

When differentiating this Lagrangian density, φ and its derivatives φ_t and φ_i, (with $i = 1$, 2, and 3), are to be considered 5 separate independent variables.

The action principle can readily be extended to allow for Lagrangian densities, {D.37}. The equations of motion for the field are then found to be

$$\frac{\partial}{\partial t}\left(\frac{\partial \pounds}{\partial \varphi_t}\right) + \frac{\partial}{\partial r_i}\left(\frac{\partial \pounds}{\partial \varphi_i}\right) = \frac{\partial \pounds}{\partial \varphi}$$

Working this out much like for the equation of motion of the spoton gives, taking ε_1 to the other side,

$$\frac{1}{c^2}\frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial r_i^2} = \frac{\sigma_\mathrm{p}}{\epsilon_1} \tag{A.116}$$

This is the so-called "Saxwell wave equation" of selectodynamics. If there is also a selecton, say, its sarge density can simply be added to the spoton one in the right hand side.

To check the Saxwell equation, first consider the case that the system is steady, i.e. independent of time. In that case the Saxwell wave equation becomes the Poisson equation of the previous subsection as it should. (The second term is summed over the three Cartesian directions i. That gives $\nabla^2\varphi$.) So the spoton produces the same steady Koulomb field (A.107) as before. So far, so good.

How about the force on a selecton in this field? Of course, the force on a selecton is a Sorentz force of the same form as (A.113),

$$F_{x_{\rm e}} = \frac{\partial\varphi(r_{{\rm e}_1}, r_{{\rm e}_2}, r_{{\rm e}_3}; t)}{\partial r_{{\rm e}_1}} s_{\rm e} \tag{A.117}$$

In the steady case, the relevant potential at the selecton is the electrostatic one (A.107) produced by the spoton as given in the previous subsection (Strictly speaking you should also include the field produced by the selecton itself. But this "self-interaction" produces no net force. That is fortunate because if the selecton was really a point sarge, the self-interaction is mathematically singular.) Now minus the potential (A.107) times the selecton sarge $s_{\rm e}$ gave the energy $V_{\rm ep}$ of the spoton-selecton interaction in the previous subsection. And minus the derivative of that gave the force on the selecton. A look at the force above then shows it is the same.

So in the steady case the Saxwell equation combined with the Sorentz force does reproduce selectostatics correctly. That means that the given Lagrangian (A.114) contains all of selectostatics in a single concise mathematical expression. At the minimum. Neat, isn't it?

Consider next the case that the time dependence cannot be ignored. Then the time derivative in the Saxwell equation (A.116) cannot be ignored. In that case the left hand side in the equation is the complete unsteady Klein-Gordon equation. Since there is a nonzero right-hand side, mathematically the Saxwell equation is an inhomogeneous Klein-Gordon equation. Now it is known from the theory of partial differential equations that the Klein-Gordon equation respects the speed of light. As an example, imagine that at time $t = 0$ you briefly shake the spoton at the origin and then put it back where it was. The right hand side of the Saxwell equation is then again back to what it was. But near the origin, the foton field φ will now contain additional disturbances. These disturbances evolve according to the homogeneous Saxwell equation, i.e. the equation with zero right hand side. And it is easy to check by substitution that the homogeneous equation has solutions of the form

$$\varphi = f(x - ct)$$

That are waves traveling in the x-direction with the speed of light c. The wave shape is the arbitrary function f and is preserved in time. And note that the x-direction is arbitrary. So waves like this can travel in any direction. The perturbations near the origin caused by shaking the spoton will consist of such waves. Since they travel with the speed of light, they need some time to reach the selecton. The selecton will not notice anything until this happens. However, when the perturbations in the foton field do reach the selecton, they will change the foton field φ at the selecton. That then will change the force (A.117) on the selecton.

It follows that selectodynamics, as described by the Lagrangian (A.114), also respects the speed of light limitation.

A.22.3 Quantum selectostatics

The previous subsections derived the Koulomb force between sarged particles. This force was due to foton exchange. While the derivations used some ideas from quantum mechanics, they were classical. The effect of the fotons took the form of a potential φ that the sarged particles interacted with. This potential was a classical field; it had a definite numerical value at each point. To be picky, there really was an undetermined constant in the potential φ. But its gradient $\nabla\varphi$ produced the fully determined Sorentz force per unit sarge (A.117). This force can be "observed" by a sarged spoton or selecton.

However, that very fact violates the fundamental postulates of quantum mechanics as formulated at the beginning of this book, chapter 3.4. Observable values should be the eigenvalues of Hermitian operators that act on wave functions. While the foton potential was loosely associated with a foton wave function, wave functions should not be observable.

Now if classically every position has its own observable local potential φ, then in a proper quantum description every position must be associated with its own Hermitian operator $\widehat{\varphi}$. In the terminology of addendum {A.15.9}, the foton field $\widehat{\varphi}$ must be a "quantum field;" an infinite amount of operators, one for each position.

The objective in this subsection is to deduce the form of this quantum field. And the type of wave function that it operates on. The results will then be used to verify the Koulomb force between stationary sarges as found the first subsection. It is imperative to figure out whether like sarges still attract in a proper quantum description.

Doing this directly would not be easy. It helps a lot if the field is written in terms of linear momentum eigenstates.

In fact, typical quantum field theories depend very heavily on this trick. However, often such theories use relativistic combined energy-momentum states in four-dimensional space-time. This subsection will use simpler purely spatial momentum states. The basic idea is the same. And it is essential for understanding the later Fermi derivation of the Coulomb potential.

Linear momentum states are complex exponentials of the form $e^{\mathrm{i}\vec{k}\cdot\vec{r}}$. Here $\vec{k}$ is a constant vector called the wave number vector. The momentum of such a state is given in terms of the wave number vector by the de Broglie relation as $\vec{p} = \hbar\vec{k}$. (It may be noted that the $e^{\mathrm{i}\vec{k}\cdot\vec{r}}$ states need an additional constant to properly normalize them, chapter 6.18. But for conciseness, in this addendum that normalization constant will be absorbed in the constants multiplying the exponentials.)

If a field φ is written in terms of linear momentum states, its value at any point $\vec{r}$ is given by:

$$\varphi(\vec{r}) = \sum_{\text{all } \vec{k}} c_{\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}}$$

Note that if you know the coefficients $c_{\vec{k}}$ of the momentum states, it is equivalent to knowing the field φ. Then the field at any point can in principle be found by doing the sum.

The expression above assumes that the entire system is confined to a very large periodic box, as in chapter 6.17. In infinite space the sum becomes an integral, section 7.9. That would be much more messy. (But that is the way you will usually find it in a typical quantum field analysis.) The precise values of the wave number vectors to sum over for a given periodic box were given in chapter 6.18 (6.28); they are all points in figure 6.17.

The first subsection found the selectostatic potential φ^{p} that was produced by a spoton, (A.107). This potential was a *classical* field; it had a definite numerical value for each position. The first step will be to see how this potential looks in terms of momentum states. While the final objective is to rederive the classical potential using proper quantum mechanics, the correct answer will need to be recognized when written in terms of momentum states. Not to mention that the answer will reappear in the discussion of the Coulomb potential. For simplicity it will be assumed that the spoton is at the origin.

According to the first subsection, the classical potential was the solution to a Poisson equation; a steady Klein-Gordon equation with forcing by the spoton:

$$-\nabla^2 \varphi^{\mathrm{p}}_{\mathrm{cl}} = \frac{s_{\mathrm{p}}}{\epsilon_1} \psi^*_{\mathrm{p}} \psi_{\mathrm{p}}$$

As a reminder that φ^{p} is a classical potential, not a quantum one, a subscript "cl" has been added. Also note that since this is a now a quantum description, the spoton sarge density σ_{p} has been identified as the spoton sarge s_{p} times the square magnitude of the spoton wave function $|\psi_{\mathrm{p}}|^2 = \psi^*_{\mathrm{p}}\psi_{\mathrm{p}}$.

Now the classical potential is to be written in the form

$$\varphi^{\mathrm{p}}_{\mathrm{cl}}(\vec{r}) = \sum_{\text{all } \vec{k}} c_{\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}}$$

To figure out the coefficients $c_{\vec{k}}$, plug it in the Poisson equation above. That gives

$$\sum_{\text{all } \vec{k}} k^2 c_{\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} = \frac{s_{\mathrm{p}}}{\epsilon_1} \psi^*_{\mathrm{p}} \psi_{\mathrm{p}}$$

Note that in the left hand side each ∇ produced a factor $\mathrm{i}\vec{k}$ for $-k^2$ total.

Now multiply this equation at both sides by some sample complex-conjugate momentum eigenfunction $e^{-\mathrm{i}\vec{\underline{k}}\cdot\vec{r}}$ and integrate over the entire volume $\mathcal{V}$ of the

periodic box. In the left hand side, you only get something nonzero for the term in the sum where $\vec{k} = \underline{\vec{k}}$ because eigenfunctions are orthogonal. For that term, the exponentials multiply to 1. So the result is

$$\underline{k}^2 c_{\underline{\vec{k}}} \mathcal{V} = \frac{s_{\rm p}}{\epsilon_1} \int e^{-{\rm i}\underline{\vec{k}}\cdot\vec{r}} \psi_{\rm p}^* \psi_{\rm p} \, {\rm d}^3\vec{r}$$

Now in the right hand side, assume again that the spoton is almost exactly at the origin. In other words, assume that its wave function is zero except very close to the origin. In that case, the exponential in the integral is approximately 1 when the spoton wave function is not zero. Also, the square wave function integrates to 1. So the result is, after clean up,

$$c_{\underline{\vec{k}}} = \frac{s_{\rm p}}{\epsilon_1 \mathcal{V} \underline{k}^2}$$

This expression applies for any wave number vector $\underline{\vec{k}}$, so you can leave the underline away. It fully determines $\varphi_{\rm cl}^{\rm p}$ in terms of the momentum states:

$$\varphi_{\rm cl}^{\rm p}(\vec{r}) = \sum_{\text{all } \vec{k}} \frac{s_{\rm p}}{\epsilon_1 \mathcal{V} k^2} e^{{\rm i}\vec{k}\cdot\vec{r}} \tag{A.118}$$

This solution is definitely one to remember. Note in particular that the coefficients of the momentum states are a constant divided by k^2. Recall also that for a unit value of $s_{\rm p}/\epsilon_1$, this solution is the fundamental solution, or Green's function, of the Poisson equation with point wise forcing at the origin.

If the requirement that the spoton wave function is completely at the origin is relaxed, the integral involving the spoton wave function stays:

$$\varphi_{\rm cl}^{\rm p}(\vec{r}) = \sum_{\text{all } \vec{k}} \frac{s_{\rm p}}{\epsilon_1 \mathcal{V} k^2} \langle \psi_{\rm p} | e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}} \psi_{\rm p} \rangle e^{{\rm i}\vec{k}\cdot\vec{r}} \tag{A.119}$$

where

$$\langle \psi_{\rm p} | e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}} \psi_{\rm p} \rangle \equiv \int \psi_{\rm p}^*(\vec{r}_{\rm p}) e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}} \psi_{\rm p}(\vec{r}_{\rm p}) \, {\rm d}^3\vec{r}_{\rm p}$$

Note that the integration variable over the spoton wave function has been renamed $\vec{r}_{\rm p}$ to avoid confusion with the position $\vec{r}$ at which the potential is evaluated. The above result is really better to work with in this subsection, since it does not suffer from some convergence issues that the Green's function solution has. And it is exact for a spoton wave function that is somewhat spread out.

Now the objective is to reproduce this classical result using a proper quantum field theory. And to find the force when a selecton is added to the system.

To do so, consider initially a system of fotons and a single spoton. The spoton will be treated as a nonrelativistic particle. Then its wave function $\psi_{\rm p}$ describes exactly one spoton. The spoton wave function will be treated as given.

Imagine something keeping the spoton in a ground state squeezed around the origin. Maxwell's demon would work. He has not been doing much anyway after he failed his thermo test.

Next the fotons. Their description will be done based upon linear momentum states. Such a state corresponds to a single-foton wave function of the form $e^{\mathrm{i}\vec{k}\cdot\vec{r}}$.

To keep it simple, for now only a single momentum state will be considered. In other words, only a single wave number vector $\vec{k}$ will be considered. But there might be multiple fotons in the state, or even uncertainty in the number of fotons.

Of course, at the end of the day the results must still be summed over all values of $\vec{k}$.

Some notations are needed now. A situation in which there are no fotons in the considered state will be indicated by the "Fock space ket" $|0\rangle$. If there is one foton in the state, it is indicated by $|1\rangle$, two by $|2\rangle$, etcetera. In the mathematics of quantum field theory, kets are taken to be orthonormal, {A.15}:

$$\langle i_1||i_2\rangle = \begin{cases} 1 \text{ if } i_1 = i_2 \\ 0 \text{ otherwise} \end{cases} \tag{A.120}$$

In words, the inner product of kets is 0 unless the numbers of fotons are equal. Then it is 1.

The ground state wave function for the combined spoton-fotons system is then assumed to be of the form

$$\psi_{\varphi\mathrm{p}} = C_0\psi_\mathrm{p}|0\rangle + C_1\psi_\mathrm{p}|1\rangle + C_2\psi_\mathrm{p}|2\rangle + \ldots \qquad |C_0|^2 + |C_1|^2 + |C_2|^2 + \ldots = 1 \tag{A.121}$$

That is a linear combination of system states with 0, 1, 2, ... fotons. So it is assumed that there may be uncertainty about the number of fotons in the considered foton state. The normalization condition for the constants expresses that the total probability of finding the system in some state or the other is 1.

(It may be noted that in typical quantum field theories, a charged relativistic particle would also be described in terms of kets and some quantum field $\widehat{\psi}$. However, unlike for a photon, for a charged particle $\widehat{\psi}$ would normally be a complex quantum field. Then $\widehat{\psi}^*\widehat{\psi}$ or something along these lines provides a real probability for a photon to "observe" the particle. That resembles the Born interpretation of the nonrelativistic wave function somewhat, especially for a spinless particle. Compare ⟦17, pp. 49, 136, 144⟧. The field $\widehat{\psi}$ will describe both the particle and its oppositely charged antiparticle. The spoton wave function ψ_p as used here represents some nonrelativistic limit in which the antiparticle has been approximated away from the field, ⟦17, pp. 41-45⟧. Such a nonrelativistic limit simply does not exist for a real scalar field like the Koulomb one.)

Now, of course, the Hamiltonian is needed. The Hamiltonian determines the energy. It consists of three parts:

$$H = H_\mathrm{p} + H_\varphi + H_{\varphi\mathrm{p}}$$

The first part is the Hamiltonian for the spoton in isolation. It consists of the kinetic energy of the spoton, as well as the potential provided by the fingers of the demon. By definition

$$H_{\rm p}\psi_{\rm p} = E_{\rm p}\psi_{\rm p} \tag{A.122}$$

where $E_{\rm p}$ is the energy of the spoton in isolation.

The second part is the Hamiltonian of the free foton field. Each foton in the considered state should have an energy $\hbar\omega$ with $\omega = kc$. That is the energy that you get if you substitute the momentum eigenfunction $e^{{\rm i}\vec{k}\cdot\vec{r}}$ into the Klein-Gordon eigenvalue problem (A.101) for a massless particle. And if one foton has an energy $\hbar\omega$, then i of them should have energy $i\hbar\omega$, so

$$H_\varphi|i\rangle = i\hbar\omega|i\rangle \tag{A.123}$$

Note that specifying what the Hamiltonian does to each separate ket tells you all you need to know about it. (Often there is an additional ground state energy shown in the above expression, but that does not make a difference here. It reflects the choice of the zero of energy.)

Finally, the third part of the total Hamiltonian is the interaction between the spoton and the foton field. This is the tricky one. First recall the classical expression for the interaction energy. According to the previous subsection, (A.111), it was $-s_{\rm p}\varphi_{\rm p}$. Here $\varphi_{\rm p}$ was the classical foton potential, evaluated at the position of the spoton.

In quantum field theory, the observable field φ gets replaced by a quantum field $\widehat{\varphi}$. The interaction Hamiltonian then becomes

$$H_{\varphi{\rm p}} = -s_{\rm p}\widehat{\varphi}_{\rm p} \tag{A.124}$$

This Hamiltonian needs to operate on the wave function (A.121) involving the spoton wave function and Fock space kets for the fotons. The big question is now: what is that quantum field $\widehat{\varphi}$?

To answer that, first note that sarged particles can create and destroy fotons. The above interaction Hamiltonian must express that somehow. After all, it is the Hamiltonian that determines the time evolution of systems in quantum mechanics.

Now in quantum field theories, creation and destruction of particles are accounted for through creation and annihilation operators, {A.15}. A creation operator $\widehat{a}_{\vec{k}}$ creates a single particle in a momentum state $e^{{\rm i}\vec{k}\cdot\vec{r}}$. An annihilation operator $\widehat{a}_{\vec{k}}$ annihilates a single particle from such a state. More precisely, the operators are defined as

$$\widehat{a}_{\vec{k}}|i\rangle = \sqrt{i}|i{-}1\rangle \qquad \widehat{a}^\dagger_{\vec{k}}|i{-}1\rangle = \sqrt{i}|i\rangle \tag{A.125}$$

Here $|i\rangle$ is the Fock-space ket that indicates that there are i fotons in the considered momentum state. Except for the numerical factor $\sqrt{i}$, the annihilation

operator takes a foton out of the state. The creation operator puts it back in, adding another numerical factor $\sqrt{i}$.

Note incidentally that the foton field Hamiltonian given earlier can now be rewritten as

$$H_\varphi = \hbar\omega \widehat{a}^\dagger_{\vec{k}} \widehat{a}_{\vec{k}} \tag{A.126}$$

That is because

$$\hbar\omega \widehat{a}^\dagger_{\vec{k}} \widehat{a}_{\vec{k}} |i\rangle = \hbar\omega \widehat{a}^\dagger_{\vec{k}} \sqrt{i} |i{-}1\rangle = \hbar\omega \sqrt{i}^2 |i\rangle = i\hbar\omega |i\rangle = H_\varphi |i\rangle$$

In general this Hamiltonian will still need to be summed over all values of $\vec{k}$.

But surely, the creation and annihilation of particles should also depend on where the spoton is. Fotons in the considered state have a spatially varying wave function. That should be reflected in the quantum field $\widehat{\varphi}$ somehow. To find the correct expression, it is easiest to first perform a suitable normalization of the foton state. Now the full wave function corresponding to the single-foton momentum eigenstate in empty space is

$$\varphi_{\rm f} = C e^{-{\rm i}\omega t} e^{{\rm i}\vec{k}\cdot\vec{r}}$$

Here C is some normalization constant to be chosen. The above wave function can be verified by putting it into the Klein-Gordon equation (A.102). The energy of the foton is given in terms of its wave function above as $\hbar\omega$. But the energy in the foton field is also related to the observable field φ; classical selectostatics gives that relation as (A.103). If you plug the foton wave function above into that classical expression, you do not normally get the correct energy $\hbar\omega$. There is no need for it; the foton wave function is not observable. However, it makes things simpler if you choose C so that the classical energy does equal $\hbar\omega$. That gives a energy-normalized wave function

$$\varphi_{\vec{k}} = \frac{\varepsilon_k}{k} e^{{\rm i}\vec{k}\cdot\vec{r}} \qquad \varepsilon_k \equiv \sqrt{\frac{\hbar\omega}{\epsilon_1 \mathcal{V}}} \tag{A.127}$$

In those terms, the needed quantum field turns out to be

$$\widehat{\varphi} = \frac{1}{\sqrt{2}} (\varphi_{\vec{k}} \widehat{a}_{\vec{k}} + \varphi^*_{\vec{k}} \widehat{a}^\dagger_{\vec{k}}) = \frac{\varepsilon_k}{\sqrt{2}k} (e^{{\rm i}\vec{k}\cdot\vec{r}} \widehat{a}_{\vec{k}} + e^{-{\rm i}\vec{k}\cdot\vec{r}} \widehat{a}^\dagger_{\vec{k}}) \qquad \varepsilon_k \equiv \sqrt{\frac{\hbar\omega}{\epsilon_1 \mathcal{V}}} \tag{A.128}$$

The first term in the right hand side is the normalized single-foton wave function at wave number $\vec{k}$ times the corresponding annihilation operator. The second term is the complex-conjugate foton wave function times the creation operator. There is also the usual factor $1/\sqrt{2}$ that appears when you take a linear combination of two states.

You might of course wonder about that second term. Mathematically it is needed to make the operator Hermitian. Recall that operators in quantum

mechanics need to be Hermitian to ensure that observable quantities have real, rather than complex values, chapter 2.6. To check whether an operator is Hermitian, you need to check that it is unchanged when you take it to the other side of an inner product. Now the wave function is a numerical quantity that changes into its complex conjugate when taken to the other side. And $\widehat{a}$ changes into $\widehat{a}^\dagger$ and vice-versa when taken to the other side, {A.15.2}. So each term in $\widehat{\varphi}$ changes into the other one, leaving the sum unchanged. So the operator as shown is indeed Hermitian.

But what to make physically of the two terms? One way of thinking about it is that the observed field is real because it does not just involve an interaction with an $e^{\mathrm{i}(\vec{k}\cdot\vec{r}-\omega t)}$ foton, but also with an $e^{-\mathrm{i}(\vec{k}\cdot\vec{r}-\omega t)}$ antifoton.

In general, the quantum field above would still need to be summed over all wave numbers $\vec{k}$. (Or integrated over $\vec{k}$ in infinite space). It may be noted that for given $\vec{r}$ the sum of the creation operator terms over all $\vec{k}$ can be understood as a field operator that creates a particle at position $\vec{r}$, [35, p. 24]. That is a slightly different definition of the creation field operator than given in {A.15.9}, [43, pp. 22]. But for nonrelativistic particles (which have nonzero rest mass) it would not make a difference.

With the quantum field φ now identified, the Hamiltonian of the spoton-fotons interaction becomes finally

$$H_{\varphi\mathrm{p}} = -s_\mathrm{p}\widehat{\varphi}_\mathrm{p} = -\frac{s_\mathrm{p}\varepsilon_k}{\sqrt{2}k}(e^{\mathrm{i}\vec{k}\cdot\vec{r}_\mathrm{p}}\widehat{a}_{\vec{k}} + e^{-\mathrm{i}\vec{k}\cdot\vec{r}_\mathrm{p}}\widehat{a}^\dagger_{\vec{k}}) \qquad \varepsilon_k \equiv \sqrt{\frac{\hbar\omega}{\epsilon_1\mathcal{V}}} \tag{A.129}$$

Note that the spoton has uncertainty in position. The spoton position in the Hamiltonian above is just a possible spoton position. In usage it will still get multiplied by the square spoton wave function magnitude that gives the probability for that position. Still, at face value the interaction of the spoton with the field takes place at the location of the spoton. Interactions in quantum field theories are "local." At least on macroscopic scales that is needed to satisfy the limitation of the speed of light.

Having a Hamiltonian allows quantum selectodynamics to be explored. That will be done to some detail for the case of the electromagnetic field in subsequent addenda. However, here the only question that will be addressed is whether classical selectostatics as described in the first subsection was correct. In particular, do equal sarges still attract in the quantum description?

Selectostatics of the spoton-fotons system should correspond to the ground state of the system. The ground state has the lowest possible energy. You can therefore find the ground state by finding the state of lowest possible system energy. That is the same trick as was used to find the ground states of the hydrogen molecular ion and the hydrogen molecule in chapters 4.6 and 5.2. The "expectation value" of the system energy is defined by the inner product

$$\langle E\rangle = \langle\psi_{\varphi\mathrm{p}}|(H_\mathrm{p} + H_\varphi + H_{\varphi\mathrm{p}})\psi_{\varphi\mathrm{p}}\rangle$$

Here the Hamiltonians have already been described above.

Now to find the ground state, the lowest possible value of the expectation energy above is needed. To get that, the inner products between the kets in the factors $\psi_{\varphi\mathrm{p}}$ must be multiplied out. First apply the Hamiltonians (A.122), (A.123), and (A.129) on the wave function $\psi_{\varphi\mathrm{p}}$, (A.121), using (A.125). Then apply the orthogonality relations (A.120) of kets. Do not forget the complex conjugate on the left side of an inner product. That produces

$$\begin{aligned}\langle E\rangle &= E_{\mathrm{p}} \\ &\quad + |C_1|^2\hbar\omega + |C_2|^2 2\hbar\omega + \ldots \\ &\quad - \frac{s_{\mathrm{p}}\varepsilon_k}{\sqrt{2}k}\langle\psi_{\mathrm{p}}|e^{\mathrm{i}\vec{k}\cdot\vec{r}_{\mathrm{p}}}\psi_{\mathrm{p}}\rangle\left(C_0^*C_1 + \sqrt{2}C_1^*C_2 + \ldots\right) \\ &\quad - \frac{s_{\mathrm{p}}\varepsilon_k}{\sqrt{2}k}\langle\psi_{\mathrm{p}}|e^{-\mathrm{i}\vec{k}\cdot\vec{r}_{\mathrm{p}}}\psi_{\mathrm{p}}\rangle\left(C_1^*C_0 + \sqrt{2}C_2^*C_1 + \ldots\right)\end{aligned}$$

Here the dots stand for terms involving the coefficients $C_3, C_4, \ldots$ of states with three or more fotons.

Note that the first term in the right hand side above is the energy E_{p} of the spoton by itself. That term is a given constant. The question is what foton states produce the lowest energy for the remaining terms. The answer is easy if the spoton sarge s_{p} is zero. Then the terms in the last two lines are zero. So the second line shows that $C_1, C_2, \ldots$ must all be zero. Then there are no fotons; only the state with zero fotons is then left in the system wave function (A.121).

If the spoton sarge is nonzero however, the interaction terms in the last two lines can lower the energy for suitable nonzero values of the constants C_1, C_2, ... To simplify matters, it will be assumed that the spoton sarge is nonzero but small. Then so will be the constants. In that case only the C_1 terms need to be considered; the other terms in the last two lines involve the product of two small constants, and those cannot compete. Further the normalization condition in (A.121) shows that $|C_0|$ will be approximately 1 since even C_1 is small. Then C_0 may be assumed to be 1, because any eigenfunction is indeterminate by a factor of magnitude 1 anyway.

Further, since any complex number may always be written as its magnitude times some exponential of magnitude 1, the second last line of the energy above can be written as

$$-\frac{s_{\mathrm{p}}\varepsilon_k}{\sqrt{2}k}\langle\psi_{\mathrm{p}}|e^{\mathrm{i}\vec{k}\cdot\vec{r}_{\mathrm{p}}}\psi_{\mathrm{p}}\rangle\, C_1 = -\frac{s_{\mathrm{p}}\varepsilon_k}{\sqrt{2}k}\left|\langle\psi_{\mathrm{p}}|e^{\mathrm{i}\vec{k}\cdot\vec{r}_{\mathrm{p}}}\psi_{\mathrm{p}}\rangle\right| e^{\mathrm{i}\alpha}\,|C_1|e^{\mathrm{i}\beta}$$

Replacing i everywhere by $-\mathrm{i}$ gives the corresponding expression for the last line. The complete expression for the energy then becomes

$$\begin{aligned}E &= E_{\mathrm{p}} + |C_1|^2\hbar\omega \\ &\quad - \frac{s_{\mathrm{p}}\varepsilon_k}{\sqrt{2}k}\left|\langle\psi_{\mathrm{p}}|e^{\mathrm{i}\vec{k}\cdot\vec{r}_{\mathrm{p}}}\psi_{\mathrm{p}}\rangle\right||C_1|e^{\mathrm{i}(\alpha+\beta)} - \frac{s_{\mathrm{p}}\varepsilon_k}{\sqrt{2}k}\left|\langle\psi_{\mathrm{p}}|e^{\mathrm{i}\vec{k}\cdot\vec{r}_{\mathrm{p}}}\psi_{\mathrm{p}}\rangle\right||C_1|e^{-\mathrm{i}(\alpha+\beta)}\end{aligned}$$

In the last term, note that the sign of i inside an absolute value does not make a difference. Using the Euler formula (2.5) on the trailing exponentials gives

$$E = E_{\rm p} + |C_1|^2\hbar\omega - 2\frac{s_{\rm p}\varepsilon_k}{\sqrt{2}k}\Big|\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}\psi_{\rm p}\rangle\Big||C_1|\cos(\alpha+\beta)$$

But in the ground state, the energy should be minimal. Clearly, that requires that the cosine is at its maximum value 1. So it requires taking C_1 so that $\beta = -\alpha$.

That still leaves the magnitude $|C_1|$ to be found. Note that the final terms in expression above are now of the generic form

$$a|C_1|^2 - 2b|C_1|$$

where a and b are positive constants. That is a quadratic function of $|C_1|$. By differentiation, it is seen that the minimum occurs at $|C_1| = b/a$ and has a value $-b^2/a$. Putting in what a and b are then gives

$$C_1 = \frac{s_{\rm p}\varepsilon_k}{\sqrt{2}k\hbar\omega}\langle\psi_{\rm p}|e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}\psi_{\rm p}\rangle \qquad E = E_{\rm p} - \frac{s_{\rm p}^2}{2\epsilon_1\mathcal{V}k^2}|\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}\psi_{\rm p}\rangle|^2$$

The second part of the energy is the energy lowering achieved by having a small probability $|C_1|^2$ of a single foton in the considered momentum state.

This energy-lowering still needs to be summed over all states $\vec{k}$ to get the total:

$$E - E_{\rm p} = -\sum_{\vec{k}}\frac{s_{\rm p}^2}{2\epsilon_1\mathcal{V}k^2}|\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}\psi_{\rm p}\rangle|^2 = -\sum_{\vec{k}}\frac{s_{\rm p}^2}{2\epsilon_1\mathcal{V}k^2}\langle\psi_{\rm p}|e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}\psi_{\rm p}\rangle\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}}\psi_{\rm p}\rangle$$

Note that the final two inner products represent separate integrations. Therefore to avoid confusion, the subscript p was dropped from one integration variable. In the sum, the classical field (A.119) can now be recognized:

$$E - E_{\rm p} = -\frac{s_{\rm p}}{2}\langle\psi_{\rm p}|\varphi_{\rm cl}^{\rm p}(\vec{r})\psi_{\rm p}\rangle = -\tfrac{1}{2}\int\varphi_{\rm cl}^{\rm p}(\vec{r})s_{\rm p}\psi_{\rm p}^*(\vec{r})\psi_{\rm p}(\vec{r})\,{\rm d}^3\vec{r}$$

Ignoring the differences in notation, the energy lowering is exactly the same as (A.108) found in the classical analysis. The classical analysis, while not really justified, did give the right answer.

However now an actual picture of the quantum ground state has been obtained. It is a quantum superposition of system states. The most likely state is the one where there are no fotons at all. But there are also small probabilities for system states where there is a single foton in a single linear momentum foton state. This picture does assume that the spoton sarge is small. If that was not true, things would get much more difficult.

Another question is whether the observable values of the foton potential are the same as those obtained in the classical analysis. This is actually a trick question because even the classical foton potential is not observable. There is still an undetermined constant in it. What is observable are the *derivatives* of the potential: they give the observable selectic force per unit sarge on sarged particles.

Now, in terms of momentum modes, the derivatives of the classical potential can be found by differentiating (A.119). That gives

$$\varphi_{i_{\rm p,cl}} = \sum_{\text{all } \vec{k}} \frac{s_{\rm p}}{\epsilon_1 \mathcal{V} k^2} \langle \psi_{\rm p} | e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}} \psi_{\rm p} \rangle {\rm i} k_i e^{{\rm i}\vec{k}\cdot\vec{r}}$$

Recall again the convention introduced in the previous subsection that a subscript i on φ indicates the derivative $\partial/\partial r_i$, where r_1, r_2, and r_3 correspond to x, y, and z respectively. So the above expression gives the selectic force per unit sarge in the x, y, or z direction, depending on whether i is 1, 2, or 3.

The question is now whether the quantum analysis predicts the same observable forces. Unfortunately, the answer here is no. The observable forces have quantum uncertainty that the classical analysis missed. However, the Ehrenfest theorem of chapter 7.2.1 suggests that the expectation forces should still match the classical ones above.

The quantum expectation force per unit sarge in the i-direction is given by

$$\langle \varphi_i \rangle = \langle \Psi_{\varphi{\rm p}} | \widehat{\varphi}_i \Psi_{\varphi{\rm p}} \rangle \qquad \Psi_{\varphi{\rm p}} = e^{-{\rm i}Et/\hbar} \psi_{\varphi{\rm p}}$$

Here E is the ground state energy. Note that in this case the full, time-dependent wave function $\Psi_{\varphi{\rm p}}$ is used. That is done because in principle an observed field could vary in time as well as in space. Substituting in the r_i-derivative of the quantum field (A.128) gives

$$\langle \varphi_i \rangle = \frac{\varepsilon_k}{\sqrt{2}k} \langle \psi_{\varphi{\rm p}} | ({\rm i}k_i e^{{\rm i}\vec{k}\cdot\vec{r}} \widehat{a}_{\vec{k}} - {\rm i}k_i e^{-{\rm i}\vec{k}\cdot\vec{r}} \widehat{a}^\dagger_{\vec{k}}) \psi_{\varphi{\rm p}} \rangle$$

Note that here $\vec{r}$ is not a possible position of the spoton, but a given position at which the selectic force per unit sarge is to be found. Also note that the time dependent exponentials have dropped out against each other; the expectation forces are steady like for the classical field.

The above expression can be multiplied out as before. Using the obtained expression for C_1, and the fact that $\langle \psi_{\rm p} | \psi_{\rm p} \rangle = 1$ because wave functions are normalized, that gives.

$$\langle \varphi_i \rangle = \frac{s_{\rm p}}{2\epsilon_1 \mathcal{V} k^2} \langle \psi_{\rm p} | e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}} \psi_{\rm p} \rangle {\rm i} k_i e^{{\rm i}\vec{k}\cdot\vec{r}} - \frac{s_{\rm p}}{2\epsilon_1 \mathcal{V} k^2} \langle \psi_{\rm p} | e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}} \psi_{\rm p} \rangle {\rm i} k_i e^{-{\rm i}\vec{k}\cdot\vec{r}}$$

Summed over all $\vec{k}$, the two terms in the right hand side produce the same result, because opposite values of $\vec{k}$ appear equally in the summation. In other words,

for every $\vec{k}$ term in the first sum, there is a $-\vec{k}$ term in the second sum that produces the same value. And that then means that the expectation selectic forces are the same as the classical ones. The classical analysis got that right, too.

To see that there really is quantum uncertainty in the forces, it suffices to look at the expectation *square* forces. If there was no uncertainty in the forces, the expectation square forces would be just the square of the expectation forces. To see that that is not true, it is sufficient to simply take the spoton sarge zero. Then the expectation field is zero too. But the expectation square field is given by

$$\left\langle \varphi_i^2 \right\rangle = \frac{\varepsilon_k^2}{2k^2} \langle \psi_{\varphi \mathrm{p}} | (\mathrm{i}k_i e^{\mathrm{i}\vec{k}\cdot\vec{r}} \widehat{a}_{\vec{k}} - \mathrm{i}k_i e^{-\mathrm{i}\vec{k}\cdot\vec{r}} \widehat{a}_{\vec{k}}^\dagger)^2 \psi_{\varphi \mathrm{p}} \rangle \qquad \psi_{\varphi \mathrm{p}} = \psi_{\mathrm{p}} |0\rangle$$

Multiplying this out gives

$$\left\langle \varphi_i^2 \right\rangle = \frac{\varepsilon_k^2 k_i^2}{2k^2} = \frac{\hbar\omega k_i^2}{2\epsilon_1 \mathcal{V} k^2}$$

Since Planck's constant is not zero, this is not zero either. So even without the spoton, a selectic force measurement will give a random, but nonzero value. The average of a large number of such force measurements will be zero, but not the individual measurements.

The above expression can be compared with the corresponding $|\varphi_{\vec{k},i}|^2$ of a single foton, as given by (A.127). That comparison indicates that even in the ground state in empty space, there is still half a foton of random field energy left. Recall now the Hamiltonian (A.126) for the foton field. Usually, this Hamiltonian would be defined as

$$H_\varphi = \sum_{\vec{k}} \hbar\omega (\widehat{a}_{\vec{k}}^\dagger \widehat{a}_{\vec{k}} + \tfrac{1}{2})$$

The additional $\frac{1}{2}$ expresses the half foton of energy left in the ground state. The ground state energy does not change the dynamics. However, it is physically reflected in random nonzero values if the selectic field is measured in vacuum.

The bad news is that if you sum these ground state energies over all values of $\vec{k}$, you get infinite energy. The exact same thing happens for the photons of the electromagnetic field. Quantum field theories are plagued by infinite results; this "vacuum energy" is just a simple example. What it really means physically is as yet not known either. More on this can be found in {A.23.4}.

The final issue to be addressed is the attraction between a spoton and a selecton. That can be answered by simply adding the selecton to the spoton-fotons analysis above, {D.37.2}. The answer is that the spoton-selecton interaction energy is the same as found in the classical analysis.

So equal sarges still attract.

A.22.4 Poincaré and Einstein try to save the universe

The Koulomb universe is a grim place. In selectodynamics, particles with the same sarge attract. So all selectons clump together into one gigantic ball. Assuming that spotons have the opposite sarge, they clump together into another big ball at the other end of the universe. But actually there is no justification to assume that spotons would have a different sarge from selectons. That then means that all matter clumps together into a single gigantic satom. A satom like that will form one gigantic, obscene, black hole. It is hardly conductive to the development of life as we know it.

Unfortunately, the Koulomb force is based on highly plausible, apparently pretty unavoidable assumptions. The resulting force simply makes sense. None of these things can be said about the Coulomb force.

But maybe, just maybe, the Koulomb juggernaut can be tripped up by some legal technicality. Things like that have happened before.

Now in a time not really that very long ago, there lived a revolutionary of mathematics called Poincaré. Poincaré dreamt of countless shining stars that would sweep through a gigantic, otherwise dark universe. And around these stars there would be planets populated by living beings called "observers." But if the stars all moved in random directions, with random speeds, then which star would be the one at rest? Which star would be the king around which the other stars danced? Poincaré thought long and hard about that problem. "No!" he thundered eventually; "It shall not be. I hereby proclaim that all stars are created equal. Any observer at any star can say at any given time that its star is at rest and that the other stars are moving. On penalty of dead, nothing in physics may indicate that observer to be wrong."

Now nearby lived a young physicist called Einstein who was very lazy. For example, he almost never bothered to write the proper summation symbols in his formulae. Of course, that made it difficult for him to find a well paying job in some laboratory where they smash spotons into each other. Einstein ended up working in some patent office for little pay. But, fortunate for our story, working in a patent office did give Einstein a fine insight in legal technicalities.

First Einstein noted that the Proclamation of Poincaré meant that observers at different stars had to disagree seriously about the locations and times of events. However, it would not be complete chaos. The locations and times of events as perceived by different observers would still be related. The relation would be a transformation that the famous physicist Lorentz had written down earlier, chapter 1.2.1 (1.6).

And the Proclamation of Poincaré also implied that different observers had to agree about the same laws of physics. So the laws of physics should remain the same when you change them from one observer to the next using the Lorentz transformation. Nowadays we would say that the laws of physics should be "Lorentz invariant." But at the time, Einstein did not want to use the name of

Lorentz in vain.

Recall now the classical "action principle" of subsection A.22.2. The so-called action integral had to be unchanged under small deviations from the correct physics. The Proclamation of Poincaré demands that all observers must agree that the action is unchanged. If the action is unchanged for an observer at one star, but not for one at another star, then not all stars are created equal.

To see what that means requires a few fundamental facts about special relativity, the theory of systems in relative motion.

The Lorentz transformation badly mixes up spatial positions and times of events as seen by different observers. To deal with that efficiently, it is convenient to combine the three spatial coordinates and time into a single four-dimensional vector, a four-vector, chapter 1.2.4. Time becomes the "zeroth coordinate" that joins the three spatial coordinates. In various notations, the four-vector looks like

$$\vec{\vec{r}} \equiv \begin{pmatrix} ct \\ \\ \vec{r} \end{pmatrix} \equiv \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ r_3 \end{pmatrix} \equiv \begin{pmatrix} r_0 \\ \\ \{r_i\} \end{pmatrix} \equiv \begin{pmatrix} x^0 \\ x^1 \\ x^2 \\ x^3 \end{pmatrix} \equiv \{x^\mu\} \rightarrow x^\mu$$

First of all, note that the zeroth coordinate receives an additional factor c, the speed of light. That is to ensure that it has units of length just like the other components. It has already been noted before that the spatial coordinates x, y, and z are indicated by r_1, r_2, and r_3 in this addendum. That allows a generic component to be indicated by r_i for $i = 1$, 2, or 3. Note also that curly brackets are a standard mathematical way of indicating a set or collection. So $\{r_i\}$ stands for the set of all three r_i values; in other words, it stands for the complete position vector $\vec{r}$. That is the primary notation that will be used in this addendum.

However, in virtually any quantum field book, you will find four-vectors indicated by x^μ. Here μ is an index that can have the values 0, 1, 2, or 3. (Except that some books make time the fourth component instead of the zeroth.) An x^μ by itself probably really means $\{x^\mu\}$, in other words, the complete four-vector. Physicists have trouble typing curly brackets, so they leave them away. When more than one index is needed, another Greek symbol will be used, like x^ν. However, x^i would stand for just the spatial components, so for the position vector $\{r_i\}$. The give-away is here that a roman superscript is used. Roman superscript j would mean the same thing as i; the spatial components only.

There are similar notations for the derivatives of a function f:

$$\vec{\vec{\nabla}} f \equiv \begin{pmatrix} \partial f/c\partial t \\ \\ \nabla f \end{pmatrix} \equiv \begin{pmatrix} \partial f/c\partial t \\ \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{pmatrix} \equiv \begin{pmatrix} f_t/c \\ \\ \{f_i\} \end{pmatrix} \equiv \begin{pmatrix} \partial_0 f \\ \partial_1 f \\ \partial_2 f \\ \partial_3 f \end{pmatrix} \equiv \{\partial_\mu f\} \rightarrow \partial_\mu f$$

Note again that time derivatives in this addendum are indicated by a subscript t and spatial derivatives by a subscript i for $i = 1$, 2, or 3.

Quantum field books use $\partial_\mu f$ for derivatives. They still have problems with typing curly brackets, so $\partial_\mu f$ by itself probably means the set of all four derivatives. Similarly $\partial_i f$ would probably mean the spatial gradient ∇f.

The final key fact to remember about special relativity is:

> *In dot products between four-vectors, the product of the zeroth components gets a minus sign.*

Dot products between four-vectors are very important because all observers agree about the values of these dot products. They are Lorentz invariant. (In nonrelativistic mechanics, all observers agree about the usual dot products between spatial vectors. That is no longer true at relativistic speeds.)

One warning. In almost all modern quantum field books, the products of the *spatial* components get the minus sign instead of the time components. The purpose is to make the relativistic dot product incompatible with the nonrelativistic one. After all, "backward compatibility" is so, well, backward. (One source that does use the compatible dot product is [49]. This is a truly excellent book written by a Nobel Prize winning pioneer in quantum field theory. It may well be the best book on the subject available. Unfortunately it is also very mathematical and the entire thing spans three volumes. Then again, you could certainly live without supersymmetry.)

One other convention might be mentioned. Some books put a factor $\mathrm{i} = \sqrt{-1}$ in the zeroth components of four-vectors. That takes care of the minus sign in dot products automatically. But modern quantum field books do not this.

Armed with this knowledge about special relativity, the Koulomb force can now be checked. Action is defined as

$$\mathcal{S} \equiv \int_{t_1}^{t_2} \mathcal{L} \, \mathrm{d}t$$

Here the time range from t_1 to t_2 should be chosen to include the times of interest. Further $\mathcal{L}$ is the so-called Lagrangian.

If all observers agree about the value of the action in selectodynamics, then selectodynamics is Lorentz invariant. Now the Lagrangian of classical selectodynamics was of the form, subsection A.22.2,

$$\mathcal{L}_{\varphi \mathrm{p}} = \int \pounds_\varphi \, \mathrm{d}^3\vec{r} + \tfrac{1}{2} m_\mathrm{p} \vec{v}_\mathrm{p}^2 + \varphi_\mathrm{p} s_\mathrm{p}$$

Here the Lagrangian density of the foton field φ was

$$\pounds_\varphi = -\frac{\epsilon_1}{2} \left(-\frac{1}{c^2} \varphi_t^2 + \varphi_i^2 \right)$$

To this very day, a summation symbol may *not* be used to reveal to nonphysicists that the last term needs to be summed over all three values of i. That is in honor of the lazy young physicist, who tried to save the universe.

Note that the parenthetical term in the Lagrangian density is simply the square of the four-vector of derivatives of φ. Indeed, the relativistic dot product puts the minus sign in front of the product of the time derivatives. Since all observers agree about dot products, they all agree about the values of the Lagrangian density. It is Lorentz invariant.

To be sure, it is the action and not the Lagrangian density that must be Lorentz invariant. But note that in the action, the Lagrangian density gets integrated over both space and time. Such integrals are the same for any two observers. You can easily check that from the Lorentz transformation chapter 1.2.1 (1.6) by computing the Jacobian of the $\mathrm{d}x\mathrm{d}t$ integration between observers.

(OK, the limits of integration are not really the same for different observers. One simple way to get around that is to assume that the field vanishes at large negative and positive times. Then you can integrate over all space-time. A more sophisticated argument can be given based on the derivation of the action principle {A.1.5}. From that derivation it can be seen that it suffices to consider small deviations from the correct physics that are localized in both space and time. It implies that the limits of integration in the action integral are physically irrelevant.)

(Note that this subsection does no longer mention periodic boxes. In relativity periodicity is not independent of the observer, so the current arguments really need to be done in infinite space.)

The bottom line is that the first, integral, term in the Lagrangian produces a Lorentz-invariant action. The second term in the Lagrangian is the *non*relativistic kinetic energy of the spoton. Obviously the action produced by this term will not be Lorentz invariant. But you can easily fix that up by substituting the corresponding relativistic term as given in chapter 1.3.2. So the lack of Lorentz invariance of this term will simply be ignored in this addendum. If you want, you can consider the spoton mass to be the moving mass in the resulting equations of motion.

The final term in the Lagrangian is the problem. It represents the spoton-fotons interaction. The term by itself would be Lorentz invariant, but it gets integrated with respect to time. Now in relativity time intervals $\mathrm{d}t$ are *not* the same for different observers. So the action for this term is not Lorentz invariant. Selectodynamics cannot be correct. The Koulomb juggernaut has been stopped by a small legal technicality.

(To be sure, any good lawyer would have pointed out that there is no problem if the spoton sarge density, instead of the spoton sarge s_{p}, is the same for different observers. But the Koulomb force was so sure about its invincibility that it never bothered to seek competent legal counsel.)

The question is now of course how to fix this up. That will hopefully produce

a more appealing universe. One in which particles like protons and electrons have charges q rather than sarges s. Where these charges allow them to interact with the photons of the electromagnetic field. And where these photons assure that particles with like charges repel, rather than attract.

Consider the form of the problem term in the Koulomb action:

$$\int_{t_1}^{t_2} \varphi_{\mathrm{p}} s_{\mathrm{p}} \, \mathrm{d}t$$

It seems logical to try to write this in relativistic terms, like

$$\int_{t_1}^{t_2} \left(\frac{\varphi_{\mathrm{p}}}{c}\right)\left(s_{\mathrm{p}} \, \mathrm{d}ct\right) = \int_{t_1}^{t_2} \left(\frac{\varphi_{\mathrm{p}}}{c}\right)\left(s_{\mathrm{p}} \, \mathrm{d}r_{\mathrm{p}_0}\right)$$

Here $\mathrm{d}r_{\mathrm{p}_0}$ is the zeroth component of the change in spoton four-vector $\mathrm{d}\overrightarrow{r}_{\mathrm{p}}$. The product of the two parenthetical factors is definitely not Lorentz invariant. But suppose that you turn each of the factors into a complete four-vector? Dot products are Lorentz invariant. And the four-vector corresponding to $\mathrm{d}r_{\mathrm{p}_0}$ is clearly $\mathrm{d}\overrightarrow{r}_{\mathrm{p}}$.

But the photon potential must also become a four-vector, instead of a scalar. That is what it takes to achieve Lorentz invariance. So electrodynamics defines a four-vector of potentials of the form

$$\overrightarrow{A} \equiv \begin{pmatrix} \varphi/c \\ \\ \vec{A} \end{pmatrix} \equiv \begin{pmatrix} \varphi/c \\ A_x \\ A_y \\ A_z \end{pmatrix} \equiv \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix} \equiv \begin{pmatrix} A^0 \\ A^1 \\ A^2 \\ A^3 \end{pmatrix} \equiv \{A^\mu\} \rightarrow A^\mu$$

Here $\vec{A}$ is the so-called "vector potential" while φ is now the electrostatic potential.

The interaction term in the action now becomes, replacing the spoton sarge s_{p} by minus the proton charge q_{p},

$$\int_{t_1}^{t_2} \overrightarrow{A}_{\mathrm{p}} \cdot q_{\mathrm{p}} \, \mathrm{d}\overrightarrow{r}_{\mathrm{p}} = \int_{t_1}^{t_2} \overrightarrow{A}_{\mathrm{p}} \cdot q_{\mathrm{p}} \frac{\mathrm{d}\overrightarrow{r}_{\mathrm{p}}}{\mathrm{d}t} \, \mathrm{d}t$$

In writing out the dot product, note that the spatial components of $\mathrm{d}\overrightarrow{r}_{\mathrm{p}}/\mathrm{d}t$ are simply the proton velocity components v_{p_j}. That gives the interaction term in the action as

$$\int_{t_1}^{t_2} \left(-\varphi_{\mathrm{p}} q_{\mathrm{p}} + A_{j_{\mathrm{p}}} q_{\mathrm{p}} v_{\mathrm{p}_j}\right) \, \mathrm{d}t$$

Once again nonphysicists may not be told that the second term in parentheses must be summed over all three values of j since j appears twice.

The integrand above is the interaction term in the electromagnetic Lagrangian,

$$\mathcal{L}_{\rm int} = -\varphi_{\rm p} q_{\rm p} + A_{j_{\rm p}} q_{\rm p} v_{{\rm p}_j}$$

For now at least.

The Lagrangian density of the photon field is also needed. Since the photon field is a four-vector rather than a scalar, the self-evident electromagnetic density is

$$\pounds_{\rm seem} = -\frac{\epsilon_0}{2}\left(-A_{j_t}^{\ 2} + c^2 A_{j_i}^{\ 2} + \frac{1}{c^2}\varphi_t^2 - \varphi_i^2\right)$$

Here the constant ϵ_0 is called the "permittivity of space." Note that the second term in parentheses must be summed over both i and j. The curious sign pattern for the parenthetical terms arises because it involves two dot products: one from the square four-gradient (derivatives), and one from the square four-potential. Simply put, having electrostatic potentials is worth a minus sign, and having time derivatives is too.

It might be noted that in principle the proper Lagrangian density could be minus the above expression. But a minus sign in a Lagrangian does not change the motion. The convention is to choose the sign so that the corresponding Hamiltonian describes energies that can be increased by arbitrarily large amounts, not lowered by arbitrarily large amounts. Particles can have unlimited amounts of positive kinetic energy, not negative kinetic energy.

Still, it does seem worrisome that the proper sign of the Lagrangian density is not self-evident. But that issue will have to wait until the next subsection.

Collecting things together, the self-evident Lagrangian for electromagnetic field plus proton is

$$\mathcal{L}_{\rm seem+p} = \int \pounds_{\rm seem}\, {\rm d}^3\vec{r} + \tfrac{1}{2} m_{\rm p} v_{{\rm p}_j}^{\ 2} - \varphi_{\rm p} q_{\rm p} + A_{j_{\rm p}} q_{\rm p} v_{{\rm p}_j}$$

Here $\pounds_{\rm seem}$ was given above.

The first thing to check now is the equation of motion for the proton. Following subsection A.22.2, it can be found from

$$\frac{\rm d}{{\rm d}t}\left(\frac{\partial \mathcal{L}}{\partial v_{{\rm p}_i}}\right) = \left(\frac{\partial \mathcal{L}}{\partial r_{{\rm p}_i}}\right)$$

Substituting in the Lagrangian above gives

$$\frac{\rm d}{{\rm d}t}\left(m_{\rm p} v_{{\rm p}_i} + A_{i_{\rm p}} q_{\rm p}\right) = -\varphi_{i_{\rm p}} q_{\rm p} + A_{j_{i_{\rm p}}} q_{\rm p} v_{{\rm p}_j}$$

This can be cleaned up, {D.6}. In short, first an "electric" field $\vec{\mathcal{E}}$ and a "magnetic" field $\vec{\mathcal{B}}$ are defined as, in vector notation,

$$\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial \vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla \times \vec{A} \tag{A.130}$$

The individual components are

$$\mathcal{E}_i = -\varphi_i - A_{i_t} \qquad \mathcal{B}_i = A_{\bar{i}_{\bar{i}}} - A_{\bar{i}_{\bar{\bar{i}}}} \tag{A.131}$$

Here $i = 1$, 2, or 3 corresponds to the x, y, or z components respectively. Also $\bar{i}$ follows i in the periodic sequence $\ldots 123123 \ldots$ and $\bar{\bar{i}}$ precedes i. In these terms, the simplified equation of motion of the proton becomes, in vector notation,

$$\frac{\mathrm{d}m_{\mathrm{p}}\vec{v}_{\mathrm{p}}}{\mathrm{d}t} = q_{\mathrm{p}} \left(\vec{\mathcal{E}}_{\mathrm{p}} + \vec{v}_{\mathrm{p}} \times \vec{\mathcal{B}}_{\mathrm{p}} \right) \tag{A.132}$$

The left hand side is mass times acceleration. Relativistically speaking, the mass should really be the moving mass here, but OK. The right hand side is known as the "Lorentz force."

Note that there are 4 potentials with 4 derivatives each, for a total of 16 derivatives. But matter does not observe all 16 individually. Only the 3 components of the electric field and the 3 of the magnetic field are actually observed. That suggests that there may be changes to the fields that can be made that are not observable. Such changes are called "gage (or gauge) changes." The name arises from the fact that a gage is a measuring device. You and I would then of course say that these changes should be called *non*gage changes. They are not measurable. But "gage" is really shorthand for "Take that, you stupid gage."

Consider the most general form of such gage changes. Given potentials φ and $\vec{A}$, equivalent potentials can be created as

$$\varphi' = \varphi - \chi_t \qquad \vec{A}' = \vec{A} + \nabla\chi$$

Here χ can be any function of space and time that you want.

The potentials φ' and $\vec{A}'$ give the exact same electric and magnetic fields as φ and $\vec{A}$. (These claims are easily checked using a bit of vector calculus. Use Stokes to show that they are the most general changes possible.)

The fact that you can make unmeasurable changes to the potentials like that is called the "gage" (or gauge) property of the electromagnetic field. Nonphysicists probably think it is something you read off from a voltage gage. Hilarious, isn't it?

Observable or not, the evolution equations of the four potentials are also needed. To find them it is convenient to spread the proton charge out a bit. That is the same trick as was used in subsection A.22.2. For the spread-out charge, a "charge density" ρ_{p} can be defined as the charge per unit volume. It is also convenient to define a "current density" $\vec{\jmath}_{\mathrm{p}}$ as the charge density times its velocity. Then the proton-photons interaction terms in the Lagrangian are:

$$\int \Big(-\varphi\rho_{\mathrm{p}} + A_j \jmath_{\mathrm{p}_j} \Big)\, \mathrm{d}^3\vec{r} \approx -\varphi_{\mathrm{p}} q_{\mathrm{p}} + A_{j_{\mathrm{p}}} q_{\mathrm{p}} v_{\mathrm{p}_j} \tag{A.133}$$

Here the right hand side is an approximation if the proton charge is almost concentrated at a single point, or exact for a point charge.

The interaction terms can now be included in the Lagrangian density to give the total Lagrangian

$$\mathcal{L}_{\text{seem+p}} = \int \Big(\pounds_{\text{seem}} + \pounds_{\text{int}} \Big) \, \mathrm{d}^3\vec{r} + \tfrac{1}{2} m_{\text{p}} {v_{\text{p}}}_j^{\,2} \tag{A.134}$$

$$\pounds_{\text{seem}} = -\frac{\epsilon_0}{2} \left(-{A_j}_t^{\,2} + c^2 {A_j}_i^{\,2} + \frac{1}{c^2} \varphi_t^2 - \varphi_i^2 \right) \qquad \pounds_{\text{int}} = -\varphi \rho_{\text{p}} + A_j {\jmath_{\text{p}}}_j$$

If there are more charged particles than just a proton, their charge and current densities will combine into a net ρ and $\vec{\jmath}$.

The field equations now follow similarly as in subsection A.22.2. For the electrostatic potential:

$$\frac{\partial}{\partial t} \left(\frac{\partial \pounds}{\partial \varphi_t} \right) + \frac{\partial}{\partial r_i} \left(\frac{\partial \pounds}{\partial \varphi_i} \right) = \frac{\partial \pounds}{\partial \varphi}$$

where $\pounds$ is the combined Lagrangian density. Worked out and converted to vector notation, that gives

$$\frac{1}{c^2} \frac{\partial^2 \varphi}{\partial t^2} - \nabla^2 \varphi = \frac{\rho}{\epsilon_0} \tag{A.135}$$

This is the same equation as for the Koulomb potential earlier.

Similarly, for the components of the vector potential

$$\frac{\partial}{\partial t} \left(\frac{\partial \pounds}{\partial {A_j}_t} \right) + \frac{\partial}{\partial r_i} \left(\frac{\partial \pounds}{\partial {A_j}_i} \right) = \frac{\partial \pounds}{\partial \varphi}$$

That gives

$$\frac{\partial^2 \vec{A}}{\partial t^2} - c^2 \nabla^2 \vec{A} = \frac{\vec{\jmath}}{\epsilon_0} \tag{A.136}$$

The above equations are again Klein-Gordon equations, so they respect the speed of light. And since the action is now Lorentz invariant, all observers agree with the evolution. That seems very encouraging.

Consider now the steady case, with no charge motion. The current density $\vec{\jmath}$ is then zero. That leads to zero vector potentials. Then there is no magnetic field either, (A.130).

The steady equation (A.135) for the electrostatic field φ is exactly the same as the one for the Koulomb potential. But note that the electric force per unit charge is now minus the gradient of the electrostatic potential, (A.130) and (A.132). And that means that like charges repel, not attract. All protons in the universe no longer clump together into one big ball. And neither do electrons. That sounds great.

But wait a second. How come that apparently protons suddenly manage to create fields that are repulsive to protons? What happened to energy minimization? It seems that all is not yet well in the universe.

A.22.5 Lorenz saves the universe

The previous subsection derived the self-evident equations of electromagnetics. But there were some worrisome aspects. A look at the Hamiltonian can clarify the problem.

Given the Lagrangian (A.134) of the previous subsection, the Hamiltonian can be found as, {A.1.5}:

$$H_{\text{seem+p}} = \int \left(\frac{\partial \pounds}{\partial A_{j_t}} A_{j_t} + \frac{\partial \pounds}{\partial \varphi_t} \varphi_t \right) \mathrm{d}^3\vec{r} + \frac{\partial \mathcal{L}}{\partial v_{\text{p}_j}} v_{\text{p}_j} - \mathcal{L}$$

That gives

$$H_{\text{seem+p}} = \frac{\epsilon_0}{2} \int \left(A_{j_t}^{\ 2} + c^2 A_{j_i}^{\ 2} - \frac{1}{c^2}\varphi_t^2 - \varphi_i^2 \right) \mathrm{d}^3\vec{r} + \tfrac{1}{2} m_{\text{p}} v_{\text{p}_j}^{\ 2} + \int \varphi \rho_{\text{p}} \, \mathrm{d}^3\vec{r} \tag{A.137}$$

(This would normally still need to be rewritten in terms of canonical momenta, but that is not important here.)

Note that the electrostatic potential φ produces *negative* electromagnetic energy. That means that the electromagnetic energy can have arbitrarily large negative values for large enough φ.

That then answers the question of the previous subsection: "How come a proton produces an electrostatic field that repels it? What happened to energy minimization?" There is no such thing as energy minimization here. If there is no lowest energy, then there is no ground state. Instead the universe should evolve towards larger and larger electrostatic fields. That would release infinite amounts of energy. It should blow life as we know it to smithereens. (The so-called second law of thermodynamics says, simply put, that thermal energy is easier to put into particles than to take out again. See chapter 11.)

In fact, the Koulomb force would also produce repulsion between equal sarges, if its field energy was negative instead of positive. Just change the sign of the constant ϵ_1 in classical selectodynamics. Then its universe should blow up too. Unlike what you will read elsewhere, the difference between the Koulomb force, (or its more widely known sibling, the Yukawa force of {A.42}), and the Coulomb force is not simply that the photon wave function is a four-vector. It is whether negative field energy appears in the most straightforward formulation.

As the previous subsection noted, you might assume that the electrodynamic Lagrangian, and hence the Hamiltonian, would have the opposite sign. But that does not help. In that case the vector potentials A_j would produce the negative energies. Reversing the sign of the Hamiltonian is like reversing the direction of time. In either direction, the universe gets blown to smithereens.

To be sure, it is not completely sure that the universe will be blown to smithereens. A negative field energy only says that it is in theory possible to

extract limitless amounts of energy out of the field. But you would still need some actual mechanism to do so. There might not be one. Nature might be carefully constrained so that there is no dynamic mechanism to extract the energy.

In that case, there might then be some mathematical expression for the constraint. As another way to look at that, suppose that you would indeed have a highly unstable system. And suppose that there is still something recognizable left at the end of the first day. Then surely you would expect that whatever is left is special in some way. That it obeys some special mathematical condition.

So presumably, the electromagnetic field that we observe obeys some special condition, some constraint. What could this constraint be? Since this is very basic physics, you would guess it to be relatively simple. Certainly it must be Lorentz invariant. The simplest condition that meets this requirement is that the dot product of the four-gradient $\overset{\leftrightarrow}{\nabla}$ with the four-potential $\overset{\leftrightarrow}{A}$ is zero. Written out that produces the so-called "Lorenz condition:"

$$\boxed{\frac{1}{c}\frac{\partial \varphi/c}{\partial t} + \nabla \cdot \vec{A} = 0} \tag{A.138}$$

This condition implies that only a very special subset of possible solutions of the Klein-Gordon equations given in the previous subsection is actually observed in nature.

Please note that the Lorenz condition is named after the Danish physicist Ludvig Loren*z*, not the Dutch physicist Hendrik Loren*tz*. Almost all my sources mislabel it the Lorentz condition. The savior of the universe deserves more respect. Always remember: the Loren*z* condition is Loren*tz* invariant.

(You might wonder why the first term in the Lorenz condition does not have the minus sign of dot products. One way of thinking about it is that the four-gradient in its "natural" condition already has a minus sign on the time derivative. Physicists call it a "covariant" four-vector rather than a "contravariant" one. A better way to see it is to grind it out; you can use the Lorentz transform (1.6) of chapter 1.2.1 to show directly that the above form is the same for different observers. But those familiar with index notation will recognize immediately that the Lorenz condition is Lorentz invariant from the fact that it equals $\partial_\mu A^\mu = 0$, and that has μ as both subscript and superscript. See chapter 1.2.5 for more.)

To be sure, the Lorenz condition can only be true if the interaction with matter does not produce violations. To check that, the evolution equation for the Lorenz condition quantity can be obtained from the Klein-Gordon equations of the previous subsection. In particular, in vector notation take $\partial/\partial t$ (A.135) plus ∇ (A.136) to get

$$\left[\frac{\partial^2}{\partial t^2} - c^2\nabla^2\right]\left(\frac{1}{c^2}\frac{\partial \varphi}{\partial t} + \nabla \cdot \vec{A}\right) = \frac{1}{\epsilon_0}\left(\frac{\partial \rho}{\partial t} + \nabla \cdot \vec{\jmath}\right) \tag{A.139}$$

The parenthetical expression in the left hand side should be zero according to the Lorenz condition. But that is only possible if the left hand side is zero too, so

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \vec{\jmath}$$

This important result is known as "Maxwell's continuity equation." It expresses conservation of charge. (To see that, take any arbitrary volume. Integrate both sides of the continuity equation over that volume. The left hand side then becomes the time derivative of the charge inside the volume. The right hand side becomes, using the [divergence] [Gauss] [Ostrogradsky] theorem, the net inflow of charge. And if the charge inside can only change due to inflow or outflow, then no charge can be created out of nothing or destroyed.) So charge conservation can be seen as a consequence of the need to maintain the Lorenz condition.

Note that the Lorenz condition (A.138) looks mathematically just like the continuity equation. It produces conservation of the integrated electrostatic potential. In subsection A.22.7 it will be verified that it is indeed enough to produce a stable electromagnetic field. One with meaningfully defined energies that do not run off to minus infinity.

Note that charge conservation by itself is not quite enough to ensure that the Lorenz condition is satisfied. However, if in addition the Lorenz quantity and its time derivative are zero at just a single time, it is OK. Then (A.139) ensures that the Lorenz condition remains true for all time.

A.22.6 Gupta-Bleuler condition

The ideas of the previous subsection provide one way to quantize the electromagnetic field, ⟦17, 6⟧.

As already seen in subsection A.22.3 (A.128), in quantum field theory the potentials become quantum fields, i.e. operator fields. For electromagnetics the quantum field four-vector is a bit more messy

$$\widehat{\vec{A}} = \begin{pmatrix} \widehat{\varphi}/c \\ \widehat{\vec{A}} \end{pmatrix} = \sum_{\vec{k}} \frac{\varepsilon_k}{\sqrt{2}k} \sum_{\nu=0}^{3} \overrightarrow{e}_{\vec{k}}^{\,\nu} \left(e^{\mathrm{i}\vec{k}\cdot\vec{r}} \widehat{a}_{\vec{k}\nu} + e^{-\mathrm{i}\vec{k}\cdot\vec{r}} \widehat{a}_{\vec{k}\nu}^\dagger \right)$$

Since a four-vector has four components, a general four-vector can be written as a linear combination of four chosen basis four-vectors $\overrightarrow{e}_{\vec{k}}^{\,0}$, $\overrightarrow{e}_{\vec{k}}^{\,1}$, $\overrightarrow{e}_{\vec{k}}^{\,2}$, and $\overrightarrow{e}_{\vec{k}}^{\,3}$. (That is much like a general vector in three dimensions can be written as a linear combination of $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$.) The four basis vectors physically represent different possible "polarizations" of the electromagnetic field. That is why they are typically aligned with the momentum of the wave rather than with some Cartesian axis system and its time axis. Note that each polarization vector has

its own annihilation operator $\widehat{a}_{\vec{k}\nu}$ and creation operator $\widehat{a}^\dagger_{\vec{k}\nu}$. These annihilate respectively create photons with that wave number vector $\vec{k}$ and polarization.

(Electromagnetic waves in empty space are special; for them only two independent polarizations are possible. Or to be precise, even in empty space the Klein-Gordon equations with Lorenz condition allow a third polarization. But these waves produce no electric and magnetic fields and contain no net electromagnetic energy. So they are physically irrelevant. You can call them "gage equivalent to the vacuum." That sounds better than irrelevant.)

The Lorenz condition of the previous subsection is again needed to get rid of negative energy states. The question is now exactly how to phrase the Lorenz condition in quantum terms.

(There is an epidemic among my, highly authorative, sources that come up with negative norm states without Lorenz condition. Now the present author himself is far from an expert on quantum field theories. But he knows one thing: *norms cannot be negative.* If you come up with negative norms, it tells you nothing about the physics. It tells you that you are doing the mahematics wrong. I believe the correct argument goes something like this: "Suppose that we can do our usual stupid canonical quantization tricks for this system. Blah Blah. That gives negative norm states. Norms cannot be negative. Ergo: we cannot do our usual stupid canonical quantization tricks for this system." If you properly define the creation and annihilation operators to put photons in negative energy states, there is no mathematical problem. The commutator relation for the negative energy states picks up a minus sign and the norms are positive as they should. Now the mathematics is sound and you can start worrying about problems in the physics. Like that there are negative energy states. And maybe lack of Lorentz invariance, although the original system is Lorentz invariant, and I do not see what would not be Lorentz invariant about putting particles in the negative energy states.)

The simplest idea would be to require that the quantum field above satisfies the Lorenz condition. But the quantum field determines the dynamics. Like in the classical case, you do not want to change the dynamics. Instead you want to throw certain solutions away. That means that you want to throw certain wave functions $|\Psi\rangle$ away.

The strict condition would be to require (in the Heisenberg picture {A.12})

$$\Big(\frac{1}{c}\frac{\partial\widetilde{\varphi}/c}{\partial t} + \nabla\cdot\widetilde{\vec{A}}\Big)|\Psi\rangle = 0$$

for all physically observable states $|\Psi\rangle$. Here the parenthetical expression is the operator of the Lorenz quantity that must be zero. The above requirement makes $|\Psi\rangle$ an eigenvector of the Lorenz quantity with eigenvalue zero. Then according to the rules of quantum mechanics, chapter 3.4, the only measurable value of the Lorenz quantity is zero.

But the above strict condition is too restrictive. Not even the vacuum state with no photons would be physically observable. That is because the creation operators in $\widehat{\varphi}$ and $\widehat{\vec{A}}$ will create nonzero photon states when applied on the vacuum state. That suggests that only the annihilation terms should be included. That then gives the "Gupta-Bleuler condition:"

$$\Big(\frac{1}{c}\frac{\partial\widetilde{\varphi}^+/c}{\partial t} + \nabla\cdot\widetilde{\vec{A}}^+\Big)|\Psi\rangle = 0$$

for physically observable states $|\Psi\rangle$. Here the superscript + on the quantum fields means that only the $\widehat{a}_{\vec{k}\nu}$ annihilation operator terms are included.

You might of course wonder why the annihilation terms are indicated by a plus sign, instead of the creation terms. After all, it are the creation operators that create more photons. But the plus sign actually stands for the fact that the annihilation terms are associated with an $e^{-\mathrm{i}\omega t}$ time dependence instead of $e^{\mathrm{i}\omega t}$. Yes true, $e^{-\mathrm{i}\omega t}$ has a minus sign, not a plus sign. But $e^{-\mathrm{i}\omega t}$ has the normal sign, and "normal" is represented by a plus sign. Is not addition more normal than subtraction? Please do not pull at your hair like that, there are less drastic ways to save on professional hair care.

Simply dropping the creation terms may seem completely arbitrary. But it actually has some physical logic to it. Consider the inner product

$$\langle\Psi'|\Big(\frac{1}{c}\frac{\partial\widetilde{\varphi}/c}{\partial t} + \nabla\cdot\widetilde{\vec{A}}\Big)|\Psi\rangle = 0$$

This is the amount of state $|\Psi'\rangle$ produced by applying the Lorenz quantity on the physically observable state $|\Psi\rangle$. The strict condition is equivalent to saying that this inner product must always be zero; no amount of any state may be produced. For the Gupta-Bleuler condition, the above inner product remains zero if $|\Psi'\rangle$ is a physically observable state. (The reason is that the creation terms can be taken to the other side of the inner product as annihilation terms. Then they produce zero if $|\Psi'\rangle$ is physically observable.) So the Gupta-Bleuler condition implies that no amount of any physically observable state may be produced by the Lorenz quantity.

There are other ways to do quantization of the electromagnetic field. The quantization following Fermi, as discussed in subsection A.22.8, can be converted into a modern quantum field theory. But that turns out to be a very messy process indeed, [[17, 6]]. The derivation is essentially to mess around at length until you more or less prove that you can use the Lorenz condition result instead. You might as well start there.

It does turns out that the so-called "path-integral" formulation of quantum mechanics does a very nice job here, [53, pp. 30ff]. It avoids many of the contortions of canonical quantization like the ones above.

In fact, a popular quantum field textbook, [35, p. 79], refuses to do canonical quantization of the electromagnetic field at all, calling it an awkward subject.

This book is typically used during the second year of graduate study in physics, so it is not that its readers are unsophisticated.

A.22.7 The conventional Lagrangian

Returning to the classical electromagnetic field, it still needs to be examined whether the Lorenz condition has made the universe safe for life as we know it.

The answer depends on the Lagrangian, because the Lagrangian determines the evolution of a system. So far, the Lagrangian has been written in terms of the four potentials φ and A_j (with $j = 1$, 2, and 3) of the electromagnetic field. But recall that matter does not observe the four potentials directly. It only notices the electric field $\vec{\mathcal{E}}$ and the magnetic field $\vec{\mathcal{B}}$. So it may help to reformulate the Lagrangian in terms of the electric and magnetic fields. Concentrating on the observed fields is likely to show up more clearly what is actually observed.

With a bit of mathematical manipulation, {D.37.3}, the self-evident electromagnetic Lagrangian density can be written as:

$$\mathcal{L}_{\rm seem} = \frac{\epsilon_0}{2}\Big(\mathcal{E}^2 - c^2\mathcal{B}^2 - c^2\Big\{\frac{1}{c^2}\frac{\partial\varphi}{\partial t} + \nabla\cdot\vec{A}\Big\}^2\Big) + ...$$

Here the dots stand for terms that do not affect the motion. (Since in the action, Lagrangian densities get integrated over space and time, terms that are pure spatial or time derivatives integrate away. The quantities relevant to the action principle vanish at the limits of integration.)

The term inside the curly brackets is zero according to the Lorenz condition (A.138). Therefore, it too does not affect the motion. (To be precise, the term does not affect the motion because it is squared. By itself it would affect the motion. In the formal way in which the Lagrangian is differentiated, one power is lost.)

The conventional Lagrangian density is found by disregarding the terms that do not change the motion:

$$\mathcal{L}_{\rm conem} = \frac{\epsilon_0}{2}\Big(\mathcal{E}^2 - c^2\mathcal{B}^2\Big)$$

So the conventional Lagrangian density of the electromagnetic field is completely in terms of the observable fields.

As an aside, it might be noted that physicists find the above expression too intuitive. So you will find it in quantum field books in relativistic index notation as:

$$\mathcal{L}_{\rm conem} = -\frac{\epsilon_0}{4}F_{\mu\nu}F^{\mu\nu}$$

Here the "field strength tensor" is defined by

$$F_{\mu\nu} = c\,(\partial_\mu A_\nu - \partial_\nu A_\mu) \qquad \mu = 0,1,2,3 \quad \nu = 0,1,2,3$$

Note that the indices on each A are subscripts instead of superscripts as they should be. That means that you must add a minus sign whenever the index on an A is 0. If you do that correctly, you will find that from the 16 $F_{\mu\nu}$ values, some are zero, while the rest are components of the electric or magnetic fields. To go from $F_{\mu\nu}$ to $F^{\mu\nu}$, you must raise both indices, so add a minus sign for each index that is zero. If you do all that the same Lagrangian density as before results.

Because the conventional Lagrangian density is different from the self-evident one, the field equations (A.135) and (A.136) for the potentials pick up a few additional terms. To find them, repeat the analysis of subsection A.22.4 but use the conventional density above in (A.134). Note that you will need to write the electric and magnetic fields in terms of the potentials using (A.131). (Using the field strength tensor is actually somewhat simpler in converting to the potentials. If you can get all the blasted sign changes right, that is.)

Then the conventional field equations become:

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} - \nabla^2\varphi - \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} - \frac{\partial \nabla\cdot\vec{A}}{\partial t} = \frac{\rho}{\epsilon_0} \tag{A.140}$$

$$\frac{\partial^2\vec{A}}{\partial t^2} - c^2\nabla^2\vec{A} + \nabla\frac{\partial\varphi}{\partial t} + c^2\nabla(\nabla\cdot\vec{A}) = \frac{\vec{\jmath}}{\epsilon_0} \tag{A.141}$$

Here ρ is again the charge density and $\vec{\jmath}$ the current density of the charges that are around,

The additional terms in each equation above are the two before the equals signs. Note that these additional terms are zero on account of the Lorenz condition. So they do not change the solution.

The conventional field equations above are obviously more messy than the original ones. Even if you cancel the second order time derivatives in (A.140). However, they do have one advantage. If you use these conventional equations, you do not have to worry about satisfying the Lorenz condition. Any solution to the equations will give you the right electric and magnetic fields and so the right motion of the charged particles.

To be sure, the potentials will be different if you do not satisfy the Lorenz condition. But the potentials have no meaning of their own. At least not in classical electromagnetics.

To verify that the Lorenz condition is no longer needed, first recall the indeterminacy in the potentials. As subsection A.22.4 discussed, more than one set of potentials can produce the same electric and magnetic fields. In particular, given potentials φ and $\vec{A}$, you can create equivalent potentials as

$$\varphi' = \varphi - \chi_t \qquad \vec{A}' = \vec{A} + \nabla\chi$$

Here χ can be any function of space and time that you want. The potentials φ' and $\vec{A}'$ give the exact same electric and magnetic fields as φ and $\vec{A}$. Such a transformation of potentials is called a "gage transform."

Now suppose that you have a solution φ and $\vec{A}$ of the conventional field equations, but it does not satisfy the Lorenz condition. In that case, simply apply a gage transform as above to get new fields φ' and $\vec{A}'$ that do satisfy the Lorenz condition. To do so, write out the Lorenz condition for the new potentials,

$$\frac{1}{c^2}\frac{\partial \varphi'}{\partial t} + \nabla \cdot \vec{A}' = \frac{1}{c^2}\frac{\partial \varphi}{\partial t} - \frac{1}{c^2}\frac{\partial^2 \chi}{\partial t^2} + \nabla \cdot \vec{A} + \nabla^2 \chi$$

You can always choose the function χ to make this quantity zero. (Note that that gives an inhomogeneous Klein-Gordon equation for χ.)

Now it turns out that the new potentials φ' and $\vec{A}'$ still satisfy the conventional equations. That can be seen by straight substitution of the expressions for the new potentials in the conventional equations. So the new potentials are perfectly OK: they satisfy both the Lorenz condition and the conventional equations. But the original potentials φ and $\vec{A}$ produced the exact same electric and magnetic fields. So the original potentials were OK too.

The evolution equation (A.140) for the electrostatic field is worth a second look. Because of the definition of the electric field (A.130), it can be written as

$$\nabla \cdot \vec{\mathcal{E}} = \frac{\rho}{\epsilon_0} \tag{A.142}$$

That is called "Maxwell's first equation," chapter 13.2. It ties the charge density to the electric field quite rigidly.

Maxwell's first equation is a consequence of the Lorenz condition. It would not be required for the original Klein-Gordon equations without Lorenz condition. In particular, it is the Lorenz condition that allows the additional two terms in the evolution equation (A.140) for the electrostatic potential. These then eliminate the second order time derivative from the equation. That then turns the equation from a normal evolution equation into a restrictive spatial condition on the electric field.

It may be noted that the other evolution equation (A.141) is Maxwell's fourth equation. Just rewrite it in terms of the electric and magnetic fields. The other two Maxwell equations follow trivially from the definitions (A.130) of the electric and magnetic fields in terms of the potentials.

Since there is no Lorenz condition for the conventional equations, it becomes interesting to find the corresponding Hamiltonian. That should allow the stability of electromagnetics to be examined more easily.

The Hamiltonian for electromagnetic field plus a proton may be found the same way as (A.137) in subsection A.22.5, {A.1.5}. Just use the conventional Lagrangian density instead. That gives

$$H_{\text{conem+p}} = \int \left(\frac{\epsilon_0}{2}(\mathcal{E}^2 + c^2\mathcal{B}^2 + 2\mathcal{E}_i\varphi_i) + \varphi\rho_{\text{p}} \right) \mathrm{d}^3\vec{r} + \tfrac{1}{2} m_{\text{p}} {v_{\text{p}}}_j^2$$

But the proton charge density $\rho_{\rm p}$ may eliminated using Maxwell's first equation above. An additional integration by parts of that term then causes it to drop away against the previous term. That gives the conventional energy as

$$E_{\rm conem+p} = \frac{\epsilon_0}{2} \int (\mathcal{E}^2 + c^2 \mathcal{B}^2) \, {\rm d}^3 \vec{r} + \tfrac{1}{2} m_{\rm p} \vec{v}_{\rm p}^{\,2} \tag{A.143}$$

The first term is the energy in the observable fields and the final term is the kinetic energy of the proton.

The simplified energy above is no longer really a Hamiltonian; you cannot write Hamilton's equations based on it as in {A.1.5}. But it does still give the energy that is conserved.

The energy above is always positive. So it can no longer be lowered by arbitrary amounts. The system will not blow up. And that then means that the original Klein-Gordon equations (A.135) and (A.136) for the fields are stable too as long as the Lorenz condition is satisfied. They produce the same evolution. And they satisfy the speed of light restriction and are Lorentz invariant. Lorenz did it!

Note also the remarkable result that the interaction energy between proton charge and field has disappeared. The proton can no longer minimize any energy of interaction between itself and the field it creates. Maxwell's first equation is too restrictive. All the proton can try to do is minimize the energy in the electric and magnetic fields.

A.22.8 Quantization following Fermi

Quantizing the electromagnetic field is not easy. The previous subsection showed a couple of problems. The gage property implies that the electromagnetic potentials φ and $\vec{A}$ are indeterminate. Also, taking the Lorenz condition into account, the second order time derivative is lost in the Klein-Gordon equation for the electrostatic potential φ. The equation turns into Maxwell's first equation,

$$\nabla \cdot \vec{\mathcal{E}} = \frac{\rho}{\epsilon_0}$$

That is not an evolution equation but a spatial constraint for the electric field $\vec{\mathcal{E}}$ in terms of the charge density ρ.

Various ways to deal with that have been developed. The quantization procedure discussed in this subsection is a simplified version of the one found in Bethe's book, [6, pp. 255-271]. It is due to Fermi, based on earlier work by Dirac and Heisenberg & Pauli. This derivation was a great achievement at the time, and fundamental to more advanced quantum field approaches, [6, p. 266]. Note that all five mentioned physicists received a Nobel Prize in physics at one time or the other.

The starting point in this discussion will be the original potentials φ and $\vec{A}$ of subsection A.22.4. The ones that satisfied the Klein-Gordon equations (A.135) and (A.136) as well as the Lorenz condition (A.138).

It was Fermi who recognized that you can make things a lot simpler for yourself if you write the potentials as sums of exponentials of the form $e^{\mathrm{i}\vec{k}\cdot\vec{r}}$:

$$\varphi = \sum_{\text{all } \vec{k}} c_{\vec{k}} e^{\mathrm{i}\vec{k}\cdot\vec{r}} \qquad \vec{A} = \sum_{\text{all } \vec{k}} \vec{d}_{\vec{k}} e^{\mathrm{i}\vec{k}\cdot\vec{r}}$$

That is the same trick as was used in quantizing the Koulomb potential in subsection A.22.3. However, in classical mechanics you do not call these exponentials momentum eigenstates. You call them "Fourier modes." The principle is the same. The constant vector $\vec{k}$ that characterizes each exponential is still called the wave number vector. Since the potentials considered here vary with time, the coefficients $c_{\vec{k}}$ and $\vec{d}_{\vec{k}}$ are functions of time.

Note that the coefficients $\vec{d}_{\vec{k}}$ are vectors. These will have three independent components. So the vector potential can be written more explicitly as

$$\vec{A} = \sum_{\text{all } \vec{k}} d_{1,\vec{k}}\, \vec{e}_{1,\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} + d_{2,\vec{k}}\, \vec{e}_{2,\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} + d_{3,\vec{k}}\, \vec{e}_{3,\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}}$$

where $\vec{e}_{1,\vec{k}}$, $\vec{e}_{2,\vec{k}}$, and $\vec{e}_{3,\vec{k}}$ are unit vectors. Fermi proposed that the smart thing to do is to take the first of these unit vectors in the same direction as the wave number vector $\vec{k}$. The corresponding electromagnetic waves are called "longitudinal." The other two unit vectors should be orthogonal to the first component and to each other. That still leaves a bit choice in direction. Fortunately, in practice it does not really make a difference exactly how you take them. The corresponding electromagnetic waves are called "transverse."

In short, the fields can be written as

$$\varphi = \sum_{\text{all } \vec{k}} c_{\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} \qquad \vec{A}_{\parallel} = \sum_{\text{all } \vec{k}} d_{1,\vec{k}}\, \vec{e}_{1,\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} \qquad \vec{A}_{\perp} = \sum_{\text{all } \vec{k}} d_{2,\vec{k}}\, \vec{e}_{2,\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} + d_{3,\vec{k}}\, \vec{e}_{3,\vec{k}}\, e^{\mathrm{i}\vec{k}\cdot\vec{r}} \tag{A.144}$$

where

$$\vec{e}_{1,\vec{k}} = \frac{\vec{k}}{k} \qquad \vec{e}_{2,\vec{k}} \cdot \vec{k} = \vec{e}_{3,\vec{k}} \cdot \vec{k} = \vec{e}_{2,\vec{k}} \cdot \vec{e}_{3,\vec{k}} = 0$$

From those expressions, and the directions of the unit vectors, it can be checked by straight substitution that the "curl" of the longitudinal potential is zero:

$$\operatorname{curl} \vec{A}_{\parallel} \equiv \nabla \times \vec{A}_{\parallel} = 0 \quad \text{(irrotational)}$$

A vector field with zero curl is called "irrotational." (The term can be understood from fluid mechanics; there the curl of the fluid velocity field gives the local average angular velocity of the fluid.)

The same way, it turns out that that the "divergence" of the transverse potential is zero

$$\operatorname{div} \vec{A}_\perp \equiv \nabla \cdot \vec{A}_\perp = 0 \quad \text{(solenoidal)}$$

A field with zero divergence is called "solenoidal." (This term can be understood from magnetostatics; a magnetic field, like the one produced by a solenoid, an electromagnet, has zero divergence.)

To be fair, Fermi did not really discover that it can be smart to take vector fields apart into irrotational and solenoidal parts. That is an old trick known as the "Helmholtz decomposition."

Since the transverse potential has no divergence, the longitudinal potential is solely responsible for the Lorenz condition (A.138). The transverse potential can do whatever it wants.

The real problem is therefore with the longitudinal potential $\vec{A}_\parallel$ and the electrostatic potential φ. Bethe [6] deals with these in terms of the Fourier modes. However, that requires some fairly sophisticated analysis. It is actually easier to return to the potentials themselves now.

Reconsider the expressions (A.130) for the electric and magnetic fields in terms of the potentials. They show that the electrostatic potential produces no magnetic field. And neither does the longitudinal potential because it is irrotational.

They do produce a combined electric field $\vec{\mathcal{E}}_{\varphi\parallel}$. But this electric field is irrotational, because the longitudinal potential is, and the gradient ∇ of any scalar function is. That helps, because then the Stokes theorem of calculus implies that the electric field $\vec{\mathcal{E}}_{\varphi\parallel}$ is minus the gradient of some scalar potential:

$$\vec{\mathcal{E}}_{\varphi\parallel} = -\nabla\varphi_{\mathrm{C}}$$

Note that normally φ_{C} is not the same as the electrostatic potential φ, since there is also the longitudinal potential. To keep them apart, φ_{C} will be called the "Coulomb potential."

As far as the divergence of the electric field $\vec{\mathcal{E}}_{\varphi\parallel}$ is concerned, it is the same as the divergence of the complete electric field. The reason is that the transverse field has no divergence. And the divergence of the complete electric field is given by Maxwell's first equation. Together these observations give

$$\vec{\mathcal{E}}_{\varphi\parallel} = -\nabla\varphi_{\mathrm{C}} \qquad \vec{\mathcal{B}}_{\varphi\parallel} = 0 \qquad \nabla \cdot \vec{\mathcal{E}}_{\varphi\parallel} = -\nabla^2\varphi_{\mathrm{C}} = \frac{\rho}{\epsilon_0}$$

Note that the final equation is a Poisson equation for the Coulomb potential.

Now suppose that you replaced the electrostatic field φ with the Coulomb potential φ_{C} and had no longitudinal field $\vec{A}_\parallel$ at all. It would give the same electric and magnetic fields. And they are the only ones that are observable. They give the forces on the particles. The potentials are just mathematical tools in classical electromagnetics.

So why not? To be sure, the combination of the Coulomb potential $\varphi_{\rm C}$ and remaining vector potential $\vec{A}_\perp$ will no longer satisfy the Lorenz condition. But who cares?

Instead of the Lorenz condition, the combination of Coulomb potential plus transverse potential satisfies the so-called "Coulomb condition:"

$$\boxed{\nabla \cdot \vec{A} = 0} \tag{A.145}$$

The reason is that now $\vec{A} = \vec{A}_\perp$ and the transverse vector potential has no divergence. Physicists like to say that the original potentials used the "Lorenz gage," while the new ones use the "Coulomb gage."

Because the potentials $\varphi_{\rm C}$ and $\vec{A}_\perp$ do no longer satisfy the Lorenz condition, the Klein-Gordon equations (A.135) and (A.136) do no longer apply. But the conventional equations (A.140) and (A.141) do still apply; they do not need the Lorenz condition.

Now consider the Coulomb potential somewhat closer. As noted above it satisfies the Poisson equation

$$-\nabla^2 \varphi_{\rm C} = \frac{\rho}{\epsilon_0}$$

The solution to this equation was already found in the first subsection, (A.107). If the charge distribution ρ consists of a total of I point charges, it is

$$\varphi_{\rm C}(\vec{r};t) = \sum_{i=1}^{I} \frac{q_i}{4\pi\epsilon_0 |\vec{r} - \vec{r}_i|} \tag{A.146}$$

Here q_i is the charge of point charge number i, and $\vec{r}_i$ its position.

If the charge distribution ρ is smoothly distributed, simply take it apart in small "point charges" $\rho(\underline{\vec{r}};t){\rm d}^3\underline{\vec{r}}$. That gives

$$\varphi_{\rm C}(\vec{r};t) = \int_{\rm all\ \underline{\vec{r}}} \frac{\rho(\underline{\vec{r}};t)}{4\pi\epsilon_0 |\vec{r} - \underline{\vec{r}}|}\, {\rm d}^3\underline{\vec{r}} \tag{A.147}$$

The key point to note here is that the Coulomb potential has no life of its own. It is rigidly tied to the positions of the charges. That then provides the most detailed answer to the question: "What happened to energy minimization?" Charged particles have no option of minimizing any energy of interaction with the field. Maxwell's first equation, the Poisson equation above, forces them to create a Coulomb field that is repulsive to them. Whether they like it or not.

Note further that all the mechanics associated with the Coulomb field is quasi-steady. The Poisson equation does not depend on how fast the charged particles evolve. The Coulomb electric field is minus the spatial gradient of the potential, so that does not depend on the speed of evolution either. And the

Coulomb force on the charged particles is merely the electric field times the charge.

It is still not obvious how to quantize the Coulomb potential, even though there is no longer a longitudinal field. But who cares about the Coulomb potential in the first place? The important thing is how the charged particles are affected by it. And the forces on the particles caused by the Coulomb potential can be computed using the electrostatic potential energy, {D.37.4},

$$V_{\rm C} = \tfrac{1}{2} \sum_{i=1}^{I} \sum_{\substack{\underline{i}=1 \\ \underline{i}\neq i}}^{I} \frac{q_i q_{\underline{i}}}{4\pi\epsilon_0 |\vec{r}_i - \vec{r}_{\underline{i}}|} \tag{A.148}$$

For example, this is the Coulomb potential energy that was used to find the energy levels of the hydrogen atom in chapter 4.3. It can still be used in unsteady motion because everything associated with the Coulomb potential is quasi-steady. Sure, it is due to the interaction of the particles with the electromagnetic field. But where in the above mathematical expression does it say electromagnetic field? All it contains are the coordinates of the charged particles. So what difference does it make where the potential energy comes from? Just add the energy above to the Hamiltonian and then pretend that there *are no* electrostatic and longitudinal fields.

Incidentally, note the required omission of the terms with $\underline{i} = i$ in the potential energy above. Otherwise you would get infinite energy. In fact, a point charge in classical electromagnetics does have infinite Coulomb energy. Just take any of the point charges and mentally chop it up into two equal parts sitting at the same position. The interaction energy between the halves is infinite.

The issue does not exist if the charge is smoothly distributed. In that case the Coulomb potential energy is, {D.37.4},

$$V_{\rm C} = \tfrac{1}{2} \int_{\text{all } \vec{r}} \int_{\text{all } \underline{\vec{r}}} \frac{\rho(\vec{r};t)\rho(\underline{\vec{r}};t)}{4\pi\epsilon_0 |\vec{r} - \underline{\vec{r}}|} \, \mathrm{d}^3\vec{r} \mathrm{d}^3\underline{\vec{r}} \tag{A.149}$$

While the integrand is infinite at $\underline{\vec{r}} = \vec{r}$, the integral remains finite.

So the big idea is to throw away the electrostatic and longitudinal potentials and replace them with the Coulomb energy $V_{\rm C}$, origin unknown. Now it is mainly a matter of working out the details.

First, consider the Fermi Lagrangian. It is found by throwing out the electrostatic and longitudinal potentials from the earlier Lagrangian (A.134) and subtracting $V_{\rm C}$. That gives, using the point charge approximation (A.133) and in vector notation,

$$\mathcal{L}_{\rm F} = \frac{\epsilon_0}{2} \int \left[\left| \vec{A}_{\perp_t} \right|^2 - \sum_{\underline{j}=1}^{3} c^2 \left| \vec{A}_{\perp_{\underline{j}}} \right|^2 \right] \mathrm{d}^3\vec{r} + \sum_{i=1}^{I} \left[q_i \vec{v}_i \cdot \vec{A}_{\perp_i} + \tfrac{1}{2} m_i \vec{v}_i^{\,2} \right] - V_{\rm C} \tag{A.150}$$

Note that it is now assumed that there are I particles instead of just the single proton in (A.134). Because i is already used to index the particles, $\underline{j}$ is used to index the three directions of spatial differentiation. The Coulomb energy $V_{\rm C}$ was already given in (A.148). The velocity of particle i is $\vec{v}_i$, while q_i is its charge and m_i its mass. The subscript i on the transverse potential in the interaction term indicates that it is evaluated at the location of particle i.

You may wonder how you can achieve that only the transverse potential $\vec{A}_\perp$ is left. That would indeed be difficult to do if you work in terms of spatial coordinates. The simplest way to handle it is to work in terms of the transverse waves (A.144). They are transverse by construction.

The unknowns are now no longer the values of the potential at the infinitely many possible positions. Instead the unknowns are now the coefficients $d_{2,\vec{k}}$ and $d_{3,\vec{k}}$ of the transverse waves. Do take into account that since the field is real,

$$d_{2,-\vec{k}} = d^*_{2,\vec{k}} \qquad d_{3,-\vec{k}} = d^*_{3,\vec{k}}$$

So the number of independent variables is half of what it seems. The most straightforward way of handling this is to take the unknowns as the real and imaginary parts of the $d_{2,\vec{k}}$ and $d_{3,\vec{k}}$ for half of the $\vec{k}$ values. For example, you could restrict the $\vec{k}$ values to those for which the first nonzero component is positive. The corresponding unknowns must then describe both the $\vec{k}$ and $-\vec{k}$ waves.

(The $\vec{k} = 0$ terms are awkward. One way to deal with it is to take an adjacent periodic box and reverse the sign of all the charges and fields in it. Then take the two boxes together to be a new bigger periodic box. The net effect of this is to shift the mesh of $\vec{k}$-values figure 6.17 by half an interval. That means that the $\vec{k} = 0$ terms are gone. And other problems that may arise if you sum over all boxes, like to find the total Coulomb potential, are gone too. Since the change in $\vec{k}$ values becomes zero in the limit of infinite box size, all this really amounts to is simply ignoring the $\vec{k} = 0$ terms.)

The Hamiltonian can be obtained just like the earlier one (A.137), {A.1.5}. (Or make that {A.1.4}, since the unknowns, $d_{2,\vec{k}}$ and $d_{3,\vec{k}}$, are now indexed by the discrete values of the wave number $\vec{k}$.) But this time it really needs to be done right, because this Hamiltonian is supposed to be actually used. It is best done in terms of the components of the potential and velocity vectors. Using j to index the components, the Lagrangian becomes

$$\mathcal{L}_{\rm F} = \frac{\epsilon_0}{2} \int \sum_{j=1}^{3} \Big[\left|A_{\perp j_t}\right|^2 - \sum_{\underline{j}=1}^{3} c^2 \left|A_{\perp j_{\underline{j}}}\right|^2 \Big] \,{\rm d}^3\vec{r} + \sum_{i=1}^{I} \sum_{j=1}^{3} \Big[q_i v_{i_j} A_{\perp j_i} + \tfrac{1}{2} m_i v_{i_j}^2 \Big] - V_{\rm C}$$

Now Hamiltonians should *not* be in terms of particle velocities, despite what (A.137) said. Hamiltonians should be in terms of "canonical momenta,"

{A.1.4}. The canonical momentum corresponding to the velocity component v_{i_j} of a particle i is defined as

$$p^{\rm c}_{i_j} \equiv \frac{\partial \mathcal{L}}{\partial v_{i_j}}$$

Differentiating the Lagrangian above gives

$$p^{\rm c}_{i_j} = m_i v_{i_j} + q_i A_{j_i}$$

It is this canonical momentum that in quantum mechanics gets replaced by the operator $\hbar\partial/\mathrm{i}\partial r_j$. That is important since, as the above expression shows, canonical momentum is not just linear momentum in the presence of an electromagnetic field.

The time derivatives of the real and imaginary parts of the coefficients $d_{2,\vec{k}}$ and $d_{3,\vec{k}}$ should be replaced by similarly defined canonical momenta. However, that turns out to be a mere rescaling of these time derivatives.

The Hamiltonian then becomes, following {A.1.4} and in vector notation,

$$\begin{aligned} H_{\rm F} &= \frac{\epsilon_0}{2}\int \Big[\Big|\vec{A}_{\perp_t}\Big|^2 + \sum_{\underline{j}=1}^{3} c^2 \Big|\vec{A}_{\perp_{\underline{j}}}\Big|^2 \Big]\, \mathrm{d}^3\vec{r} \\ &\quad + \sum_{i=1}^{I} \frac{(\vec{p}^{\,\rm c}_i - q_i\vec{A}_{\perp_i})^2}{2m_i} + \tfrac{1}{2}\sum_{i=1}^{I}\sum_{\substack{\underline{i}=1 \\ \underline{i}\neq i}}^{I} \frac{q_i q_{\underline{i}}}{4\pi\epsilon_0|\vec{r}_i - \vec{r}_{\underline{i}}|} \end{aligned} \qquad \text{(A.151)}$$

Note in particular that the center term is the kinetic energy of the particles, but in terms of their canonical momenta.

In terms of the waves (A.144), the integral falls apart in separate contributions from each $d_{2,\vec{k}}$ and $d_{3,\vec{k}}$ mode. That is a consequence of the orthogonality of the exponentials, compare the Parseval identity in {A.26}. (Since the exponentials are complex, the absolute values in the integral are now required.) As a result, the equations for different coefficients are only indirectly coupled by the interaction with the charged particles. In particular, it turns out that each coefficient satisfies its own harmonic oscillator equation with forcing by the charged particles, {A.1.4},

$$\epsilon_0\mathcal{V}(\ddot{d}_{j,\vec{k}} + k^2c^2 d_{j,\vec{k}}) = \sum_i q_i\vec{v}_i\cdot\vec{e}_{j,\vec{k}}e^{-\mathrm{i}\vec{k}\cdot\vec{r}_i} \quad \text{for } j = 2 \text{ and } 3$$

If the speed of the particle gets comparable to the speed of light, you may want to use the relativistic energy (1.2);

$$\frac{(\vec{p}^{\,\rm c}_i - q_i\vec{A}_{\perp_i})^2}{2m_i} \quad\Longrightarrow\quad \sqrt{(m_ic^2)^2 + (\vec{p}^{\,\rm c}_i - q_i\vec{A}_{\perp_i})^2c^2}$$

Sometimes, it is convenient to assume that the system under consideration also experiences an external electromagnetic field. For example, you might consider an atom or atomic nucleus in the magnetic field produced by an electromagnet. You probably do not want to include every electron in the wires of the electromagnet in your model. That would be something else. Instead you can simply add the vector potential $\vec{A}_{\text{ext}_i}$ that they produce to $\vec{A}_{\perp_i}$ in the Hamiltonian. If there is also an external electrostatic potential, add a separate term $q_i\varphi_{\text{ext}i}$ to the Hamiltonian for each particle i. The external fields will be solutions of the homogeneous evolution equations (A.140) and (A.141), (i.e. the equations without charge and current densities). However, the external fields will not vanish at infinity; that is why they can be nonzero without charge and current densities.

Note that the entire external vector potential is needed, not just the transverse part. The longitudinal part is not included in V_{C}. Bethe [6, p. 266] also notes that the external field should satisfy the Lorenz condition. No further details are given. However, at least in various simple cases, a gage transform that kills off the Lorenz condition may be applied. See for example the gage property for a pure external field {A.19.5}. In the classical case a gage transform of the external fields does not make a difference either, because it does not change either the Lagrangian equations for the transverse field nor those for the particles. Using the Lorenz condition cannot hurt, anyway.

Particle spin, if any, is not included in the above Hamiltonian. At nonrelativistic speeds, its energy can be described as a dot produce with the local magnetic field, chapter 13.4.

So far all this was classical electrodynamics. But the interaction between the charges and the transverse waves can readily be quantized using essentially the same procedure as used for the Koulomb potential in subsection A.22.3. The details are worked out in addendum {A.23} for the fields. It allows a relativistic description of the emission of electromagnetic radiation by atoms and nuclei, {A.24} and {A.25}.

While the transverse field must be quantized, the Coulomb potential can be taken unchanged into quantum mechanics. That was done, for example, for the nonrelativistic hydrogen atom in chapter 4.3 and for the relativistic one in addendum {D.81}.

Finally, any external fields are assumed to be given; they are not quantized either.

Note that the Fermi quantization is not fully relativistic. In a fully relativistic theory, the particles too should be described by quantum fields. The Fermi quantization does not do that. So even the relativistic hydrogen atom is not quite exact, even though it is orders of magnitude more accurate than the already very accurate nonrelativistic one. The energy levels are still wrong by the so-called "Lamb shift," {A.39} But this is an extremely tiny effect. Little in life is perfect, isn't it?

A.22.9 The Coulomb potential and the speed of light

The Coulomb potential

$$\varphi_{\rm C}(\vec{r};t) = \sum_{i=1}^{I} \frac{q_i}{4\pi\epsilon_0|\vec{r}-\vec{r}_i|}$$

does not respect the speed of light c. Move a charge, and the Coulomb potential immediately changes everywhere in space. However, special relativity says that an event may not affect events elsewhere unless these events are reachable by the speed of light. Something else must prevent the use of the Coulomb potential to transmit observable effects at a speed greater than that of light.

To understand what is going on, assume that at time zero some charges at the origin are given a well-deserved kick. As mentioned earlier, the Klein-Gordon equations respect the speed of light. Therefore the *original* potentials φ and $\vec{A}$, the ones that satisfied the Klein-Gordon equations and Lorenz condition, are unaffected by the kick beyond a distance ct from the origin. The original potentials do respect the speed of light.

The Coulomb potential above, however, includes the longitudinal part $\vec{A}_\parallel$ of the vector potential $\vec{A}$. As the Coulomb potential reflects, $\vec{A}_\parallel$ does change immediately all the way up to infinity. But the transverse part $\vec{A}_\perp$ also changes immediately all the way up to infinity. Beyond the limit dictated by the speed of light, the two parts of the potential exactly cancel each other. As a result, beyond the speed of light limit, the net vector potential $\vec{A}$ does not change.

The bottom line is

> *The mathematics of the Helmholtz decomposition of $\vec{A}$ into $\vec{A}_\parallel$ and $\vec{A}_\perp$ hides, but of course does not change, the limitation imposed by the speed of light.*

The limitation is still there, it is just much more difficult to see. The change in current density $\vec{\jmath}$ caused by kicking the charges near the origin is restricted to the immediate vicinity of the origin. But both the longitudinal part $\vec{\jmath}_\parallel$ and the transverse part $\vec{\jmath}_\perp$ extend all the way to infinity. And then so do the longitudinal and transverse potentials. It is only when you add the two that you see that the sum is zero beyond the speed of light limit.

A.23 Quantization of radiation

Long ago, the electromagnetic field was described in terms of classical physics by Maxwell, chapter 13. His equations have stood up well to special relativity. However, they need correction for quantum mechanics. According to the Planck-Einstein relation, the electromagnetic field comes in discrete particles of energy

$\hbar\omega$ called photons. A classical electromagnetic field cannot explain that. This addendum will derive the quantum field in empty space. While the description tries to be reasonably self-contained, to really appreciate the details you may have to read some other addenda too. It may also be noted that the discussion here is quite different from what you will find in other sources, {N.12}.

First, representing the electromagnetic field using the photons of quantum mechanics is called "second quantization." No, there is no earlier quantization of the electromagnetic field involved. The word "second" is there for historical reasons. Historically, physicists have found it hysterical to confuse students.

In the quantum description, the electromagnetic field is an *observable* property of photons. And the key assumption of quantum mechanics is that observable properties of particles are the eigenvalues of Hermitian operators, chapter 3. Furthermore, these operators act on wave functions that are associated with the particles.

Therefore, second quantization is basically straightforward. Find the nature of the wave function of photons. Then identify the Hermitian operators that give the observable electromagnetic field.

However, to achieve this in a reasonable manner requires a bit of preparation. To understand photon wave functions, an understanding of a few key concepts of classical electromagnetics is essential. And the Hermitian operators that act on these wave functions are quite different from the typical operators normally used in this book. In particular, they involve operators that create and annihilate photons. Creation and annihilation of particles is a purely relativistic effect, described by so-called quantum field theory.

Also, after the field has been quantized, then of course you want to see what the effects of the quantization really are, in terms of observable quantities.

A.23.1 Properties of classical electromagnetic fields

Classical electromagnetics is discussed in considerable detail in chapter 13.2 and 13.3. Here only a few selected results are needed.

The classical electromagnetic field is a combination of a so-called electric field $\vec{\mathcal{E}}$ and a magnetic field $\vec{\mathcal{B}}$. These are measures for the forces that the field exerts on any charged particles inside the field. The classical expression for the force on a charged particle is the so-called Lorentz force law

$$q\left(\vec{\mathcal{E}} + \vec{v} \times \vec{\mathcal{B}}\right)$$

where q is the charge of the particle and $\vec{v}$ its velocity. However, this is not really important here since quantum mechanics uses neither forces nor velocities.

What is important is that electromagnetic fields carry energy. That is how the sun heats up the surface of the earth. The electromagnetic energy in a

volume $\mathcal{V}$ is given by, chapter 13.2 (13.11):

$$E_{\mathcal{V}} = \tfrac{1}{2}\epsilon_0 \int_{\mathcal{V}} \left(\vec{\mathcal{E}}^2 + c^2 \vec{\mathcal{B}}^2 \right) \mathrm{d}^3\vec{r} \tag{A.152}$$

where $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is called the permittivity of space and $c = 3\ 10^8$ m/s the speed of light. As you might guess, the energy per unit volume is proportional to the square fields. After all, no fields, no energy; also the energy should always be positive. The presence of the permittivity of space is needed to get proper units of energy. The additional factor $\frac{1}{2}$ is not so trivial; it is typically derived from examining the energy stored inside condensers and coils. That sort of detail is outside the scope of this book.

Quantum mechanics is in terms of potentials instead of forces. As already noted in chapter 1.3.2, in classical electromagnetics there is both a scalar potential φ as well as a vector potential $\vec{A}$. In classical electromagnetics these potentials by themselves are not really important. What is important is that their derivatives give the fields. Specifically:

$$\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial \vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla \times \vec{A} \tag{A.153}$$

Here the operator

$$\nabla = \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

is called nabla or del. As an example, for the z components of the fields:

$$\mathcal{E}_z = -\frac{\partial \varphi}{\partial z} - \frac{\partial A_z}{\partial t} \qquad \mathcal{B}_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}$$

Quantum mechanics is all about the potentials. But the potentials are not unique. In particular, for any arbitrary function χ of position and time, you can find two different potentials φ' and $\vec{A}'$ that produce the exact same electric and magnetic fields as φ and $\vec{A}$. These potentials are given by

$$\varphi' = \varphi - \frac{\partial \chi}{\partial t} \qquad \vec{A}' = \vec{A} + \nabla\chi \tag{A.154}$$

(To check that the fields for these potentials are indeed the same, note that $\nabla \times \nabla\chi$ is zero for any function χ.) This indeterminacy in potentials is the famous "gauge property" of the electromagnetic field. The arbitrary function χ is the "gauge function."

Classical relativistic mechanics likes to combine the four scalar potentials in a four-dimensional vector, or four-vector, chapter 1.3.2:

$$\vec{A} = \begin{pmatrix} \varphi/c \\ \vec{A} \end{pmatrix} = \begin{pmatrix} \varphi/c \\ A_x \\ A_y \\ A_z \end{pmatrix}$$

A.23.2 Photon wave functions

The wave function of photons was discussed in addendum {A.21}. A summary of the key results will be given here.

Superficially, the photon wave function $\overrightarrow{A}_\gamma$ takes the same form as an electromagnetic potential four-vector like in the previous subsection. However, where classical potentials are real, the photon wave function is in general complex. And unlike physical potentials, the derivatives of the photon wave function are not physically observable.

Furthermore, to use the photon wave function in an reasonably efficient manner, it is essential to simplify it. The gauge property of the previous subsection implies that the wave function is not unique. So among all the possible alternatives, it is smart to select the simplest. And in empty space, as discussed here, the simplest photon wave function is of the form:

$$\overrightarrow{A}_\gamma = \begin{pmatrix} 0 \\ \vec{A}_\gamma \end{pmatrix} \qquad \nabla \cdot \vec{A}_\gamma = 0$$

The gauge function corresponding to this wave function is called the Coulomb-Lorenz gauge. Seen from a moving coordinate system, this form of the wave function gets messed up, so do not do that.

The real interest is in quantum states of definite energy. Now for a nonrelativistic particle, wave functions with definite energy must be eigenfunctions of the so-called Hamiltonian eigenvalue problem. That eigenvalue problem is also known as the time-independent Schrödinger equation. However, a relativistic particle of zero rest mass like the photon must satisfy a different eigenvalue problem, {A.21}:

$$-\nabla^2 \vec{A}^{\rm e}_\gamma = k^2 \vec{A}^{\rm e}_\gamma \qquad k \equiv \frac{E}{\hbar c} = \frac{p}{\hbar} \qquad E = \hbar\omega \quad p = \hbar k \tag{A.155}$$

Here $\vec{A}^{\rm e}_\gamma$ is the energy eigenfunction. Further p is the magnitude of the linear momentum of the photon. The second-last equation is the so-called Planck-Einstein relation that gives the photon energy E in terms of its frequency ω, while the last equation is the de Broglie relation that gives the photon momentum p in terms of its wave number k. The relation between frequency and wave number is $\omega = kc$ with c the speed of light.

The simplest energy eigenfunctions are those that have definite linear momentum. A typical example of such an energy eigenfunction is

$$\vec{A}^{\rm e}_\gamma = \hat{\imath} e^{{\rm i}ky} \tag{A.156}$$

This wave function has definite linear momentum $\hat{\jmath}\hbar k$, as you can see from applying the components of the linear momentum operator $\hbar\nabla/{\rm i}$ on it. And it is an energy eigenfunction, as you can verify by substitution in (A.155). Further,

since the wave function vector is in the z-direction, it is called linearly polarized in the z-direction.

Now the photon wave function $\vec{\hookrightarrow}_\gamma$ is not a classical electromagnetic potential. The "electric and magnetic fields" $\vec{\mathcal{E}}_\gamma$ and $\vec{\mathcal{B}}_\gamma$ that you would find by using the classical expressions (A.153) are not physically observable quantities. So they do not have to obey the classical expression (A.152) for the energy in an electromagnetic field.

However, you make life a lot simpler for yourself if you normalize the photon wave functions so that they *do* satisfy it. That produces a normalized wave function and corresponding unobservable fields of the form, {A.21}:

$$\vec{A}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}kc}\vec{A}^{\rm e}_\gamma \qquad \vec{\mathcal{E}}^{\rm n}_\gamma = \varepsilon_k \vec{A}^{\rm e}_\gamma \qquad c\vec{\mathcal{B}}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}k}\nabla \times \vec{A}^{\rm e}_\gamma \tag{A.157}$$

where the constant ε_k is found by substitution into the normalization condition:

$$\tfrac{1}{2}\epsilon_0 \int_{\rm all} \left(\left|\vec{\mathcal{E}}^{\rm n}_\gamma\right|^2 + c^2 \left|\vec{\mathcal{B}}^{\rm n}_\gamma\right|^2 \right) {\rm d}^3\vec{r} = \hbar\omega \tag{A.158}$$

Consider how this works out for the example eigenfunction of definite linear momentum mentioned above. That eigenfunction cannot be normalized in infinite space since it does not go to zero at large distances. To normalize it, you have to assume that the electromagnetic field is confined to a big periodic box of volume $\mathcal{V}$. In that case the normalized eigenfunction and unobservable fields become:

$$\vec{A}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}kc}\hat{k}e^{{\rm i}ky} \qquad \vec{\mathcal{E}}^{\rm n}_\gamma = \varepsilon_k \hat{k}e^{{\rm i}ky} \qquad c\vec{\mathcal{B}}^{\rm n}_\gamma = \varepsilon_k \hat{\imath}e^{{\rm i}ky} \qquad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}} \tag{A.159}$$

Another interesting example is given in {A.21.6}. It is a photon state with definite angular momentum $\hbar$ in the direction of motion. Such a photon state is called right-circularly polarized. In this example the linear momentum is taken to be in the z-direction, rather than the y-direction. The normalized wave function and unobservable fields are:

$$\vec{A}^{\rm n}_\gamma = \frac{\varepsilon_k}{{\rm i}kc}\frac{\hat{\imath}+{\rm i}\hat{\jmath}}{\sqrt{2}}e^{{\rm i}kz} \quad \vec{\mathcal{E}}^{\rm n}_\gamma = \varepsilon_k \frac{\hat{\imath}+{\rm i}\hat{\jmath}}{\sqrt{2}}e^{{\rm i}kz} \quad c\vec{\mathcal{B}}^{\rm n}_\gamma = \varepsilon_k \frac{-{\rm i}\hat{\imath}+\hat{\jmath}}{\sqrt{2}}e^{{\rm i}kz} \quad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}} \tag{A.160}$$

It will be interesting to see what the observable fields for this photon state look like.

A.23.3 The electromagnetic operators

The previous subsection has identified the form of the wave function of a photon in an energy eigenstate. The next step is to identify the Hamiltonian operators of the observable electric and magnetic fields.

But first there is a problem. If you have exactly one photon, the wave functions as discussed in the previous subsection would do just fine. If you had exactly two photons, you could readily write a wave function for them too. But a state with an exact number i of photons is an energy eigenstate. It has energy $i\hbar\omega$, taking the zero of energy as the state of no photons. Energy eigenstates are stationary, chapter 7.1.4. They never change. All the interesting mechanics in nature is due to uncertainty in energy. As far as photons are concerned, that requires uncertainty in the number of photons. And there is no way to write a wave function for an uncertain number of particles in classical quantum mechanics. The mathematical machinery is simply not up to it.

The mathematics of quantum field theory is needed, as discussed in addendum {A.15}. The key concepts will be briefly summarized here. The mathematics starts with Fock space kets. Consider a single energy eigenfunction for photons, like one of the examples given in the previous subsection. The Fock space ket

$$|i\rangle$$

indicates the wave function if there are exactly i photons in the considered energy eigenfunction. The number i is called the occupation number of the state. (If more than one photon state is of interest, an occupation number is added to the ket for each state. However, the discussion here will stay restricted to a single state.)

The Fock space ket formalism allows wave functions to be written for any number of particles in the state. And by taking linear combinations of kets with different occupation numbers, uncertainty in the number of photons can be described. So uncertainty in energy can be described.

Kets are taken to be orthonormal. The inner product $\langle i_1|i_2\rangle$ of two kets with different occupation numbers i_1 and i_2 is zero. The inner product of a ket with itself is taken to be one. That is exactly the same as for energy eigenfunctions in classical quantum mechanics,

Next, it turns out that operators that act on photon wave functions are intrinsically linked to operators that annihilate and create photons. Mathematically, at least. These operators are defined by the relations

$$\widehat{a}|i\rangle = \sqrt{i}|i{-}1\rangle \qquad \widehat{a}^\dagger|i{-}1\rangle = \sqrt{i}|i\rangle \tag{A.161}$$

for any number of photons i. In words, the annihilation operator $\widehat{a}$ takes a state of i photons and turns it into a state with one less photon. The creation operator $\widehat{a}^\dagger$ puts the photon back in. The scalar factors $\sqrt{i}$ are a matter of convenience. If you did not put them in here, you would have to do it elsewhere.

At first it may seem just weird that there are physical operators like that. But a bit more thought may make it more plausible. First of all, nature pretty much forces the Fock space kets on us. Classical quantum mechanics would like to number particles such as photons just like classical physics likes to do: photon

1, photon 2, ... But nature makes a farce out of that with the symmetrization requirement. It allows absolutely no difference in the way one photon occupies a state compared to another one. Indeed, nature goes to such a length of preventing us to, God forbid, make a distinction between one photon and another that she puts every single photon in the universe partly in every single microscopic photon state on earth. Now Fock space kets are the only way to express *how many* photons are in a given state without saying *which* photons. And if the symbols of nature are then apparently Fock space kets, the operators of nature are pretty unavoidably annihilation and creation operators. There is not much else you can do with Fock space kets than change the number of particles.

The annihilation and creation operators are not Hermitian. They cannot be taken unchanged to the other side of an inner product of kets. However, they are Hermitian conjugates: they change into each other when taken to the other side of an inner product:

$$\left\langle |i_2\rangle \middle| \widehat{a}|i_1\rangle \right\rangle = \left\langle \widehat{a}^\dagger|i_2\rangle \middle| |i_1\rangle \right\rangle \equiv \langle i_2|\widehat{a}|i_1\rangle \quad \left\langle |i_1\rangle \middle| \widehat{a}^\dagger|i_2\rangle \right\rangle = \left\langle \widehat{a}|i_1\rangle \middle| |i_2\rangle \right\rangle \equiv \langle i_1|\widehat{a}^\dagger|i_2\rangle$$

To see this, note that the inner products are only nonzero if $i_2 = i_1 - 1$ because of the orthogonality of kets with different numbers of photons. And if $i_2 = i_1 - 1$, then (A.161) shows that all four inner products above equal $\sqrt{i_1}$.

That is important because it shows that Hermitian operators can be formed from combinations of the two operators. For example, $\widehat{a} + \widehat{a}^\dagger$ is a Hermitian operator. Each of the two operators changes into the other when taken to the other side of an inner product. So the sum stays unchanged. More generally, if c is any real or complex number, $\widehat{a}c + \widehat{a}^\dagger c^*$ is Hermitian.

And that then is the basic recipe for finding the operators of the observable electric and magnetic fields. Take $\widehat{a}$ times the unobservable field of the normalized photon state, (A.157) with (A.158) and (A.155). Add the Hermitian conjugate of that. And put in the usual $1/\sqrt{2}$ factor of quantum mechanics for averaging states. In total

$$\boxed{\widehat{\vec{\mathcal{E}}} = \frac{1}{\sqrt{2}}\left(\widehat{a}\vec{\mathcal{E}}^{\rm n}_\gamma + \widehat{a}^\dagger \vec{\mathcal{E}}^{\rm n*}_\gamma\right) \qquad \widehat{\vec{\mathcal{B}}} = \frac{1}{\sqrt{2}}\left(\widehat{a}\vec{\mathcal{B}}^{\rm n}_\gamma + \widehat{a}^\dagger \vec{\mathcal{B}}^{\rm n*}_\gamma\right)} \tag{A.162}$$

You might wonder why there are two terms in the operators, one with a complex conjugate wave function. Mathematically that is definitely needed to get Hermitian operators. That in turn is needed to get real observed fields. But what does it mean physically? One way of thinking about it is that the observed field is real because it does not just involve an interaction with an $e^{{\rm i}(\vec{k}\cdot\vec{r}-\omega t)}$ photon, but also with an $e^{-{\rm i}(\vec{k}\cdot\vec{r}-\omega t)}$ antiphoton.

Of course, just because the above operators are Hermitian does not prove that they are the right ones for the observable electric and magnetic fields. Unfortunately, there is no straightforward way to deduce quantum mechanics

operators from mere knowledge of the classical approximation. Vice-versa is not a problem: given the operators, it is fairly straightforward to deduce the corresponding classical equations for a macroscopic system. It is much like at the start of this book, where it was postulated that the momentum of a particle corresponds to the operator $\hbar\partial/\mathrm{i}\partial x$. That was a leap of faith. However, it was eventually seen that it did produce the correct classical momentum for macroscopic systems, chapter 7.2.1 and 7.10, as well as correct quantum results like the energy levels of the hydrogen atom, chapter 4.3. A similar leap of faith will be needed to quantize the electromagnetic field.

A.23.4 Properties of the observable electromagnetic field

The previous subsection postulated the operators (A.162) for the observable electric and magnetic fields. This subsection will examine the consequences of these operators, in order to gain confidence in them. And to learn something about the effects of quantization of the electromagnetic field.

Consider first a simple wave function where there are exactly i photons in the considered photon state. In terms of Fock space kets, the wave function is then:

$$\Psi = c_i e^{-\mathrm{i}i\omega t}|i\rangle$$

where c_i is a constant with magnitude 1. This follows the Schrödinger rule that the time dependence in a wave function of definite energy E is given by $e^{-\mathrm{i}Et/\hbar}$, with in this case $E = i\hbar\omega$.

The expectation value of the electric field at a given position and time is then by definition

$$\langle\vec{\mathcal{E}}\rangle = \left\langle\Psi\middle|\widehat{\vec{\mathcal{E}}}\Psi\right\rangle = \frac{1}{\sqrt{2}}\langle i|e^{\mathrm{i}i\omega t}c_i^*\Big|\left(\widehat{a}\vec{\mathcal{E}}_\gamma^{\,\mathrm{n}} + \widehat{a}^\dagger\vec{\mathcal{E}}_\gamma^{\,\mathrm{n}*}\right)c_i e^{-\mathrm{i}i\omega t}|i\rangle$$

That is zero because $\widehat{a}$ and $\widehat{a}^\dagger$ turn the right hand ket into $|i{-}1\rangle$ respectively $|i{+}1\rangle$. These are orthogonal to the left hand $\langle i|$. The same way, the expectation magnetic field will be zero too.

Oops.

Zero electric and magnetic fields were not exactly expected if there is a nonzero number of photons present.

No panic please. This is an energy eigenstate. Often these do not resemble classical physics at all. Think of a hydrogen atom in its ground state. The expectation value of the linear momentum of the electron is zero in that state. That is just like the electric and magnetic fields are zero here. But the expectation value of the *square* momentum of the hydrogen electron is not zero. In fact, that gives the nonzero expectation value of the kinetic energy of the electron, 13.6 eV. So maybe the square fields need to be examined here.

Come to think of it, the first thing to check should obviously have been the energy. It better be $i\hbar\omega$ for i photons. Following the Newtonian analogy for the classical energy integral (A.152), the Hamiltonian should be

$$H = \tfrac{1}{2}\epsilon_0 \int_{\text{all}} \left(\widehat{\vec{\mathcal{E}}}^2 + c^2 \widehat{\vec{\mathcal{B}}}^2 \right) \mathrm{d}^3\vec{r}$$

This can be greatly simplified by plugging in the given expressions for the operators and identifying the integrals, {D.40}:

$$\boxed{H = \tfrac{1}{2}\hbar\omega(\widehat{a}^\dagger\widehat{a} + \widehat{a}\widehat{a}^\dagger)} \tag{A.163}$$

Now apply this Hamiltonian on a state with i photons. First, using the definitions (A.161) of the annihilation and creation operators

$$\widehat{a}^\dagger\widehat{a}|i\rangle = \widehat{a}^\dagger(\sqrt{i}|i{-}1\rangle) = i|i\rangle \qquad \widehat{a}\widehat{a}^\dagger|i\rangle = \widehat{a}(\sqrt{i+1}|i{+}1\rangle) = (i+1)|i\rangle$$

This shows that the operators $\widehat{a}$ and $\widehat{a}^\dagger$ do not commute; their order makes a difference. In particular, according to the above their commutator equals

$$[\widehat{a}, \widehat{a}^\dagger] \equiv \widehat{a}\widehat{a}^\dagger - \widehat{a}^\dagger\widehat{a} = 1 \tag{A.164}$$

Anyway, using the above relations the expression for the Hamiltonian applied on a Fock space ket becomes

$$H|i\rangle = \hbar\omega(i + \tfrac{1}{2})|i\rangle$$

The factor in front of the final ket is the energy eigenvalue. It is more or less like expected, which was $i\hbar\omega$ for i photons. But there is another one-half photon worth of energy.

That, however, may be correct. It closely resembles what happened for the harmonic oscillator, chapter 4.1. Apparently the energy in the electromagnetic field is never zero, just like a harmonic oscillator is never at rest. The energy increases by $\hbar\omega$ for each additional photon as it should.

Actually, the half photon "vacuum energy" is somewhat of a problem. If you start summing these half photons over all infinitely many frequencies, you end up, of course, with infinity. Now the ground state energy does not affect the dynamics. But if you do "measurements" of the electric or magnetic fields in vacuum, you will get nonzero values. So apparently there is real energy there. Presumably that should affect gravity. Maybe the effect would not be infinite, if you cut off the sum at frequencies at which quantum mechanics might fail, but it should certainly be extremely dramatic. So why is it not observed? The answer is unknown. See chapter 8.7 for one suggestion.

The vacuum energy also has consequences if you place two conducting plates extremely closely together. The conducting plates restrict the vacuum field

between the plates. (Or at least the relatively low energy part of it. Beyond say the X-ray range photons will not notice the plates.) Because of the restriction of the plates, you would expect the vacuum energy to be less than expected. Because of energy conservation, that must mean that there is an attractive force between the plates. That is the so-called "Casimir force." This weird force has actually been measured experimentally. Once again it is seen that the half photon of vacuum energy in each state is not just a mathematical artifact.

Because of the infinite energy, some authors describe the vacuum as a "seething cauldron" of electromagnetic waves. These authors may not be aware that the vacuum state, being a ground state, is stationary. Or they may not have access to a dictionary of the English language.

The next test of the field operators is to reconsider the expectation electric field when there is uncertainty in energy. Also remember to add another half photon of energy now. Then the general wave function takes the form:

$$\Psi = \sum_i c_i e^{-\mathrm{i}(i+\frac{1}{2})\omega t}|i\rangle$$

The expectation value of the electric field follows as

$$\langle\vec{\mathcal{E}}\rangle = \sum_i \sum_j \frac{1}{\sqrt{2}}\langle i|e^{\mathrm{i}(i+\frac{1}{2})\omega t}c_i^*\left|\left(\widehat{a}\vec{\mathcal{E}}_\gamma^{\,\mathrm{n}} + \widehat{a}\vec{\mathcal{E}}_\gamma^{\,\mathrm{n}*}\right)c_j e^{-\mathrm{i}(j+\frac{1}{2})\omega t}|j\rangle\right.$$

Using the definitions of the annihilation and creation operators and the orthonormality of the kets, this can be worked out further to

$$\langle\vec{\mathcal{E}}\rangle = Ce^{-\mathrm{i}\omega t}\vec{\mathcal{E}}_\gamma^{\,\mathrm{n}} + C^*e^{\mathrm{i}\omega t}\vec{\mathcal{E}}_\gamma^{\,\mathrm{n}*} \qquad C \equiv \frac{1}{\sqrt{2}}\sum_i c_{i-1}^* c_i\sqrt{i} \tag{A.165}$$

Well, the field is no longer zero. Note that the first term in the electric field is more or less what you would expect from the unobservable field of a single photon. But the observable field adds the complex conjugate. That makes the observable field real.

The properties of the observable fields can now be determined. For example, consider the photon wave function (A.159) given earlier. This wave function had its linear momentum in the y-direction. It was "linearly polarized" in the z-direction. According to the above expression, the observable electric field is:

$$\langle\vec{\mathcal{E}}\rangle = \hat{k}\varepsilon_k\left(Ce^{\mathrm{i}(ky-\omega t)} + C^*e^{-\mathrm{i}(ky-\omega t)}\right)$$

The first term is roughly what you would want to write down for the unobservable electric field of a single photon. The second term, however, is the complex conjugate of that. It makes the observable field real. Writing C in the form $|C|e^{\mathrm{i}\alpha}$ and using the Euler formula 2.5 to clean up gives:

$$\langle\vec{\mathcal{E}}\rangle = \hat{k}2\varepsilon_k|C|\cos(ky-\omega t+\alpha)$$

That is a real electromagnetic wave. It is still polarized in the z-direction, and it travels in the y-direction.

The corresponding magnetic field goes exactly the same way. The only difference in (A.159) is that $\hat{k}$ gets replaced by $\hat{\imath}$. Therefore

$$\langle c\vec{\mathcal{B}}\rangle = \hat{\imath}2\varepsilon_k|C|\cos(ky - \omega t + \alpha)$$

Note that like for the photon wave function, the observable fields are normal to the direction of wave propagation, and to each other.

As another example, consider the "circularly polarized" photon wave function (A.160). This wave function had its linear momentum in the z-direction, and it had definite angular momentum $\hbar$ around the z-axis. Here the observable fields are found to be

$$\langle\vec{\mathcal{E}}\rangle = \sqrt{2}\varepsilon_k|C|\left[\hat{\imath}\cos(kz - \omega t + \alpha) - \hat{\jmath}\sin(kz - \omega t + \alpha)\right]$$

$$\langle c\vec{\mathcal{B}}\rangle = \sqrt{2}\varepsilon_k|C|\left[\hat{\imath}\sin(kz - \omega t + \alpha) + \hat{\jmath}\cos(kz - \omega t + \alpha)\right]$$

Like for linearly polarized light, the electric and magnetic fields are normal to the direction of wave propagation and to each other. But here the electric and magnetic field vectors rotate around in a circle when seen at a fixed position z. Seen at a fixed time, the end points of the electric vectors that start from the z-axis form a helix. And so do the magnetic ones.

The final question is under what conditions you would get a classical electromagnetic field with relatively little quantum uncertainty. To answer that, first note that the square quantum uncertainty is given by

$$\sigma_{\vec{\mathcal{E}}}{}^2 = \langle\vec{\mathcal{E}}^2\rangle - \langle\vec{\mathcal{E}}\rangle^2$$

(This is the square of chapter 4.4.3 (4.44) multiplied out and identified.)

To evaluate this uncertainty requires the expectation value of the square electric field. That can be found much like the expectation value (A.165) of the electric field itself. The answer is

$$\langle\vec{\mathcal{E}}^2\rangle = 2D_0|\vec{\mathcal{E}}^{\rm n}_\gamma|^2 + D_1 e^{-2{\rm i}\omega t}(\vec{\mathcal{E}}^{\rm n}_\gamma)^2 + D_1^* e^{2{\rm i}\omega t}(\vec{\mathcal{E}}^{\rm n*}_\gamma)^2$$

where

$$D_0 \equiv \frac{1}{2}\sum_i |c_i|^2(i+\tfrac{1}{2}) \qquad D_1 \equiv \frac{1}{2}\sum_i c_{i-1}^* c_{i+1}\sqrt{i}\sqrt{i+1}$$

Note that when this is substituted into the integral (A.152) for the energy, the D_0 term gives half the expectation value of the energy. In particular, the coefficient D_0 itself is half the expectation value of the $i+\frac{1}{2}$ number of photons of energy. The other half comes from the corresponding term in the magnetic field. The D_1 terms above integrate away against the corresponding terms in the magnetic field, {D.40}.

To determine the uncertainty in the electric field, it is convenient to write the expectation square electric field above in real form. To do so, the coefficient D_1 is written in the form $|D_1|e^{2\mathrm{i}\beta}$. Also, the square unobservable electric field $(\vec{\mathcal{E}}^{\mathrm{n}}_\gamma)^2$ is written in the form $|\vec{\mathcal{E}}^{\mathrm{n}}_\gamma|^2 e^{2\mathrm{i}\gamma}$. Here γ will normally depend on position; for example $\gamma = ky$ for the given example of linearly polarized light.

Then the expectation square electric field becomes, using the Euler formula (2.5) and some trig,

$$\langle \vec{\mathcal{E}}^2 \rangle = 2(D_0 - |D_1|)|\vec{\mathcal{E}}^{\mathrm{n}}_\gamma|^2 + 4|D_1||\vec{\mathcal{E}}^{\mathrm{n}}_\gamma|^2 \cos^2(\gamma - \omega t + \beta)$$

with D_0 and $D_1 = |D_1|e^{2\mathrm{i}\beta}$ as given above. Similarly the square of the expectation electric field, as given earlier in (A.165), can be written as

$$\langle \vec{\mathcal{E}} \rangle^2 = 4|C|^2|\vec{\mathcal{E}}^{\mathrm{n}}_\gamma|^2 \cos^2(\gamma - \omega t + \alpha) \qquad C = |C|e^{\mathrm{i}\alpha} = \frac{1}{\sqrt{2}} \sum_i c^*_{i-1} c_i \sqrt{i}$$

For a field without quantum uncertainty, $\langle \vec{\mathcal{E}}^2 \rangle$ and $\langle \vec{\mathcal{E}} \rangle^2$ as given above must be equal. Note that first of all this requires that $|D_1| = D_0$, because otherwise $\langle \vec{\mathcal{E}}^2 \rangle$ does not become zero periodically like $\langle \vec{\mathcal{E}} \rangle^2$ does. Also β will have to be α, up to a whole multiple of π, otherwise the zeros are not at the same times. Finally, $|C|$ will have to be equal to D_0 too, or the amplitudes will not be the same.

However, regardless of uncertainty, the coefficients must always satisfy

$$|D_1| \leqslant D_0 \qquad |C|^2 \leqslant \tfrac{1}{2}(D_0 + |D_1|)$$

The first inequality applies because otherwise $\langle \vec{\mathcal{E}}^2 \rangle$ would become negative whenever the cosine is zero. The second applies because $\langle \vec{\mathcal{E}} \rangle^2$ cannot be larger than $\langle \vec{\mathcal{E}}^2 \rangle$; the square uncertainty cannot be negative. For quantum certainty then, the above relations must become equalities. However, a careful analysis shows that they cannot become equalities, {D.40}.

So there is always some quantum uncertainty left. Maximum uncertainty occurs when the number of photons has a definite value. Then $D_1 = C = 0$.

If there is always at least some uncertainty, the real question is under what conditions it is relatively small. Analysis shows that the uncertainty in the fields is small under the following conditions, {D.40}:

- The expectation value of the number of photons $\langle i \rangle$ is large enough.
- Nonnegligible coefficients c_i are limited to a relatively small range around $i = \langle i \rangle$. However, that range must still contain a relatively large number of coefficients.
- The coefficients for successive values of i are different by a factor that is approximately equal to $e^{\mathrm{i}\alpha}$.

In that case classical electric and magnetic field result with little quantum uncertainty. Note that the above conditions apply for photons restricted to a single quantum state. In a real electromagnetic field, many quantum states would be occupied and things would be much messier still.

It may also be noted that the above conditions bear a striking resemblance to the conditions that produce a particle with a fairly coherent position and momentum in classical quantum mechanics, chapter 7.10.

A.24 Quantum spontaneous emission

Chapter 7.8 explained the general interaction between atoms and electromagnetic fields. However, spontaneous emission of radiation was found using a dirty trick due to Einstein. He peeked at the solution for blackbody radiation. This addendum will give a proper quantum description. Warning: while this addendum tries to be reasonably self-contained, to really appreciate the details you may have to read some other addenda too.

The problem with the descriptions of emission and absorption of radiation in chapter 7.7 and 7.8 is that they assume that the electromagnetic field is given. The electromagnetic field is not given; it changes by one photon. That is rather important for spontaneous emission, where it changes from no photons to one photon. To account for that correctly requires that the electromagnetic field is properly quantized. That is done in this note.

To keep it simple, it will be assumed that the atom is a hydrogen one. Then there is just one electron to worry about. (The general analysis can be found in {A.25}). The hydrogen atom is initially in some high energy state $\psi_{\rm H}$. Then it emits a photon and transitions to a lower energy state $\psi_{\rm L}$. The emitted photon comes out in a state with energy

$$E_\gamma = \hbar\omega \approx E_{\rm H} - E_{\rm L}$$

Recall that the photon energy is given in terms of its frequency ω by the Planck-Einstein relation. This photon energy is approximately the difference between the atomic energies. It does not have to be exact; there can be some energy slop, chapter 7.6.1.

Only a single photon energy state needs to be considered at a time. At the end of the story, the results can be summed over all possible photon states. To allow for stimulated emission, it will be assumed that initially there may already be i preexisting photons present. For spontaneous emission, $i = 0$. The initial system state will be indicated as:

$$\psi_1 = \psi_{\rm H}|i\rangle$$

Here the so-called Fock space ket $|i\rangle$ is simply a concise way of indicating that there are i photons in the considered photon quantum state.

In the final state the atom has decayed to a lower energy state $\psi_{\rm L}$. In doing so it has released 1 more photon into the considered photon state. So the final wave function is

$$\psi_2 = \psi_{\rm L}|i{+}1\rangle$$

The key to the emission process is now the set of Hamiltonian coefficients, chapter 7.6,

$$\langle E_1\rangle = \langle\psi_1|H\psi_1\rangle \qquad H_{21} = \langle\psi_2|H\psi_1\rangle \qquad \langle E_2\rangle = \langle\psi_2|H\psi_2\rangle$$

Here H is the Hamiltonian. All that really needs to be done in this note is to identify these coefficients, and in particular the so-called matrix element H_{21}. With the matrix element known, Fermi's golden rule can be used to find the precise transition rate, chapter 7.6.1.

To identify the Hamiltonian coefficients, first the Hamiltonian must be identified. Recall that the Hamiltonian is the operator of the total energy of the system. It will take the form

$$H = H_{\rm atom} + H_\gamma + H_{\rm atom,\gamma}$$

The first term in the right hand side is the inherent energy of the hydrogen atom. This Hamiltonian was written down way back in chapter 4.3. However, its precise form is of no interest here. The second term in the right hand side is the energy in the electromagnetic field. Electromagnetic fields too have inherent energy, about $\hbar\omega$ per photon in fact. The third term is the energy of the interaction between the atomic electron and the electromagnetic field.

Unlike the first term in the Hamiltonian, the other two are inherently relativistic: the number of photons is hardly a conserved quantity. Photons are readily created or absorbed by a charged particle, like the electron here. And it turns out that Hamiltonians for photons are intrinsically linked to operators that annihilate and create photons. Mathematically, at least. These operators are defined by the relations

$$\widehat{a}|i\rangle = \sqrt{i}|i{-}1\rangle \qquad \widehat{a}^\dagger|i{-}1\rangle = \sqrt{i}|i\rangle \tag{A.166}$$

for any number of photons i. In words, the annihilation operator $\widehat{a}$ takes a state of i photons and turns it into a state with one less photon. The creation operator $\widehat{a}^\dagger$ puts the photon back in. The scalar factors $\sqrt{i}$ are a matter of convenience. If you did not put them in here, you would have to do it elsewhere.

The Hamiltonian that describes the inherent energy in the electromagnetic field turns out to be, {A.23},

$$H_\gamma = \tfrac{1}{2}\hbar\omega(\widehat{a}^\dagger\widehat{a} + \widehat{a}\widehat{a}^\dagger)$$

As a sanity check, this Hamiltonian can be applied on a state of i photons. Using the definitions of the annihilation and creation operators given above,

$$H_\gamma|i\rangle = \tfrac{1}{2}\hbar\omega(\widehat{a}^\dagger\sqrt{i}|i{-}1\rangle + \widehat{a}\sqrt{i+1}|i{+}1\rangle) = \tfrac{1}{2}\hbar\omega(i|i\rangle + (i+1)|i\rangle) = \hbar\omega(i+\tfrac{1}{2})|i\rangle$$

The factor in front of the final ket is the energy eigenvalue. So the energy in the field increases by one unit $\hbar\omega$ for each photon added, exactly as it should. The additional half photon is the ground state energy of the electromagnetic field. Even in its ground state, the electromagnetic field has some energy left. That is much like a one-dimensional harmonic oscillator still has $\frac{1}{2}\hbar\omega$ of energy left in its ground state, chapter 4.1.

Finally the energy of the interaction between the electron and electromagnetic field is needed. This third part of the total Hamiltonian is the messiest. To keep it as simple as possible, it will assumed that the transition is of the normal electric dipole type. In such transitions the electron interacts only with the electric part of the electromagnetic field. In addition, just like in the analysis of chapter 7.7.1 using a classical electromagnetic field, it will be assumed that the electric field is in the z-direction and propagates in the y-direction. (The general multipole analysis can be found in {A.25}).

Now recall that in quantum mechanics, observable properties of particles are the eigenvalues of Hermitian operators, chapter 3.3. For example, the observable values of linear momentum of an electron in the y-direction are the eigenvalues of the linear momentum operator $\widehat{p}_y = \hbar\partial/\mathrm{i}\partial y$. This operator acts on the electron wave function.

Similarly, the electric field $\mathcal{E}_z$ that the electron interacts with is an *observable* property of the corresponding photons. So the observable values of the electric field must be the eigenvalues of a Hermitian electric field operator $\widehat{\mathcal{E}}_z$. And this operator acts on photon wave functions.

In the analysis using a classical electromagnetic field, the energy of interaction between the electron and the electromagnetic field was taken to be approximately $e\mathcal{E}_z z$. That is similar to the mgh potential of a particle due to gravity. The electron electric charge $-e$ takes the place of the mass m, the electric field $\mathcal{E}_z$ that of the acceleration of gravity g, and z that of the height h. Using the quantized electric field, there is no given classical field $\mathcal{E}_z$, and instead the operator $\widehat{\mathcal{E}}_z$ must be used:

$$H_{\mathrm{atom},\gamma} = e\widehat{\mathcal{E}}_z z$$

The operator $\widehat{\mathcal{E}}_z$ acts on the ket part of the combined atom-photon wave function. (And, although you may not think of it that way, the factor z is really an operator that acts on the electron wave function part. That is true even in the analysis using the classical field.)

The electric field operator $\widehat{\mathcal{E}}_z$ can be identified from the appropriate photon wave function. The photon wave function here is assumed to have its linear momentum in the y-direction and its unobservable electric field in the z-direction. The corresponding normalized wave function and unobservable electric field were given in {A.21.6} (A.95):

$$\vec{A}^{\mathrm{n}}_\gamma = \frac{\varepsilon_k}{\mathrm{i}kc}\hat{k}e^{\mathrm{i}ky} \qquad \vec{\mathcal{E}}^{\mathrm{n}}_\gamma = \varepsilon_k\hat{k}e^{\mathrm{i}ky} \qquad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}}$$

Here ϵ_0 is the permittivity of space. Also $\mathcal{V}$ is the volume of the large periodic box in which the entire system will be assumed to be located. In truly infinite space the analysis would be extremely messy, littered with ugly delta functions.

The rules to get the operator of the observable electric field were discussed in addendum {A.23}. First the unobservable electric field above is multiplied by the annihilation operator, then the Hermitian conjugate of that product is added, and the sum is divided by $\sqrt{2}$:

$$\widehat{\mathcal{E}}_z = \frac{\varepsilon_k}{\sqrt{2}}(\widehat{a}e^{\mathrm{i}ky} + \widehat{a}^\dagger e^{-\mathrm{i}ky})$$

(Note that for the usual Schrödinger approach followed here, time dependence is described by the wave function. Most sources switch here to a Heisenberg approach where the time-dependence is pushed into the operators. There is however no particular need to do so.)

In the electric dipole approximation, it is assumed that the atom is so small compared to the wave length of the photon that ky can be assumed to be zero. Therefore

$$\widehat{\mathcal{E}}_z = \frac{\varepsilon_k}{\sqrt{2}}(\widehat{a} + \widehat{a}^\dagger)$$

The combined Hamiltonian is then

$$H = H_{\rm atom} + H_\gamma + e\frac{\varepsilon_k}{\sqrt{2}}(\widehat{a} + \widehat{a}^\dagger)z$$

with the first two terms as described earlier.

Next the Hamiltonian matrix coefficients are needed. The first one is

$$\langle E_1\rangle = \langle\psi_1|H\psi_1\rangle = \langle i|\psi_{\rm H}\Big|\Big(H_{\rm atom} + H_\gamma + e\frac{\varepsilon_k}{\sqrt{2}}(\widehat{a} + \widehat{a}^\dagger)z\Big)\psi_{\rm H}|i\rangle$$

Now the atomic part of the Hamiltonian produces a mere factor $E_{\rm H}$ when it acts on the atomic part of the right hand wave function. Further, as discussed above, the electromagnetic Hamiltonian produces a factor $(i+\frac{1}{2})\hbar\omega$ when it acts on the right hand ket. Finally the interaction part of the Hamiltonian does not produce a contribution. One way to see that is from the atomic inner product. The atomic inner product is zero because negative values of z integrate away against positive ones. Another way to see it is from the electromagnetic inner product. The operators $\widehat{a}$ and $\widehat{a}^\dagger$ produce states $|i{-}1\rangle$ respectively $|i{+}1\rangle$ when they act on the right hand ket. And those are orthogonal to the left hand ket; inner products between kets with different numbers of photons are zero. Kets are by definition orthonormal.

All together then

$$\langle E_1\rangle = E_{\rm H} + (i+\tfrac{1}{2})\hbar\omega$$

The same way

$$\langle E_2\rangle = E_{\rm L} + (i+1+\tfrac{1}{2})\hbar\omega$$

Finally the matrix element:

$$H_{21} = \langle\psi_2|H\psi_1\rangle = \langle i+1|\psi_{\rm L}\Big|\Big(H_{\rm atom} + H_\gamma + e\frac{\varepsilon_k}{\sqrt{2}}(\widehat{a}+\widehat{a}^\dagger)z\Big)\psi_{\rm H}|i\rangle$$

In this case the atomic part of the Hamiltonian produces zero. The reason is that this Hamiltonian produces a simple scalar factor $E_{\rm H}$ when it acts on the right hand state. It leaves the state $\psi_{\rm H}$ itself unchanged. And this state produces zero in an inner product with the atomic state $\psi_{\rm L}$; energy eigenstates are orthonormal. Similarly, the electromagnetic Hamiltonian produces zero. It leaves the ket $|i\rangle$ in the right hand wave wave function unchanged, and that is orthogonal to the left hand $\langle i+1|$. However, in this case the interaction Hamiltonian produces a nonzero contribution:

$$H_{21} = \frac{\varepsilon_k\sqrt{i+1}}{\sqrt{2}}\langle\psi_{\rm L}|ez\psi_{\rm H}\rangle$$

The reason is that the creation operator $\widehat{a}^\dagger$ acting on the right hand ket produces a multiple $\sqrt{i+1}$ times the left hand ket. The remaining inner product $\langle\psi_{\rm L}|ez\psi_{\rm H}\rangle$ is called the "atomic matrix element," as it only depends on what the atomic states are.

The task laid out in chapter 7.6.1 has been accomplished: the relativistic matrix element has been found. A final expression for the spontaneous emission rate can now be determined.

Before doing so, however, it is good to first compare the obtained result with that of chapter 7.7.1. That section used a classical given electromagnetic field, not a quantized one. So the comparison will show up the effect of the quantization of the electromagnetic field. The section defined a modified matrix element

$$\overline{H}_{21} = H_{21}e^{{\rm i}(\langle E_2\rangle - \langle E_1\rangle)t/\hbar}$$

This matrix element determined the entire evolution of the system. For the quantized electric field discussed here, this coefficient works out to be

$$\overline{H}_{21} = \frac{\varepsilon_k\sqrt{i+1}}{\sqrt{2}}\langle\psi_{\rm L}|ez\psi_{\rm H}\rangle e^{{\rm i}(\omega-\omega_0)t} \qquad \varepsilon_k = \sqrt{\frac{\hbar\omega}{\epsilon_0\mathcal{V}}} \tag{A.167}$$

where $\omega_0 = (E_{\rm H} - E_{\rm L})/\hbar$.

That is essentially the same form as for the classical field. Recall that the second term in (7.44) for the classical field can be ignored. The first term is the same as above, within a constant. To see the real difference in the constants, note that the transition probability is proportional to the square magnitude of the matrix element. The square magnitudes are:

$$\text{quantized: } |\overline{H}_{21}^2| = \frac{(i+1)\hbar\omega}{2\epsilon_0\mathcal{V}}|\langle\psi_{\rm L}|ez\psi_{\rm H}\rangle|^2 \quad \text{classical: } |\overline{H}_{21}^2| = \frac{\mathcal{E}_{\rm f}^2}{4}|\langle\psi_{\rm L}|ez\psi_{\rm H}\rangle|^2$$

Now if there is a large number i of photons in the state, the two expressions are approximately the same. The electromagnetic energy of the wave according to classical physics, $\epsilon_0 \mathcal{E}_{\rm f}^2 \mathcal{V}/2$, {A.23}, is then approximately the number of photons $i \approx i+1$ times $\hbar\omega$.

But for spontaneous emission there is a big difference. In that case, classical physics would take the initial electromagnetic field $\mathcal{E}_{\rm f}$ to be zero. And that then implies that the atom stays in the excited state $\psi_{\rm H}$ for always. There is no electromagnetic field to move it out of the state. So there is no spontaneous emission.

Instead quantum mechanics takes the initial field to have $i = 0$ photons. But note the square matrix element above. It is not zero! The matrix element is as if there is still a full photon left in the electromagnetic field. So spontaneous emission can and does occur in the quantized electromagnetic field. Also, as noted in chapter 7.8, one full photon is exactly what is needed to explain spontaneous emission. Einstein's A coefficient has been found using pure quantum analysis. Without peeking at the black body spectrum.

That can also be seen without detouring through the messy analysis of chapter 7.7 and 7.8. To find the spontaneous emission rate directly, the matrix element above can be plugged into Fermi's Golden Rule (7.38) of chapter 7.6.1. The density of states needed in it was given earlier in chapter 6.3 (6.7) and 6.19. Do note that these modes include all directions of the electric field, not just the z-direction. To account for that, you need to average the square atomic matrix element over all three Cartesian directions. That produces the spontaneous transition rate

$$\frac{\omega^3}{\pi\hbar c^3 \epsilon_0} \frac{|\langle \psi_{\rm L}|ex\psi_{\rm H}\rangle|^2 + |\langle \psi_{\rm L}|ey\psi_{\rm H}\rangle|^2 + |\langle \psi_{\rm L}|ez\psi_{\rm H}\rangle|^2}{3}$$

The above result is the same as Einstein's, (7.47) and (7.48). (To see why a simple average works in the final term, first note that it is obviously the right average for photons with axial linear momenta and fields. Then note that the average is independent of the angular orientation of the axis system in which the photons are described. So it also works for photons that are axial in any rotated coordinate system. To verify that the average is independent of angular orientation does not really require linear algebra; it suffices to show that it is true for rotation about one axis, say the z-axis.)

Some additional observations may be interesting. You might think of the spontaneous emission as caused by excitation from the ground state electromagnetic field. But as seen earlier, the actual energy of the ground state is half a photon, not one photon. And the zero level of energy should not affect the dynamics anyway. According to the analysis here, spontaneous emission is a twilight effect, chapter 5.3. The Hamiltonian coefficient H_{21} is the energy if the atom is not excited and there is a photon if the atom is excited and there is no

photon. In quantum mechanics, the twilight term allows the excited atom to interact with the photon *that would be there if it was not excited.* Sic.

A.25 Multipole transitions

This addendum gives a description of the multipole interaction between atoms or nuclei and electromagnetic fields. In particular, the spontaneous emission of a photon of electromagnetic radiation in an atomic or nuclear transition will be examined. But stimulated emission and absorption are only trivially different.

The basic ideas were already worked out in earlier addenda, especially in {A.21} on photon wave functions and {A.24} on spontaneous emission. However, these addenda left the actual interaction between the atom or nucleus and the field largely unspecified. Only a very simple form of the interaction, called the electric dipole approximation, was worked out there.

Many transitions are not possible by the electric dipole mechanism. This addendum will describe the more general multipole interaction mechanisms. That will allow rough estimates of how fast various possible transitions occur. These will include the Weisskopf and Moszkowski estimates for the gamma decay of nuclei. It will also allow a general description exactly how the selection rules of chapter 7.4.4 relate to nuclear and photon wave functions.

The overall picture is that before the transition, the atom or nucleus is in a high energy state $\psi_{\rm H}$. Then it transitions to a lower energy state $\psi_{\rm L}$. During the transition it emits a photon that carries away the excess energy. The energy of that photon is related to its frequency ω by the Planck-Einstein relation:

$$E_{\rm H} - E_{\rm L} = \hbar\omega_0 \approx \hbar\omega$$

Here ω_0 is the nominal frequency of the photon. The actual photon frequency ω might be slightly different; there can be some slop in energy conservation. However, that will be taken care of by using Fermi's golden rule, chapter 7.6.1.

It is often useful to express the photon frequency in terms of the so-called wave number k:

$$\omega = kc$$

Here c is the speed of light. The wave number is a physically important quantity since it is inversely proportional to the wave length of the photon. If the typical size of the atom or nucleus is R, then kR is an nondimensional quantity. It describes the ratio of atom or nucleus size to photon wave length. Normally this ratio is very small, which allows helpful simplifications.

It will be assumed that only the electrons need to be considered for atomic transitions. The nucleus is too heavy to move much in such transitions. For nuclear transitions, (inside the nuclei), it is usually necessary to consider both types of nucleons, protons and neutrons. Protons and neutrons will be treated as point particles, though each is really a combination of three quarks.

As noted in chapter 7.5.3 and 7.6.1, the "driving force" in a transition is the so-called Hamiltonian matrix element:

$$H_{21} = \langle \psi_{\rm L} | H | \psi_{\rm H} \rangle$$

Here H is the Hamiltonian, which will depend on the type of transition. In particular, it depends on the properties of the emitted photon.

If the matrix element H_{21} is zero, transitions of that type are not possible. The transition is "forbidden." If the matrix element is very small, they will be very slow. (If the term "forbidden" is used without qualification, it indicates that the electric-dipole type of transition cannot occur,)

A.25.1 Approximate Hamiltonian

The big ideas in multipole transitions are most clearly seen using a simple model. That model will be explained in this subsection. However, the results in this subsection will not be quantitatively correct for multipole transitions of higher order. Later subsections will correct these deficiencies. This two-step approach is followed because otherwise it can be easy to get lost in all the mathematics of multipole transitions. Also, the terminology used in multipole transitions really arises from the simple model discussed here. And in any case, the needed corrections will turn out to be very simple.

An electromagnetic wave consists of an electric field $\vec{\mathcal{E}}$ and a magnetic field $\vec{\mathcal{B}}$. A basic plane wave takes the form, (13.10):

$$\vec{\mathcal{E}} = \hat{\imath}\sqrt{2}\mathcal{E}_0 \cos\Big(kz - \omega t - \alpha_0\Big) \qquad \vec{\mathcal{B}} = \hat{\jmath}\frac{\sqrt{2}\mathcal{E}_0}{c} \cos\Big(kz - \omega t - \alpha_0\Big)$$

For convenience the z-axis was taken in the direction of propagation of the wave. Also the x-axis was taken in the direction of the electric field. The constant c is the speed of light and the constant $\mathcal{E}_0$ is the root-mean-square value of the electric field. (The amplitude of the electric field is then $\sqrt{2}\mathcal{E}_0$, but the root mean square value is more closely related to what you end up with when the electromagnetic field is properly quantized.) Finally α_0 is some unimportant phase angle.

The above waves need to be written as complex exponentials using the Euler formula (2.5):

$$\vec{\mathcal{E}} = \hat{\imath}\frac{\mathcal{E}_0}{\sqrt{2}}\Big(e^{{\rm i}(kz-\omega t-\alpha_0)} + e^{-{\rm i}(kz-\omega t-\alpha_0)}\Big) \qquad \vec{\mathcal{B}} = \hat{\jmath}\frac{\mathcal{E}_0}{\sqrt{2}c}\Big(e^{{\rm i}(kz-\omega t-\alpha_0)} + e^{-{\rm i}(kz-\omega t-\alpha_0)}\Big)$$

Only one of the two exponentials will turn out to be relevant to the transition process. For absorption that is the first exponential. But for emission, the case discussed here, the second exponential applies.

There are different ways to see why only one exponential is relevant. Chapter 7.7 follows a classical approach in which the field is given. In that case, the evolution equation that gives the transition probability is, {D.38},

$$\mathrm{i}\hbar\dot{c}_2 \approx H_{21} e^{\mathrm{i}(E_2-E_1)t/\hbar}$$

Here $|c_2|^2$ is the transition probability. For emission, the final state is the low energy state. Then the Planck-Einstein relation gives the exponential above as $e^{-\mathrm{i}\omega_0 t}$. (By convention, frequencies are taken to be positive.) Now the Hamiltonian matrix element H_{21} will involve the electric and magnetic fields, with their exponentials. The first exponentials, combined with the exponential above, produce a time-dependent factor $e^{-\mathrm{i}(\omega_0+\omega)t}$. Since normal photon frequencies are large, this factor oscillates extremely rapidly in time. Because of these oscillations, the corresponding terms never produce a significant contribution to the transition probability. Opposite contributions average away against each other. So the first exponentials can be ignored. But the second exponentials produce a time dependent factor $e^{-\mathrm{i}(\omega_0-\omega)t}$. That does not oscillate rapidly provided that the emitted photon has frequency $\omega \approx \omega_0$. So such photons can achieve a significant probability of being emitted.

For absorption, the low energy state is the first one, instead of the second. That makes the exponential above $e^{+\mathrm{i}\omega_0 t}$, and the entire story inverts.

The better way to see that the first exponentials in the fields drop out is to quantize the electromagnetic field. This book covers that only in the addenda. In particular, addendum {A.24} described the process. Fortunately, quantization of the electromagnetic field is mainly important to figure out the right value of the constant $\mathcal{E}_0$ to use, especially for spontaneous emission. It does not directly affect the actual analysis in this addendum. In particular the conclusion remains that only the second exponentials survive.

The bottom line is that for emission

$$\vec{\mathcal{E}} = \hat{\imath}\frac{\mathcal{E}_0}{\sqrt{2}} e^{-\mathrm{i}(kz-\omega t-\alpha_0)} \qquad \vec{\mathcal{B}} = \hat{\jmath}\frac{\mathcal{E}_0}{\sqrt{2}c} e^{-\mathrm{i}(kz-\omega t-\alpha_0)} \tag{A.168}$$

Also, as far as this addendum is concerned, the difference between spontaneous and stimulated emission is only in the value of the constant $\mathcal{E}_0$.

Next the Hamiltonian is needed. For the matrix element, only the part of the Hamiltonian that describes the interaction between the atom or nucleus and the electromagnetic fields is relevant, {A.24}. (Recall that the matrix element drives the transition process; no interaction means no transition.) Assume that the electrons in the atom, or the protons and neutrons in the nucleus, are numbered using an index i. Then by approximation the interaction Hamiltonian of a single particle i with the electromagnetic field is

$$H_i \approx -q_i \vec{\mathcal{E}}_i \cdot \vec{r}_i - \frac{q_i}{2m_i}\vec{\mathcal{B}}_i \cdot \widehat{\vec{L}}_i - \frac{q_i}{2m_i} g_i \vec{\mathcal{B}}_i \cdot \widehat{\vec{S}}_i$$

In general, you will need to sum this over all particles i. But the discussion here will usually look at one particle at a time.

The first term in the Hamiltonian above is like the mgh potential of gravity, with the particle charge q_i taking the place of the mass m, the electric field that of the acceleration of gravity g, and the particle position $\vec{r}_i$ that of the height h.

The second and third terms in the Hamiltonian are due to the fact that a charged particle that is going around in circles acts as a little electromagnet. An electromagnet wants to align itself with an ambient magnetic field. That is just like a compass needle aligns itself with the magnetic field of earth.

This effect shows up as soon as there is angular momentum. Indeed, the operator $\widehat{\vec{L}}_i$ above is the orbital angular momentum of the particle and $\widehat{\vec{S}}_i$ is the spin. The factor g_i is a nondimensional number that describes the relative efficiency of the particle spin in creating an electromagnetic response. For an electron in an atom, g_i is very close to 2. That is a theoretical value expected for fundamental particles, chapter 13.4. However, for a proton in a nucleus the value is about 5.6, assuming that the effect of the surrounding protons and neutrons can be ignored. (Actually, it is quite well established that normally the surrounding particles *cannot* be ignored. But it is difficult to say what value for g_i to use instead, except that it will surely be smaller than 5.6, and greater than 2.)

A special case needs to be made for the neutrons in a nucleus. Since the neutron has no charge, $q_i = 0$, you would expect that its contribution to the Hamiltonian is zero. However, the final term in the Hamiltonian is *not* zero. A neutron has a magnetic response. (A neutron consists of three charged quarks. The combined charge of the three is zero, but the combined magnetic response is not.) To account for that, in the final term, you need to use the charge e and mass $m_{\rm p}$ of the *proton*, and take g_i about -3.8. This value of g_i ignores again the effects of surrounding protons and neutrons.

There are additional issues that are important. Often it is assumed that in a transition only a single particle changes states. If that particle is a neutron, it might then seem that the first two terms in the Hamiltonian can be ignored. But actually, the neutron and the rest of the nucleus move around their common center of gravity. And the rest of the nucleus is charged. So normally the first two terms cannot be ignored. This is mainly important for the so-called electric dipole transitions; for higher multipole orders, the electromagnetic field is very small near the origin, and the motion of the rest of the nucleus does not produce much effect. In a transition of a single proton, you may also want to correct the first term for the motion of the rest of the nucleus. But also note that the rest of the nucleus is not really a point particle. That may make a significant difference for higher multipole orders. Therefore simple corrections remain problematic. See [33] and [11] for further discussion of these nontrivial issues.

The given Hamiltonian ignores the fact that the electric and magnetic fields are unsteady and not uniform. That is the reason why the higher multipoles found in the next subsection will not be quite right. They will be good enough to show the basic ideas however. And the quantitative problems will be corrected in later subsections.

A.25.2 Approximate multipole matrix elements

The last step is to write down the matrix element. Substituting the approximate Hamiltonian and fields of the previous subsection into the matrix element of the introduction gives:

$$H_{21,i} = \langle \psi_{\rm L} | H | \psi_{\rm H} \rangle \approx -\frac{\mathcal{E}_0}{\sqrt{2}} \langle \psi_{\rm L} | e^{-{\rm i}kz_i} [q_i x_i + (q_i/2m_i c)(\widehat{L}_{i,y} + g_i \widehat{S}_{i,y})] | \psi_{\rm H} \rangle$$

This will normally need to be summed over all electrons i in the atom, or all nucleons i in the nucleus. Note that the time dependent part of the exponential is of no interest. It will in fact not even appear when the electromagnetic field is properly quantized, {A.24}. In a classical treatment, it drops out versus the $e^{{\rm i}(E_2-E_1)t/\hbar}$ exponential mentioned in the previous subsection.

To split the above matrix element into different multipole orders, write the exponential as a Taylor series:

$$e^{-{\rm i}kz_i} = \sum_{n=0}^{\infty} \frac{(-{\rm i}kz_i)^n}{n!} = \sum_{\ell=1}^{\infty} \frac{(-{\rm i}kz_i)^{\ell-1}}{(\ell-1)!}$$

In the second equality, the summation index was renotated as $n = \ell - 1$. The reason is that ℓ turns out to be what is conventionally defined as the multipole order.

Using this Taylor series, the matrix element gets split into separate electric and magnetic multipole contributions:

$$H_{21,i} = \sum_{\ell=1}^{\infty} H_{21,i}^{{\rm E}\ell} + H_{21,i}^{{\rm M}\ell} = H_{21,i}^{\rm E1} + H_{21,i}^{\rm M1} + H_{21,i}^{\rm E2} + H_{21,i}^{\rm M2} + \ldots$$

$$H_{21,i}^{{\rm E}\ell} \approx -\frac{q_i \mathcal{E}_0}{\sqrt{2}} \frac{(-{\rm i}k)^{\ell-1}}{(\ell-1)!} \langle \psi_{\rm L} | z_i^{\ell-1} x_i | \psi_{\rm H} \rangle$$

$$H_{21,i}^{{\rm M}\ell} \approx -\frac{q_i \mathcal{E}_0}{2\sqrt{2} m_i c} \frac{(-{\rm i}k)^{\ell-1}}{(\ell-1)!} \langle \psi_{\rm L} | z_i^{\ell-1} (\widehat{L}_{i,y} + g_i \widehat{S}_{i.y}) | \psi_{\rm H} \rangle$$

The terms with $\ell = 1$ are the dipole ones, $\ell = 2$ the quadrupole ones, 3 the octupole ones, 4 the hexadecapole ones, etcetera. Superscript E indicates an electric contribution, M a magnetic one. The first contribution that is nonzero gives the lowest multipole order that is allowed.

A.25.3 Corrected multipole matrix elements

The multipole matrix elements of the previous subsection were rough approximations. The reason was the approximate Hamiltonian that was used. This subsection will describe the corrections needed to fix them up. It will still be assumed that the atomic or nuclear particles involved are nonrelativistic. They usually are.

The corrected Hamiltonian is

$$\boxed{H = \sum_i \left[\frac{1}{2m_i} \left(\widehat{\vec{p}}_i - q_i \vec{A}_i \right)^2 + q_i \varphi_i - g_i \frac{q_i}{2m_i} \vec{\mathcal{B}}_i \cdot \widehat{\vec{S}}_i \right] + V} \tag{A.169}$$

where the sum is over the individual electrons in the atom or the protons and neutrons in the nucleus. In the sum, m_i is the mass of the particle, $\widehat{\vec{p}}_i$ its momentum, and q_i its charge. The potential V is the usual potential that keeps the particle inside the atom or nucleus. The remaining parts in the Hamiltonian express the effect of the additional external electromagnetic field. In particular, φ_i is the electrostatic potential of the field and $\vec{A}_i$ the so-called vector potential, each evaluated at the particle position. Finally

$$\vec{\mathcal{B}} = \nabla \times \vec{A}$$

is the magnetic part of the field. The spin $\widehat{\vec{S}}_i$ of the particle interacts with this field at the location of the particle, with a relative strength given by the nondimensional constant g_i. See chapter 1.3.2 for a classical justification of this Hamiltonian, or chapter 13 for a quantum one.

Nonrelativistically, the spin does not interact with the electric field. That is particularly limiting for the neutron, which has no net charge to interact with the electric field. In reality, a rapidly moving particle with spin will also interact with the electric field, {A.39}. See the Dirac equation and in particular {D.74} for a relativistic description of the interaction of spin with an electromagnetic field. That would be too messy to include here, but it can be found in [44]. Note also that since in reality the neutron consists of three quarks, that should allow it to interact directly with a nonuniform electric field.

If the field is quantized, you will also want to include the Hamiltonian of the field in the total Hamiltonian above. And the field quantities become operators. That goes the same way as in {A.24}. It makes no real difference for the analysis in this addendum.

It is always possible, and a very good idea, to take the unperturbed electromagnetic potentials so that

$$\varphi = 0 \qquad \nabla \cdot \vec{A} = 0$$

See for example the addendum on photon wave functions {A.21} for more on that. That addendum also gives the potentials that correspond to photons

of definite linear, respectively angular momentum. These will be used in this addendum.

The square in the above Hamiltonian may be multiplied out to give

$$H = H_0 + \sum_i \left[-\frac{q_i}{m_i} \vec{A}_i \cdot \widehat{\vec{p}}_i + \frac{q_i^2}{2m_i} \vec{A}_i^2 - g_i \frac{q_i}{2m_i} \vec{\mathcal{B}}_i \cdot \widehat{\vec{S}}_i \right]$$

The term H_0 is the Hamiltonian of the atom or nucleus in the absence of interaction with the external electromagnetic field. Like in the previous subsection, it is not directly relevant to the interaction with the electromagnetic field. Note further that $\widehat{\vec{p}}$ and $\vec{A}$ commute because $\nabla \cdot \vec{A}$ is zero. The term proportional to $\vec{A}^2$ will be ignored as it is normally very small. (It gives rise to two-photon emission, [33].)

That makes the interaction Hamiltonian of a single particle i equal to

$$\boxed{H_i = -\frac{q_i}{m_i} \vec{A}_i \cdot \widehat{\vec{p}}_i - g_i \frac{q_i}{2m_i} \vec{\mathcal{B}}_i \cdot \widehat{\vec{S}}_i} \tag{A.170}$$

Note that the final spin term has not changed from the approximate Hamiltonian written down earlier. However, the first term appears completely different from before. Still, there must obviously be a connection.

To find that connection requires considerable manipulation. First the vector potential $\vec{A}$ must be identified in terms of the simple electromagnetic wave as written down earlier in (A.168). To do so, note that the vector potential must be related to the fields as

$$\vec{\mathcal{E}} = -\frac{\partial \vec{A}}{\partial t} \qquad \vec{\mathcal{B}} = \nabla \times \vec{A}$$

See, for example, {A.21} for a discussion. That allows the vector potential corresponding to the simple wave (A.168) to be identified as:

$$\vec{A} = -\hat{\imath} \frac{\mathcal{E}_0}{\sqrt{2}\mathrm{i}\omega} e^{-\mathrm{i}(kz - \omega t - \alpha_0)}$$

This wave can be generalized to allow general directions of wave propagation and fields. That gives:

$$\vec{A} = -\hat{\imath}_{\mathcal{E}} \frac{\mathcal{E}_0}{\sqrt{2}\mathrm{i}\omega} e^{-\mathrm{i}(\vec{k}\cdot\vec{r} - \omega t - \alpha_0)} \qquad \vec{\mathcal{E}} = \hat{\imath}_{\mathcal{E}} \frac{\mathcal{E}_0}{\sqrt{2}} e^{-\mathrm{i}(\vec{k}\cdot\vec{r} - \omega t - \alpha_0)} \qquad \vec{\mathcal{B}} = \hat{\imath}_{\mathcal{B}} \frac{\mathcal{E}_0}{\sqrt{2}c} e^{-\mathrm{i}(\vec{k}\cdot\vec{r} - \omega t - \alpha_0)}$$

Here the unit vector $\hat{\imath}_{\mathcal{E}}$ is in the direction of the electric field and $\hat{\imath}_{\mathcal{B}}$ in the direction of the magnetic field. A unit vector $\hat{\imath}_k$ in the direction of wave propagation can be defined as their cross product. This defines the wave number vector as

$$\vec{k} \equiv \hat{\imath}_k k \qquad \hat{\imath}_k = \hat{\imath}_{\mathcal{E}} \times \hat{\imath}_{\mathcal{B}} \qquad \hat{\imath}_{\mathcal{E}} \cdot \hat{\imath}_{\mathcal{B}} = 0$$

The three unit vectors are orthonormal. Note that for a given direction of wave propagation $\hat{\imath}_k$, there will be two independent waves. They differ in the direction of the electric field $\hat{\imath}_\mathcal{E}$. The choice for the direction of the electric field for first wave is not unique; the field must merely be orthogonal to the direction of wave propagation. An arbitrary choice must be made. The electric field of the second wave needs to be orthogonal to that of the first wave. The example in the previous subsections took the wave propagation in the z-direction, $\hat{\imath}_k = \hat{k}$, and the electric field in the x-direction, $\hat{\imath}_\mathcal{E} = \hat{\imath}$, to give the magnetic field in the y-direction, $\hat{\imath}_\mathcal{B} = \hat{\jmath}$. In that case the second independent wave will have its electric field in the y-direction, $\hat{\imath}_\mathcal{E} = \hat{\jmath}$, and its magnetic field in the negative x-direction, $\hat{\imath}_\mathcal{B} = -\hat{\imath}$.

The single-particle matrix element is now, dropping again the time-dependent factors,

$$\begin{aligned} H_{21,i} &= \langle\psi_\mathrm{L}|H_i|\psi_\mathrm{H}\rangle \\ &= \frac{q_i}{m_i}\frac{\mathcal{E}_0}{\sqrt{2}\mathrm{i}\omega}\langle\psi_\mathrm{L}|e^{-\mathrm{i}\vec{k}\cdot\vec{r}}\hat{\imath}_\mathcal{E}\cdot\widehat{\vec{p}}_i|\psi_\mathrm{H}\rangle - g_i\frac{q_i}{2m_i}\frac{\mathcal{E}_0}{\sqrt{2}c}\langle\psi_\mathrm{L}|e^{-\mathrm{i}\vec{k}\cdot\vec{r}}\hat{\imath}_\mathcal{B}\cdot\widehat{\vec{S}}_i|\psi_\mathrm{H}\rangle \end{aligned}$$

The first term needs to be cleaned up to make sense out of it. That is an extremely messy exercise, banned to {D.43}.

However, the result is much like before:

$$H_{21,i} = \sum_{\ell=1}^{\infty} H_{21,i}^{\mathrm{E}\ell} + H_{21,i}^{\mathrm{M}\ell} = H_{21,i}^{\mathrm{E1}} + H_{21,i}^{\mathrm{M1}} + H_{21,i}^{\mathrm{E2}} + H_{21,i}^{\mathrm{M2}} + \ldots$$

where

$$\boxed{H_{21,i}^{\mathrm{E}\ell} = -\frac{q_i\mathcal{E}_0}{\sqrt{2}}\frac{(-\mathrm{i}k)^{\ell-1}}{(\ell-1)!}\langle\psi_\mathrm{L}|\frac{1}{\ell}r_{i,k}^{\ell-1}r_{i,\mathcal{E}}|\psi_\mathrm{H}\rangle} \tag{A.171}$$

$$\boxed{H_{21,i}^{\mathrm{M}\ell} \approx -\frac{q_i\mathcal{E}_0}{2\sqrt{2}m_ic}\frac{(-\mathrm{i}k)^{\ell-1}}{(\ell-1)!}\langle\psi_\mathrm{L}|r_{i,k}^{\ell-1}\Big(\frac{2}{\ell+1}\widehat{L}_{i,\mathcal{B}} + g_i\widehat{S}_{i,\mathcal{B}}\Big)|\psi_\mathrm{H}\rangle} \tag{A.172}$$

Here $r_{i,k}$ is the component of the position of the particle in the direction of motion. Similarly, $r_{i,\mathcal{E}}$ is the component of position in the direction of the electric field, while the angular momentum components are in the direction of the magnetic field.

This can now be compared to the earlier results using the approximate Hamiltonian. Those earlier results assumed the special case that the wave propagation was in the z-direction and had its electric field in the x-direction. In that case,

$$\text{example:} \qquad r_{i,k} = z_i \qquad r_{i,\mathcal{E}} = x_i \qquad \widehat{L}_{i,\mathcal{B}} = \widehat{L}_{i.y} \qquad \widehat{S}_{i,\mathcal{B}} = \widehat{S}_{i,y} \tag{A.173}$$

Noting that, it is seen that the correct electric contributions only differ from the approximate ones by a simple factor $1/\ell$. This factor is 1 for electric dipole contributions, so these were correct already. Similarly, the magnetic contribution

differs only by the additional factor $2/(\ell+1)$ for the orbital angular momentum from the approximate result. This factor is 1 for magnetic dipole contributions. So these too were already correct.

However, there is a problem with the electric contribution in the case of nuclei. A nuclear potential does not just depend on the position of the nuclear particles, but also on their momentum. That introduces an additional term in the electric contribution, {D.43}. A ballpark for that term shows that this may well make the listed electric contribution quantitatively invalid, {N.14}. Unfortunately, nuclear potentials are not known to sufficient accuracy to give a solid prediction for the contribution. In the following, this problem will usually simply be ignored, like other textbooks do.

A.25.4 Matrix element ballparks

Recall that electromagnetic transitions are driven by the matrix element. The previous subsection managed to split the matrix element into separate electric and magnetic multipole contributions. The intent in this subsection is now to show that normally, the first nonzero multipole contribution is the important one. Subsequent multipole contributions are normally small compared to the first nonzero one.

To do so, this subsection will ballpark the multipole contributions. The ballparks will show that the magnitude of the contributions decreases rapidly with increasing multipole order ℓ.

But of course ballparks are only that. If a contribution is exactly zero for some special reason, (usually a symmetry), then the ballpark is going to be wrong. That is why it is the first *nonzero* multipole contribution that is important, rather than simply the first one. The next subsection will discuss the so-called selection rules that determine when contributions are zero.

The ballparks are formulated in terms a typical size R of the atom or nucleus. For the present purposes, this size will be taken to be the average radial position of the particles away from the center of atom or nucleus. Then the magnitudes of the electric multipole contributions can be written as

$$|H^{\mathrm{E}\ell}_{21,i}| = \frac{|q_i|\mathcal{E}_0 R}{\sqrt{2}} \frac{(kR)^{\ell-1}}{\ell(\ell-1)!} |\langle \psi_{\mathrm{L}} | (r_{i,k}/R)^{\ell-1} (r_{i,\mathcal{E}}/R) | \psi_{\mathrm{H}} \rangle|$$

There is no easy way to say exactly what the inner product above will be. However, since the positions inside it have been scaled with the mean radius R, its value is supposedly some normal finite number. Unless the inner product happens to be zero for some special reason of course. Assuming that this does not happen, the inner product can be ignored for the ballpark. And that then shows that each higher nonzero electric multipole contribution is smaller than the previous one by a factor kR. Now k is inversely proportional to the wave length of the photon that is emitted or absorbed. This wave length is normally

very much larger than the size of the atom or nucleus R. That means that kR is very small. And that then implies that a nonzero multipole contribution at a higher value of ℓ will be very much less than one at a lower value. So contributions for values of ℓ higher than the first nonzero one can normally be ignored.

The magnitudes of the magnetic contributions can be written as

$$|H^{\mathrm{M}\ell}_{21,i}| \approx -\frac{|q_i|\mathcal{E}_0 R}{\sqrt{2}} \frac{(kR)^{\ell-1}}{\ell(\ell-1)!} |\langle \psi_{\mathrm{L}} | (r_{i,k}/R)^{\ell-1} \frac{1}{\hbar} \Big(\frac{2}{\ell+1} \widehat{L}_{i,\mathcal{B}} + g_i \widehat{S}_{i,\mathcal{B}} \Big) | \psi_{\mathrm{H}} \rangle | \ell \frac{\hbar}{2m_i c R}$$

Recall that angular momentum values are multiples of $\hbar$. Therefore the matrix element can again be ballparked as some finite number, if nonzero. So once again, the multipole contributions get smaller by a factor kR for each increase in order. That means that the nonzero magnetic contributions too decrease rapidly with ℓ.

That leaves the question how magnetic contributions compare to electric ones. First compare a magnetic multipole term to the electric one of the same multipole order ℓ. The above estimates show that the magnetic term is mainly different from the electric one by the factor

$$\frac{\hbar}{2m_i c R} \approx \begin{cases} \text{atoms:} & \dfrac{1 \text{ Å}}{500\, R} \\ \text{nuclei:} & \dfrac{1 \text{ fm}}{10\, R} \end{cases}$$

Atomic sizes are in the order of an Ångstrom, and nuclear ones in the order of a few femtometers. So ballpark magnetic contributions are small compared to electric ones of the same order ℓ. And more so for atoms than for nuclei. (Transition rates are proportional to the square of the first nonzero contribution. So the ballpark transition rate for a magnetic transition is smaller than an electric one of the same order by the square of the above factor.)

A somewhat more physical interpretation of the above factor can be given:

$$\frac{\hbar}{2m_i c R} = \sqrt{\frac{T_{\mathrm{bp}}}{2m_i c^2}} \qquad T_{\mathrm{bp}} \equiv \frac{\hbar^2}{2m_i R^2}$$

Here T_{bp} is a ballpark for the kinetic energy $-\hbar^2\nabla^2/2m_i$ of the particle. Note that this ballpark is exact for the hydrogen atom ground state if you take the Bohr radius as the average radius R of the atom. However, for heavier atoms and nuclei, this ballpark may be low: it ignores the exclusion effects of the other particles. Further $m_i c^2$ is the rest mass energy of the particle. Now protons and neutrons in nuclei, and at least the outer electrons in atoms are nonrelativistic; their kinetic energy is much less than their rest mass energy. It follows again that magnetic contributions are normally much smaller than electric ones of the same multipole order.

Compare the magnetic multipole term also to the electric one of the next multipole order. The trailing factor in the magnetic element can for this case be written as

$$\frac{\hbar}{2m_i cR} = kR\frac{T_{\rm bp}}{\hbar\omega}$$

The denominator in the final ratio is the energy of the emitted or absorbed photon. Typically, it is significantly less than the ballpark kinetic energy of the particle. That then makes magnetic matrix elements significantly larger than electric ones of the next-higher multipole order. Though smaller than the electric ones of the same order.

A.25.5 Selection rules

Based on the ballparks given in the previous subsection, the E1 electric dipole contribution should dominate transitions. It should be followed in size by the M1 magnetic dipole one, followed by the E2 electric quadrupole one, etcetera.

But this order gets modified because matrix elements are very often zero for special reasons. This was explained physically in chapter 7.4.4 based on the angular momentum properties of the emitted photon. This subsection will instead relate it directly to the matrix element contributions as identified in subsection A.25.3. To simplify the reasoning, it will again be assumed that the z-axis is chosen in the direction of wave motion and the y-axis in the direction of the electric field. So (A.173) applies for the multipole contributions (A.171) and (A.172).

Consider first the electric dipole contribution $H^{\rm E1}_{21,i}$. According to (A.171) and (A.173) this contribution contains the inner product

$$\langle \psi_{\rm L} | x_i | \psi_{\rm H} \rangle$$

Why would this be zero? Basically because in the inner product integrals, positive values of x_i might exactly integrate away against corresponding negative values. That can happen because of symmetries in the nuclear wave functions.

One such symmetry is parity. For all practical purposes, atomic and nuclear states have definite parity. If the positive directions of the Cartesian axes are inverted, atomic and nuclear states either stay the same (parity 1 or positive), or change sign (parity -1 or negative). Assume, for example, that $\psi_{\rm L}$ and $\psi_{\rm H}$ have both positive parity. That means that they do not change under an inversion of the axes. But the factor x_i in the inner product above has odd parity: axes inversion replaces x_i by $-x_i$. So the complete inner product above changes sign under axes inversion. But inner products are defined in a way that they do *not* change under axes inversion. (In terms of chapter 2.3, the effect of the axes inversion can be undone by a further inversion of the integration variables.) Something can only change sign and still stay the same if it is zero, (-0 is 0 but say -5 is not 5).

So if both $\psi_{\rm L}$ and $\psi_{\rm H}$ have positive parity, the electric dipole contribution is zero. The only way to get a nonzero inner product is if exactly one of $\psi_{\rm L}$ and $\psi_{\rm H}$ has negative parity. Then the factor -1 that this state picks up under axes inversion cancels the -1 from x_i, leaving the inner product unchanged as it should. So the conclusion is that in electric dipole transitions $\psi_{\rm L}$ and $\psi_{\rm H}$ must have opposite parities. In other words, the atomic or nuclear parity must "flip over" in the transition. This condition is called the parity "selection rule" for an electric dipole transition. If it is not satisfied, the electric dipole contribution is zero and a different contribution will dominate. That contribution will be much smaller than a typical nonzero electric dipole one, so the transition will be much slower.

The $H^{\rm M1}_{21,i}$ magnetic dipole contribution contains the inner product

$$\langle \psi_{\rm L} | \widehat{L}_{i,y} + g_i \widehat{S}_{i.y} | \psi_{\rm H} \rangle$$

The angular momentum operators do nothing under axes inversion. One way to see that is to think of $\psi_{\rm H}$ as written in terms of states of definite y-momentum. Then the angular momentum operators merely add scalar factors $m\hbar$ to those states. These do not affect what happens to the remainder of the inner product under axes inversion. Alternatively, note that $\widehat{L}_y = \hbar(z\partial/\partial x - x\partial/\partial z)/\mathrm{i}$ and each term has two position coordinates that change sign. And surely spin should behave the same as orbital angular momentum.

If the angular momentum operators do nothing under axes inversion, the parities of the initial and final atomic or nuclear states will have to be equal. So the parity selection rule for magnetic dipole transitions is the opposite from the one for electric dipole transitions. The parity has to stay the same in the transition.

Assuming again that the wave motion is in the z-direction, each higher multipole order ℓ adds a factor z_i to the electric or magnetic inner product. This factor changes sign under axes inversion. So for increasing ℓ, alternatingly the atomic or nuclear parity must flip over or stay the same.

If the parity selection rule is violated for a multipole term, the term is zero. However, if it is not violated, the term may still be zero for some other reason. The most important other reason is angular momentum. Atomic and nuclear states have definite angular momentum. Consider again the electric dipole inner product

$$\langle \psi_{\rm L} | x_i | \psi_{\rm H} \rangle$$

States of different angular momentum are orthogonal. That is a consequence of the fact that the momentum operators are Hermitian. What it means is that the inner product above is zero unless $x_i\psi_{\rm H}$ has at least some probability of having the same angular momentum as state $\psi_{\rm L}$. Now the factor x_i can be written in

terms of spherical harmonics using chapter 4.2.3, table 4.3:

$$x_i = \sqrt{\frac{8\pi}{3}} r_i \left(Y_1^{-1} - Y_1^1\right)$$

So it is a sum of two states, both with square angular momentum quantum number $l_x = 1$, but with z angular momentum quantum number $m_x = -1$, respectively 1.

Now recall the rules from chapter 7.4.2 for combining angular momenta:

$$Y_1^{-1}\psi_{\rm H} \qquad \Longrightarrow \qquad j_{\rm net} = j_{\rm H} - 1,\ j_{\rm H},\ \text{or } j_{\rm H}+1 \qquad m_{\rm net} = m_{\rm H} - 1$$

Here $j_{\rm H}$ is the quantum number of the square angular momentum of the atomic or nuclear state $\psi_{\rm H}$. And $m_{\rm H}$ is the quantum number of the z angular momentum of the state. Similarly $j_{\rm net}$ and $m_{\rm net}$ are the possible values for the quantum numbers of the combined state $Y_1^{-1}\psi_{\rm H}$. Note again that m_x and $m_{\rm H}$ values simply add together. However, the $j_{\rm H}$-value changes by up to $l_x = 1$ unit in either direction. (But if $j_{\rm H} = 0$, the combined state cannot have zero angular momentum.)

(It should be noted that you should be careful in combining these angular momenta. The normal rules for combining angular momenta apply to different sources of angular momentum. Here the factor x_i does not describe an additional source of angular momentum, but a particle that already has been given an angular momentum within the wave function $\psi_{\rm H}$. That means in particular that you should not try to write out $Y_1^{-1}\psi_{\rm H}$ using the Clebsch-Gordan coefficients of chapter 12.7, {N.13}. If you do not know what Clebsch-Gordan coefficients are, you have nothing to worry about.)

To get a nonzero inner product, one of the possible states of net angular momentum above will need to match the quantum numbers $j_{\rm L}$ and $m_{\rm L}$ of state $\psi_{\rm L}$. So

$$j_{\rm L} = j_{\rm H} - 1,\ j_{\rm H},\ \text{or } j_{\rm H}+1 \qquad m_{\rm L} = m_{\rm H} - 1$$

(And if $j_{\rm H} = 0$, $j_{\rm L}$ cannot be zero.) But recall that x_i also contained a Y_1^1 state. That state will allow $m_{\rm L} = m_{\rm H} + 1$. And if you take a wave that has its electric field in the z-direction instead of the x-direction, you also get a Y_1^0 state that gives the possibility $m_{\rm L} = m_{\rm H}$.

So the complete selection rules for electric dipole transitions are

$$j_{\rm L} = j_{\rm H} - 1,\ j_{\rm H},\ \text{or } j_{\rm H}+1 \qquad m_{\rm L} = m_{\rm H} - 1,\ m_{\rm H},\ \text{or } m_{\rm H}+1 \qquad \pi_{\rm L}\pi_{\rm H} = -1$$

where π means the parity. In addition, at least one of $j_{\rm L}$ or $j_{\rm H}$ must be nonzero. And as always for these quantum numbers, $j_{\rm L} \geqslant 0$ and $|m_{\rm L}| \leqslant j_{\rm L}$. Equivalent selection rules were written down for the hydrogen atom with spin-orbit interaction in chapter 7.4.4.

For magnetic dipole transitions, the relevant inner product is

$$\langle \psi_{\rm L} | \widehat{L}_{i,y} + g_i \widehat{S}_{i.y} | \psi_{\rm H} \rangle$$

Note that it is either $\widehat{L}_y$ or $\widehat{S}_y$ that is applied on $\psi_{\rm H}$, not both at the same time. It will be assumed that $\psi_{\rm H}$ is written in terms of states with definite angular momentum in the z-direction. In those terms, the effect of $\widehat{L}_y$ or $\widehat{S}_y$ is known to raise or lower the corresponding magnetic quantum number m by one unit, chapter 12.11. Which means that the net angular momentum can change by one unit. (Like when opposite orbital angular momentum and spin change into parallel ones. Note also that for the hydrogen atom in the nonrelativistic approximation of chapter 4.3, there is no interaction between the electron spin and the orbital motion. In that case, the magnetic dipole term can only change the value of m_l or m_s by one unit. Simply put, only the direction of the angular momentum changes. That is normally a trivial change as empty space has no preferred direction.)

One big limitation is that in either an electric or a magnetic dipole transition, the net atomic or nuclear angular momentum j can change by no more than one unit. Larger changes in angular momentum require higher multipole orders ℓ. These add a factor $z_i^{\ell-1}$ to the inner products. Now it turns out that:

$$z_i^{\ell-1} \sim \frac{(\ell-1)!\sqrt{4\pi(2\ell-1)}}{(2\ell-1)!!} r_i^{\ell-1} Y^0_{\ell-1} + \ldots \qquad (2\ell-1)!! \equiv \frac{(2\ell-1)!}{2^{\ell-1}(\ell-1)!} \quad \text{(A.174)}$$

Here the dots stand for spherical harmonics with lower square angular momentum. (To verify the above relation, use the Rayleigh formula of {A.6}, and expand the Bessel function and the exponential in it in Taylor series.) So the factor $z_i^{\ell-1}$ has a maximum azimuthal quantum number l equal to $\ell-1$. That means that the maximum achievable change in atomic or nuclear angular momentum increases by one unit for each unit increase in multipole order ℓ.

It follows that the first multipole term that can be nonzero has $\ell = |j_{\rm H} - j_{\rm L}|$, or $\ell = 1$ if the angular momenta are equal. At that multipole level, either the electric or the magnetic term can be nonzero, depending on parity. Normally this term will then dominate the transition process, as the terms of still higher multipole levels are ballparked to be much smaller.

A further limitation applies to orbital angular momentum. The angular momentum operators will not change the orbital angular momentum values. And the factors $z_i^{\ell-1}$ and x_i can only change it by up to $\ell-1$, respectively 1 units. So the minimum difference in possible orbital angular momentum values will have to be no larger than that:

$$\boxed{\text{electric:} \quad |l_{\rm H} - l_{\rm L}|_{\rm min} \leqslant \ell \qquad \text{magnetic:} \quad |l_{\rm H} - l_{\rm L}|_{\rm min} \leqslant \ell - 1} \qquad \text{(A.175)}$$

This is mainly important for single-particle states of definite orbital angular momentum. That includes the hydrogen atom, even with the relativistic spin-

orbit interaction. (But it does assume the nonrelativistic Hamiltonian in the actual transition process.)

The final limitation is that $j_{\rm H}$ and $j_{\rm L}$ cannot both be zero. The reason is that if $j_{\rm H}$ is zero, the possible angular momentum values of $z_i^{\ell-1}x_i j_{\rm H}$ are those of $z_i^{\ell-1}x_i$. And those values do not include zero to match $j_{\rm L}=0$. (According to the rules of quantum mechanics, the probability of zero angular momentum is given by the inner product with the spherical harmonic Y_0^0 of zero angular momentum. Since Y_0^0 is just a constant, the inner product is proportional to the average of $z_i^{\ell-1}x_i$ on a spherical surface around the origin. That average will be zero because by symmetry positive values of x_i will average away against corresponding negative ones.)

A.25.6 Ballpark decay rates

It may be interesting to find some actual ballpark values for the spontaneous decay rates. More sophisticated values, called the Weisskopf and Moszkowski estimates, will be derived in a later subsection. However, they are ballparks one way or the other.

It will be assumed that only a single particle, electron or proton, changes states. It will also be assumed that the first multipole contribution allowed by angular momentum and parity is indeed nonzero and dominant. In fact, it will be assumed that this contribution is as big as it can reasonably be.

To get the spontaneous emission rate, first the proper amplitude $\mathcal{E}_0$ of the electric field to use needs to be identified. The same relativistic procedure as in {A.24} may be followed to show it should be taken as

$$\text{spontaneous emission:}\quad \mathcal{E}_0 = \sqrt{\frac{\hbar\omega}{\epsilon_0 \mathcal{V}}}$$

That assumes that the entire system is contained in a very large periodic box of volume $\mathcal{V}$. Also, $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is the permittivity of space

Next, Fermi's golden rule of chapter 7.6.1 says that the transition rate is

$$\lambda_{\rm H\to L} = \overline{|H_{21}|^2}\frac{2\pi}{\hbar}\frac{{\rm d}N}{{\rm d}E}$$

Here H_{21} is approximated as the first allowed (nonzero) multipole contribution $H^{{\rm E}\ell}_{21,i}$ or $H^{{\rm M}\ell}_{21,i}$. So the additional higher order nonzero contributions are ignored, The overline means that this contribution needs to be suitably averaged over all directions of the electromagnetic wave. Further ${\rm d}N/{\rm d}E$ is the number of photon states in the periodic box per unit energy range. This is the density of states as given in chapter 6.3 (6.7). Using the Planck-Einstein relation it is:

$$\frac{{\rm d}N}{{\rm d}E} = \frac{\omega^2}{\hbar\pi^2 c^3}\mathcal{V}$$

Ballpark matrix coefficients were given in subsection A.25.4. However, a more accurate estimate is desirable. The main problem is the factor $r_{i,k}^{\ell-1}$ in the matrix elements (A.171) and (A.172). This factor equals $z_i^{\ell-1}$ if the z-axis is taken to be in the direction of wave motion. According to the previous subsection

$$z_i^{\ell-1} \sim \frac{(\ell-1)!\sqrt{4\pi(2\ell-1)}}{(2\ell-1)!!} r_i^{\ell-1} Y_{\ell-1}^0 + \ldots$$

The dots indicate spherical harmonics of lower angular momentum that do not do anything. Only the shown term is relevant for the contribution of lowest multipole order. So only the shown term should be ballparked. That can be done by estimating r_i as R, and $Y_{\ell-1}^0$ as $1/\sqrt{4\pi}$, (which is exact for $\ell = 1$).

The electric inner product contains a further factor x_i, taking the x-axis in the direction of the electric field. That will be accounted for by upping the value of ℓ one unit in the expression above. The magnetic inner product contains angular momentum operators. Since not much can be said about these easily, they will simply be estimated as $\hbar$.

Putting it all together, the estimated decay rates become

$$\boxed{\lambda^{\mathrm{E}\ell} \sim \alpha\omega(kR)^{2\ell}\frac{4(2\ell+1)}{(2\ell+1)!!^2} f^{\mathrm{E}\ell} \qquad \lambda^{\mathrm{M}\ell} \sim \alpha\omega(kR)^{2\ell}\frac{4(2\ell-1)}{(2\ell-1)!!^2}\left(\frac{\hbar}{2m_i cR}\right)^2 f^{\mathrm{M}\ell}} \tag{A.176}$$

Here

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137}$$

is the so-called fine structure constant. with $e = 1.6\ 10^{-19}$ C the proton or electron charge, $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m the permittivity of space, and $c = 3\ 10^8$ m/s the speed of light. This nondimensional constant gives the strength of the coupling between charged particles and photons, so it should obviously be there. The factor ω is expected for dimensional reasons; it gives the decay rate units of inverse time. The nondimensional factor kR reflects the fact that the atom or nucleus has difficulty interacting with the photon because its size is so small compared to the photon wave length. That is worse for higher multipole orders ℓ, as their photons produce less of a field near the origin. The factors $f^{\mathrm{E}\ell}$ and $f^{\mathrm{M}\ell}$ represent unknown corrections for the errors in the ballparks. These factors are hoped to be 1. (Fat chance.) As far as the remaining numerical factors are concerned, ...

The final parenthetical factor in the magnetic decay rate was already discussed in subsection A.25.4. It normally makes magnetic decays slower than electric ones of the same multipole order, but faster than electric ones of the next order.

These estimates are roughly similar to the Weisskopf ones. While they tend to be larger, that is largely compensated for by the fact that in the above

estimates R is the mean radius. In the Weisskopf estimates it is the edge of the nucleus.

In any case, actual decay rates can vary wildly from either pair of estimates. For example, nuclei satisfy an approximate conservation law for a quantity called isospin. If the transition violates an approximate conservation law like that, the transition rate will be unusually small. Also, it may happen that the initial and final wave functions have little overlap. That means that the regions where they both have significant magnitude are small. (These regions should really be visualized in the high-dimensional space of all the particle coordinates.) In that case, the transition rate can again be unexpectedly small.

Conversely, if a lot of particles change state in a transition, their individual contributions to the matrix element can add up to an unexpectedly large transition rate.

A.25.7 Wave functions of definite angular momentum

The analysis so far has represented the electromagnetic field in terms of photon states of definite linear momentum. But it is usually much more convenient to use states of definite angular momentum. That allows full use of the conservation laws of angular momentum and parity.

The states of definite angular momentum have vector potentials given by the photon wave functions of addendum {A.21.7}. For electric Eℓ and magnetic Mℓ multipole transitions respectively:

$$\vec{A}^{\rm E}_\gamma = \frac{A_0}{k}\nabla \times \vec{r} \times \nabla j_\ell(kr) Y_\ell^m(\theta,\phi) \qquad \vec{A}^{\rm M}_\gamma = A_0 \vec{r} \times \nabla j_\ell(kr) Y_\ell^m(\theta,\phi)$$

Here j_ℓ is a spherical Bessel function, {A.6} and Y_ℓ^m a spherical harmonic, chapter 4.2.3. The azimuthal angular momentum quantum number of the photon is ℓ. Its quantum number of angular momentum in the chosen z-direction is m. The electric state has parity $(-1)^\ell$ and the magnetic one $(-1)^{\ell-1}$. (That includes the intrinsic parity, unlike in some other sources). Further A_0 is a constant.

The contribution of a particle i to the matrix element is as before

$$H_{21,i} = -\frac{q_i}{m_i}\langle \psi_{\rm L}|\vec{A}_i \cdot \widehat{\vec{p}}_i|\psi_{\rm H}\rangle - \frac{q_i}{2m_i} g_i \langle \psi_{\rm L}|\vec{\mathcal{B}}_i \cdot \widehat{\vec{S}}_i|\psi_{\rm H}\rangle \qquad \vec{\mathcal{B}}_i = \nabla_i \times \vec{A}_i$$

But now, for electric transitions $\vec{A}_i$ needs to be taken as the complex conjugate of the photon wave function $\vec{A}^{\rm E}_\gamma$ above, evaluated at the position of particle i. For magnetic transitions $\vec{A}_i$ needs to be taken as the complex conjugate of $\vec{A}^{\rm M}_\gamma$. The complex conjugates are a result of the quantization of radiation, {A.24}. And they would not be there for absorption. (The classical reasons are much like the story for plane electromagnetic waves given earlier. But here

the nonquantized waves are too messy to even bother about, in this author's opinion.)

The matrix elements can be approximated assuming that the wave length of the photon is large compared to the size R of the atom or nucleus. The approximate contribution of the particle to the Eℓ electric matrix element is then, {D.43.2},

$$H^{\mathrm{E}\ell}_{21,i} \approx -\mathrm{i}q_i c A_0 \frac{(\ell+1)k^\ell}{(2\ell+1)!!} \langle \psi_{\mathrm{L}} | r_i^\ell Y_{\ell i}^{m*} | \psi_{\mathrm{H}} \rangle$$

The subscript i on the spherical harmonic means that its arguments are the coordinates of particle i. For nuclei, the above result is again suspect for the reasons discussed in {N.14}.

The approximate contribution of the particle to the Mℓ magnetic matrix element is {D.43.2},

$$H^{\mathrm{M}\ell}_{21,i} \approx \frac{q_i}{2m_i} A_0 \frac{(\ell+1)k^\ell}{(2\ell+1)!!} \langle \psi_{\mathrm{L}} | (\nabla_i r_i^\ell Y_{\ell i}^{m*}) \cdot \Big(\frac{2}{\ell+1} \widehat{\vec{L}}_i + g_i \widehat{\vec{S}}_i \Big) | \psi_{\mathrm{H}} \rangle$$

In general these matrix elements will need to be summed over all particles.

The above matrix elements can be analyzed similar to the earlier linear momentum ones. However, the above matrix elements allow you to keep the atom or nucleus in a fixed orientation. For the linear momentum ones, the nuclear orientation must be changed if the direction of the wave is to be held fixed. And in any cases, linear momentum matrix elements must be averaged over all directions of wave propagation. That makes the above matrix elements much more convenient in most cases.

Finally the matrix elements can be converted into spontaneous decay rates using Fermi's golden rule of chapter 7.6.1. In doing so, the needed value of the constant A_0 and corresponding density of states are, following {A.21.7} and {A.24},

$$A_0 = -\frac{1}{\mathrm{i}c} \sqrt{\frac{\hbar\omega}{\ell(\ell+1)\epsilon_0 r_{\max}}} \qquad \frac{\mathrm{d}N}{\mathrm{d}E} \approx \frac{1}{\hbar\pi c} r_{\max}$$

This assumes that the entire system is contained inside a very big sphere of radius $r_{\max}$. This radius $r_{\max}$ disappear in the final answer, and the final decay rates will be the ones in infinite space. (Despite the absence of $r_{\max}$ they do not apply to a finite sphere, because the density of states above is an approximation for large $r_{\max}$.)

It is again convenient to nondimensionalize the matrix elements using some suitably defined typical atomic or nuclear radius R. Recent authorative sources, like [33] and ⟦4⟧, take the nuclear radius equal to

$$R = 1.2A^{1/3} \text{ fm} \tag{A.177}$$

Here A is the number of protons and neutrons in the nucleus and a fm is 10^{-15} m.

The final decay rates are much like the ones (A.176) found earlier for linear momentum modes. In fact, linear momentum modes should give the same answer as the angular ones, if correctly averaged over all directions of the linear momentum. The decay rates in terms of angular momentum modes are:

$$\boxed{\lambda^{\mathrm{E}\ell} = \alpha\omega(kR)^{2\ell}\frac{2(l+1)}{l(2l+1)!!^2}|h_{21}^{\mathrm{E}\ell}|^2} \tag{A.178}$$

$$\boxed{\lambda^{\mathrm{M}\ell} = \alpha\omega(kR)^{2\ell}\frac{2(l+1)}{l(2l+1)!!^2}\left(\frac{\hbar}{2mcR}\right)^2|h_{21}^{\mathrm{M}\ell}|^2} \tag{A.179}$$

where $\alpha \approx 1/137$ is again the fine structure constant. The nondimensional matrix elements in these expressions are

$$\boxed{|h_{21}^{\mathrm{E}\ell}| = \sum_i \sqrt{4\pi}\,\frac{q_i}{e}\langle\psi_{\mathrm{L}}|r_i^\ell Y_{\ell i}^{m*}/R^\ell|\psi_{\mathrm{H}}\rangle} \tag{A.180}$$

$$\boxed{|h_{21}^{\mathrm{M}\ell}| = \sum_i \sqrt{4\pi}\,\frac{q_i}{e}\frac{m}{m_i}\langle\psi_{\mathrm{L}}|(\nabla_i r_i^\ell Y_{\ell i}^{m*}/R^{\ell-1})\cdot\Big(\frac{2\widehat{\vec{L}}_i}{(\ell+1)\hbar}+\frac{g_i\widehat{\vec{S}}_i}{\hbar}\Big)|\psi_{\mathrm{H}}\rangle} \tag{A.181}$$

The sum is over the electrons or protons and neutrons, with q_i their charge and m_i their mass. The reference mass m would normally be taken to be the mass of an electron for atoms and of a proton for nuclei. That means that for the electron or proton the charge and mass ratios can be set equal to 1. For an electron g_i is about 2, while for a proton, g_i would be about 5.6 if the effect of the neighboring protons and neutrons is ignored. For the neutron, the (net) charge q_i is zero. Therefore the electric matrix element is zero, and so is the first term in the magnetic one. In the second term, however, the charge and mass of the proton need to be used, along with $g_i = -3.8$, assuming again that the effect of the neighboring protons and neutrons is ignored.

A.25.8 Weisskopf and Moszkowski estimates

The Weisskopf and Moszkowski estimates are ballpark spontaneous decay rates. They are found by ballparking the nondimensional matrix elements (A.180) and (A.181) given in the previous subsection. The estimates are primarily intended for nuclei. However, they can easily be adopted to the hydrogen atom with a few straightforward changes.

It is assumed that a single proton numbered i makes the transition. The rest of the nucleus stays unchanged and can therefore be ignored in the analysis. Note that this does not take into account that the proton and the rest of the

nucleus should move around their common center of gravity. Correction factors for that can be applied, see [33] and [11] for more. In a similar way, the case that a single neutron makes the transition can be accounted for.

It is further assumed that the initial and final wave functions of the proton are of a relatively simple form, In spherical coordinates:

$$\psi_{\rm H} = R_{\rm H}(r_i)\Theta^{m_{j\rm H}}_{l_{\rm H}j_{\rm H}}(\theta_i,\phi_i) \qquad \Longrightarrow \qquad \psi_{\rm L} = R_{\rm L}(r_i)\Theta^{m_{j\rm L}}_{l_{\rm L}j_{\rm L}}(\theta_i,\phi_i)$$

These wave functions are very much like the $R_{nl}(r_i)Y_l^{m_l}(\theta_i,\phi_i)\updownarrow$ wave functions for the electron in the hydrogen atom, chapter 4.3. However, for nuclei, it turns out that you want to combine the orbital and spin states into states with definite *net* angular momentum j and definite *net* angular momentum m_j in the chosen z-direction. Such combinations take the form

$$\Theta^{m_j}_{lj}(\theta_i,\phi_i) = c_1 Y_l^{m_j-\frac{1}{2}}(\theta_i,\phi_i)\uparrow + c_2 Y_l^{m_j+\frac{1}{2}}(\theta_i,\phi_i)\downarrow$$

The coefficients c_1 and c_2 are of no interest here, but you can find them in chapter 12.8 2 if needed.

In fact even for the hydrogen atom you really want to take the initial and final states of the electron of the above form. That is due to a small relativistic effect called "spin-orbit interaction," {A.39}. It just so happens that for nuclei, the spin-orbit effect is much larger. Note however that the electric matrix element ignores the spin-orbit effect. That is a significant problem, {N.14}. It will make the ballparked electric decay rate for nuclei suspect. But there is no obvious way to fix it.

The nondimensional electric matrix element (A.180) can be written as an integral over the spherical coordinates of the proton. It then falls apart into a radial integral and an angular one:

$$|h^{{\rm E}\ell}_{21}| \approx \int R_{\rm L}(r_i)^*(r_i/R)^\ell R_{\rm H}(r_i) r_i^2 \,{\rm d}r_i \quad \sqrt{4\pi}\int \Theta^{m_{j\rm L}*}_{l_{\rm L}j_{\rm L}i} Y^{m*}_{\ell i} \Theta^{m_{j\rm H}}_{l_{\rm H}j_{\rm H}i} \sin^2\theta_i {\rm d}\theta_i {\rm d}\phi_i$$

Note that in the angular integral the product of the angular wave functions implicitly involves inner products between the spin states. Spin states are orthonormal, so their product is 0 if the spins are different and 1 if they are the same.

The bottom line is that the square electric matrix element can be written as a product of a radial factor,

$$f^{{\rm rad},\ell}_{\rm LH} \equiv \left[\int R^*_{\rm L}(r_i)(r_i/R)^\ell R_{\rm H}(r_i) r_i^2 \,{\rm d}r_i\right]^2 \tag{A.182}$$

and an angular one,

$$f^{{\rm ang},\ell}_{\rm LH} \equiv \left[\sqrt{4\pi}\int \Theta^{m_{j\rm L}*}_{l_{\rm L}j_{\rm L}i} Y^{m*}_{\ell i} \Theta^{m_{j\rm H}}_{l_{\rm H}j_{\rm H}i} \sin^2\theta_i {\rm d}\theta_i {\rm d}\phi_i\right]^2 \tag{A.183}$$

As a result, the electric multipole decay rate (A.180) becomes

$$\boxed{\lambda^{\mathrm{E}\ell} = \alpha\omega(kR)^{2\ell}\frac{2(l+1)}{l(2l+1)!!^2}f_{\mathrm{LH}}^{\mathrm{rad},\ell}f_{\mathrm{LH}}^{\mathrm{ang},\ell}} \tag{A.184}$$

Here the trailing factors represent the square matrix element.

A similar expression can be written for the nondimensional magnetic matrix element, {D.43.3}: It gives the decay rate (A.181) as

$$\boxed{\lambda^{\mathrm{M}\ell} = \alpha\omega(kR)^{2\ell}\frac{2(l+1)}{l(2l+1)!!^2}\left(\frac{\hbar}{2mcR}\right)^2 f_{\mathrm{LH}}^{\mathrm{rad},\ell-1}f_{\mathrm{LH}}^{\mathrm{ang},\ell}f_{\mathrm{LH}}^{\mathrm{mom},\ell}} \tag{A.185}$$

In this case, there is an third factor related to the spin and orbital angular momentum operators that appear in the magnetic matrix element. Also, the integrand in the radial factor is one order of r lower than in the electric element. That is due to the nabla operator ∇ in the magnetic element. It means that in terms of the radial electric factor as defined above, the value of ℓ to use is one unit below the actual multipole order.

Consider now the values of these factors. The radial factor (A.182) is the simplest one. The Weisskopf and Moszkowski estimates use a very crude approximation for this factor. They assume that the radial wave functions are equal to some constant up to the nuclear radius R and zero beyond it. (This assumption is not completely illogical for nuclei, as nuclear densities are fairly constant until the nuclear edge.) That gives, {D.43.3},

$$f_{\mathrm{LH}}^{\mathrm{rad},\ell} = \left(\frac{3}{\ell+3}\right)^2$$

Note that the magnetic decay rate uses $\ell+2$ in the denominator instead of $\ell+3$ because of the lower power of r_i.

More reasonable assumptions for the radial wave functions are possible. For a hydrogen atom instead of a nucleus, the obvious thing to do is to use the actual radial wave functions R_{nl} from chapter 4.3. That gives the radial factors listed in table A.1. These take R equal to the Bohr radius. That explains why some values are so large: the average radial position of the electron can be much larger than the Bohr radius in various excited states. In the table, n is the principal quantum number that gives the energy of the state. Further l is the azimuthal quantum number of orbital angular momentum. The two pairs of nl values correspond to those of the initial and final states; in what order does not make a difference. There are two radial factors listed for each pair of states. The first value applies to electric and multipole transitions at the lowest possible multipole order. That is usually the important one, because normally transition rates decrease rapidly with multipole order.

nl	nl	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert}$	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert+2}$	nl	nl	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert}$	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert+2}$
10	10	F	F	41	10	$2^{22}3^{3}5^{-13}$ (0.0928)	F
20	10	F	F	41	20	$2^{19}3^{-13}5$ (1.6442)	F
20	20	F	F	41	21	0	$2^{29}3^{-16}5$ (62.359)
21	10	$2^{15}3^{-9}$ (1.6648)	F	41	30	$2^{22}3^{8}5^{3}7^{-16}17^{2}$ (29.914)	F
21	20	3^{3} (27)	F	41	31	0	$2^{37}3^{10}5\,7^{-18}23^{2}$ (13182.)
21	21	1	$2^{2}3^{2}5^{2}$ (900)	41	32	$2^{33}3^{8}7^{-16}$ (1.6959)	$2^{47}3^{12}5^{2}7^{-20}11^{2}$ (2.84E6)
30	10	F	F	41	40	$2^{2}3^{3}5$ (540)	F
30	20	F	F	41	41	1	$2^{6}3^{2}5^{4}$ (360000)
30	21	$2^{15}3^{8}5^{-12}$ (0.8806)	F	42	10	$2^{32}3^{2}5^{-15}$ (1.2666)	F
30	30	F	F	42	20	$2^{29}3^{-16}5$ (62.359)	F
31	10	$2^{-13}3^{7}$ (0.2670)	F	42	21	$2^{23}3^{-15}5$ (2.9231)	0
31	20	$2^{20}3^{7}5^{-12}$ (9.3931)	F	42	30	$2^{32}3^{13}5\,7^{-18}43^{2}$ (38876.)	F
31	21	0	$2^{26}3^{10}5^{-14}$ (649.25)	42	31	$2^{31}3^{11}5\,7^{-16}$ (57.235)	$2^{49}3^{15}5^{3}7^{-20}$ (1.27E7)
31	30	$2\,3^{4}$ (162)	F	42	32	0	$2^{43}3^{13}7^{-18}$ (8611.9)
31	31	1	$2^{4}3^{4}5^{2}$ (32400)	42	40	$2^{8}3^{2}5^{3}$ (288000)	F
32	10	$2^{-15}3^{9}5$ (3.0034)	F	42	41	$2^{4}3^{3}$ (432)	$2^{8}3^{3}5^{4}7^{2}$ (2.12E8)
32	20	$2^{32}3^{9}5^{-15}$ (2770.1)	F	42	42	1	$2^{6}3^{4}7^{2}$ (254016)
32	21	$2^{22}3^{8}5^{-13}$ (22.543)	$2^{32}3^{12}5^{-17}7^{2}$ (1.47E5)	43	10	$2^{36}3^{2}5^{-17}7$ (5.6745)	F
32	30	$2\,3^{4}5^{3}$ (20250)	F	43	20	$2^{33}3^{-16}5^{3}7$ (1.75E5)	F
32	31	$2^{-2}3^{4}5$ (101.25)	$3^{6}5^{3}7^{2}$ (4.47E6)	43	21	$2^{31}3^{-17}5\,7$ (582.02)	$2^{41}3^{-17}5^{3}7$ (1.49E7)
32	32	1	$2^{2}3^{4}7^{2}$ (15876)	43	30	$2^{36}3^{17}5^{3}7^{-21}101^{2}$ (2.03E7)	F
40	10	F	F	43	31	$2^{39}3^{13}5\,7^{-19}11^{2}$ (46520.)	$2^{49}3^{23}5^{5}7^{-23}$ (6.05E9)
40	20	F	F	43	32	$2^{37}3^{11}7^{-17}$ (104.66)	$2^{47}3^{19}5^{2}7^{-21}$ (7.32E6)
40	21	$2^{21}3^{-15}$ (0.1462)	F	43	40	$2^{8}3^{2}5^{3}7^{3}$ (9.88E7)	F
40	30	F	F	43	41	$2^{8}3\,5^{2}7$ (134400)	$2^{12}3^{7}5^{2}7^{3}$ (7.7E10)
40	31	$2^{29}3^{7}7^{-16}13^{2}$ (5.9709)	F	43	42	$2^{2}3^{2}7$ (252)	$2^{8}3^{4}5^{4}7$ (9.07E7)
40	32	$2^{45}3^{9}5\,7^{-18}$ (2126.4)	F	43	43	1	$2^{6}3^{4}5^{2}$ (129600)
40	40	F	F				

Table A.1: Radial integral correction factors for hydrogen atom wave functions.

To understand the given values more clearly, first consider the relation between multipole order and orbital angular momentum. The derived matrix elements implicitly assume that the potential of the proton or electron only depends on its position, not its spin. So spin does not really affect the orbital motion. That means that the multipole order for nontrivial transitions is constrained by orbital angular momentum conservation, [33]:

$$|l_{\rm H} - l_{\rm L}| \leqslant \ell \leqslant l_{\rm H} + l_{\rm L} \tag{A.186}$$

Note that this is a consequence of (A.175) within the single-particle model. It is just like for the nonrelativistic hydrogen atom, (7.17). (M1 transitions that merely change the direction of the spin, like a $Y_0^0\uparrow$ to $Y_0^0\downarrow$ one, are irrelevant since they do not change the energy. Fermi's golden rule makes the transition rate for transitions with no energy change theoretically zero, chapter 7.6.1.)

The minimum multipole order implied by the left-hand constraint above corresponds to an electric transition because of parity. However, this transition may be impossible because of *net* angular momentum conservation or because ℓ must be at least 1. That will make the transition of lowest multipole order a magnetic one. The magnetic transition still uses the same value for the radial factor though. The second radial factor in the table is provided since the next-higher electric multipole order might reasonably compete with the magnetic one.

More realistic radial factors for nuclei can be formulated along similar lines. The simplest physically reasonable assumption is that the protons and neutrons are contained within an impenetrable sphere of radius R. A hydrogen-like numbering system of the quantum states can again be used, figure 14.14, with one difference. For hydrogen, a given energy level n allows all orbital momentum quantum numbers l up to $n-1$. For nuclei, l must be even if n is odd and vice-versa, chapter 14.12.1. Also, while for the (nonrelativistic) hydrogen atom the energy does not depend on l, for nuclei that is only a rough approximation. (It assumes that the nuclear potential is like an harmonic oscillator one, and that is really crude.)

Radial factors for the impenetrable-sphere model using this numbering system are given in table A.2.

These results illustrate the limitations of M1 transitions in the single-particle model. Because of the condition (A.186) above and parity, the orbital quantum number l cannot change in M1 transitions. A glance at the table then shows that the radial factor is zero unless the initial and final radial states are identical. (That is a consequence of the orthonormality of the energy states.) So M1 transitions cannot change the radial state. All they can do is change the *direction* of the orbital angular momentum or spin of a given state. Obviously that is ho-hum, though with a spin-orbit term it may still do something. Without a spin-orbit term, there would be no energy change, and Fermi's golden rule would make the theoretical transition rate then zero. That is similar to the

nl	nl	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert}$	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert+2}$	nl	nl	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert}$	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert+2}$	nl	nl	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert}$	$f_{\rm LH}^{{\rm rad},\lvert\Delta l\rvert+2}$
10	10	F	F	61	50	0.2066	F	72	65	0.0173	0.0150
21	10	0.2810	F	61	52	0.0918	0.0462	72	70	0.1073	F
21	21	1	0.1403	61	54	0.0211	0.0161	72	72	1	0.1270
30	10	F	F	61	61	1	0.1158	74	10	0.0018	F
30	21	0.0922	F	63	10	0.0021	F	74	21	0.0030	0.0055
30	30	F	F	63	21	0.0029	0.0087	74	30	0.0188	F
32	10	0.1116	F	63	30	0.0348	F	74	32	0.0032	0.0103
32	21	0.3727	0.0760	63	32	0.0018	0.0164	74	41	0.0495	0.0219
32	30	0.0949	F	63	41	0.1058	0.0358	74	43	0.0018	0.0168
32	32	1	0.1925	63	43	0	0.0267	74	50	0.0434	F
41	10	0.0015	F	63	50	0.0684	F	74	52	0.1322	0.0436
41	21	0	0.0317	63	52	0.3243	0.0754	74	54	0	0.0244
41	30	0.2264	F	63	54	0.0390	0.0398	74	61	0.0710	0.0302
41	32	0.0647	0.0418	63	61	0.0969	0.0395	74	63	0.3611	0.0878
41	41	1	0.1206	63	63	1	0.1620	74	65	0.0320	0.0381
43	10	0.0537	F	65	10	0.0174	F	74	70	0.0343	F
43	21	0.1707	0.0445	65	21	0.0514	0.0182	74	72	0.0897	0.0409
43	30	0.0714	F	65	30	0.0357	F	74	74	1	0.1823
43	32	0.4367	0.1085	65	32	0.1226	0.0420	76	10	0.0110	F
43	41	0.0824	0.0383	65	41	0.0549	0.0248	76	21	0.0317	0.0125
43	43	1	0.2371	65	43	0.2618	0.0866	76	30	0.0255	F
50	10	F	F	65	50	0.0274	F	76	32	0.0737	0.0281
50	21	0.0013	F	65	52	0.0682	0.0346	76	41	0.0413	0.0194
50	30	F	F	65	54	0.5231	0.1673	76	43	0.1531	0.0567
50	32	0.0157	F	65	61	0.0254	0.0164	76	50	0.0238	F
50	41	0.1175	F	65	63	0.0634	0.0408	76	52	0.0567	0.0284
50	43	0.0267	F	65	65	1	0.3093	76	54	0.2978	0.1071
50	50	F	F	70	10	F	F	76	61	0.0254	0.0154
52	10	0.0022	F	70	21	1.4E-4	F	76	63	0.0648	0.0366
52	21	0.0018	0.0151	70	30	F	F	76	65	0.5542	0.1935
52	30	0.0775	F	70	32	0.0013	F	76	70	0.0134	F
52	32	0	0.0293	70	41	0.0016	F	76	72	0.0221	0.0160
52	41	0.2803	0.0633	70	43	0.0049	F	76	74	0.0565	0.0407
52	43	0.0490	0.0412	70	50	F	F	76	76	1	0.3389
52	50	0.1042	F	70	52	0.0221	F	87	10	0.0074	F
52	52	1	0.1411	70	54	0.0092	F	87	21	0.0207	0.0088
54	10	0.0292	F	70	61	0.1297	F	87	30	0.0186	F
54	21	0.0894	0.0278	70	63	0.0345	F	87	32	0.0469	0.0195
54	30	0.0506	F	70	65	0.0121	F	87	41	0.0310	0.0152
54	32	0.2200	0.0656	70	70	F	F	87	43	0.0953	0.0387
54	41	0.0708	0.0314	72	10	2.6E-4	F	87	50	0.0200	F
54	43	0.4849	0.1390	72	21	1.6E-4	0.0015	87	52	0.0450	0.0229
54	50	0.0294	F	72	30	0.0027	F	87	54	0.1810	0.0716
54	52	0.0719	0.0402	72	32	0	0.0016	87	61	0.0234	0.0138
54	54	1	0.2756	72	41	0.0020	0.0166	87	63	0.0570	0.0311
61	10	1.3E-4	F	72	43	0.0011	0.0068	87	65	0.3292	0.1268
61	21	0	0.0016	72	50	0.0656	F	87	70	0.0134	F
61	30	0.0019	F	72	52	0	0.0352	87	72	0.0234	0.0155
61	32	0.0012	0.0077	72	54	0.0087	0.0120	87	74	0.0613	0.0379
61	41	0	0.0369	72	61	0.2449	0.0597	87	76	0.5803	0.2177
61	43	0.0113	0.0133	72	63	0.0747	0.0458	87	87	1	0.3653

Table A.2: More realistic radial integral correction factors for nuclei.

limitation of M1 transitions for the nonrelativistic hydrogen atom in chapter 7.4.4.

It may be instructive to use the more realistic radial factors of table A.2 to get a rough idea of the errors in the Weisskopf ones. The initial comparison will be restricted to changes in the principal quantum number of no more than one unit. That means that transitions between widely separated shells will be ignored. Also, only the lowest possible multipole level will be considered. That corresponds to the first of each pair of values in the table. Assuming an electric transition, ℓ is the difference between the l values in the table. Consider now the following two simple approximations of the radial factor:

$$\boxed{\text{Weisskopf: } f_{\text{LH}}^{\text{rad},\ell} = \left(\frac{3}{\ell+3}\right)^2 \qquad \text{Empirical: } f_{\text{LH}}^{\text{rad},\ell} = \left(\frac{1.5}{\ell+3}\right)^2 \text{ or } 1} \tag{A.187}$$

The coefficient 1.5 comes from a least square approximation of the data. For M1 transitions, the exact value 1 should be used.

For the given data, it turns out that the Weisskopf estimate is on average too large by a factor 5. In the worst case, the Weisskopf estimate is too large by a factor 18. The empirical formula is on average off by a factor 2, and in the worst case by a factor 4.

If any arbitrary change in principal quantum number is allowed, the possible errors are much larger. In that case the Weisskopf estimates are off by average factor of 20, and a maximum factor of 4 000. The empirical estimates are off by an average factor of 8, and a maximum one of 1 000. Including the next number in table A.2 does not make much of a difference here.

These errors do depend on the change in principal quantum numbers. For changes in principal quantum number no larger than 2 units, the empirical estimate is off by a factor no greater that 10. For 3 or 4 unit changes, the estimate is off by a factor no greater than about 100. The absolute maximum error factor of 1 000 occurs for a 5 unit change in the principal quantum number. For the Weiskopf estimate, multiply these maximum factors by 4.

These data exclude the M1 transitions mentioned earlier, for which the radial factor is either 0 or 1 exactly. The value 0 implies an infinite error factor for a Weisskopf-type estimate of the radial factor. But that requires an M1 transition with at least a two unit change in the principal quantum number. In other words, it requires an M1 transition with a huge energy change.

Consider now the angular factor in the decay rates (A.184) and (A.185). It arises from integrating the spherical harmonics, (A.183). But the actual angular factor really used in the transition rates (A.184) and (A.185) also involves an averaging over the possible angular orientations of the initial atom. (This orientation is reflected in its magnetic quantum number $m_{j\text{H}}$.) And it involves a summation over the different angular orientations of the final nucleus that can be decayed to. The reason is that experimentally, there is usually no control

over the orientation of the initial and final nuclei. An average initial nucleus will have an average orientation. But each final orientation that can be decayed to is a separate decay process, and the decay rates add up. (The averaging over the initial orientations does not really make a difference; all orientations decay at the same rate, since space has no preferred direction. The summation over the final orientations is critical.)

j_L	j_H:	$\frac{1}{2}$		$\frac{3}{2}$		$\frac{5}{2}$		$\frac{7}{2}$		$\frac{9}{2}$		$\frac{11}{2}$		$\frac{13}{2}$	
$\frac{1}{2}$		F	1	1	1	1	1	1	1	1	1	1	1	1	1
$\frac{3}{2}$		2	2	F	$\frac{1}{5}$	$\frac{6}{5}$	$\frac{2}{7}$	$\frac{9}{7}$	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{4}{11}$	$\frac{15}{11}$	$\frac{5}{13}$	$\frac{18}{13}$	$\frac{2}{5}$
$\frac{5}{2}$		3	3	$\frac{9}{5}$	$\frac{3}{7}$	F	$\frac{3}{35}$	$\frac{9}{7}$	$\frac{1}{7}$	$\frac{10}{7}$	$\frac{2}{11}$	$\frac{50}{33}$	$\frac{30}{143}$	$\frac{225}{143}$	$\frac{3}{13}$
$\frac{7}{2}$		4	4	$\frac{18}{7}$	$\frac{2}{3}$	$\frac{12}{7}$	$\frac{4}{21}$	F	$\frac{1}{21}$	$\frac{4}{3}$	$\frac{20}{231}$	$\frac{50}{33}$	$\frac{50}{429}$	$\frac{700}{429}$	$\frac{20}{143}$
$\frac{9}{2}$		5	5	$\frac{10}{3}$	$\frac{10}{11}$	$\frac{50}{21}$	$\frac{10}{33}$	$\frac{5}{3}$	$\frac{25}{231}$	F	$\frac{1}{33}$	$\frac{15}{11}$	$\frac{25}{429}$	$\frac{225}{143}$	$\frac{35}{429}$
$\frac{11}{2}$		6	6	$\frac{45}{11}$	$\frac{15}{13}$	$\frac{100}{33}$	$\frac{60}{143}$	$\frac{25}{11}$	$\frac{25}{143}$	$\frac{18}{11}$	$\frac{10}{143}$	F	$\frac{3}{143}$	$\frac{18}{13}$	$\frac{6}{143}$
$\frac{13}{2}$		7	7	$\frac{63}{13}$	$\frac{7}{5}$	$\frac{525}{143}$	$\frac{7}{13}$	$\frac{1225}{429}$	$\frac{35}{143}$	$\frac{315}{143}$	$\frac{49}{429}$	$\frac{21}{13}$	$\frac{7}{143}$	F	$\frac{1}{65}$
$\frac{15}{2}$		8	8	$\frac{28}{5}$	$\frac{28}{17}$	$\frac{56}{13}$	$\frac{56}{85}$	$\frac{490}{143}$	$\frac{70}{221}$	$\frac{392}{143}$	$\frac{392}{2431}$	$\frac{28}{13}$	$\frac{196}{2431}$	$\frac{8}{5}$	$\frac{8}{221}$

Table A.3: Angular integral correction factors $f_{\text{LH}}^{\text{ang},|\Delta j|}$ and $f_{\text{LH}}^{\text{ang},|\Delta j|+1}$ for the Weisskopf electric unit and the Moszkowski magnetic one. The correction for the Weisskopf magnetic unit is to cross it out and write in the Moszkowski unit.

Values for the angular factor are in table A.3. For the first and second number of each pair respectively:

$$\ell = |j_H - j_L| \qquad \ell = |j_H - j_L| + 1$$

More generally, the angular factor is given by, [33, p. 878],

$$\boxed{f_{\text{LH}}^{\text{ang},\ell} = (2j_L + 1)[\langle \ell\, 0 || j_H\, \tfrac{1}{2} \rangle | j_L\, -\tfrac{1}{2} \rangle]^2} \tag{A.188}$$

Here the quantity in square brackets is called a Clebsch-Gordan coefficient. For small angular momenta, values can be found in figure 12.5. For larger values, refer to {N.13}. The leading factor is the reason that the values in the table are not the same if you swap the initial and final states. When the final state has the higher angular momentum, there are more nuclear orientations that an atom can decay to.

It may be noted that [11, p. 9-178] gives the above factor for electric transitions as

$$f_{\text{LH}}^{\text{ang},\ell} = (2j_L + 1)(2\ell + 1)(2l_L + 1)(2l_H + 1) \begin{pmatrix} l_L & l_H & l \\ 0 & 0 & 0 \end{pmatrix}^2 \begin{Bmatrix} l_L & j_L & \frac{1}{2} \\ j_H & l_H & \ell \end{Bmatrix}^2$$

Here the array in parentheses is the so-called Wigner 3j symbol and the one in curly brackets is the Wigner 6j symbol, {N.13}. The idea is that this expression will take care of the selection rules automatically. And so it does, if you assume that the multiply-defined l is ℓ, as the author seems to say. Of course, selection rules might be a lot easier to evaluate than 3j and 6j symbols.

For magnetic multipole transitions, with $\ell = |j_{\rm H} - j_{\rm L}| = |l_{\rm H} - l_{\rm L}| + 1$, the same source comes up with

$$f_{\rm LH}^{{\rm ang},\ell} = (2j_{\rm L}+1)\frac{3(2\ell+1)^2(2\ell-1)}{2\ell}(2l_{\rm L}+1)(2l_{\rm H}+1) \times \begin{pmatrix} l_{\rm L} & l_{\rm H} & \ell-1 \\ 0 & 0 & 0 \end{pmatrix}^2 \begin{Bmatrix} l_{\rm L} & j_{\rm L} & \frac{1}{2} \\ j_{\rm H} & l_{\rm H} & \ell-1 \end{Bmatrix}^2 \begin{Bmatrix} l_{\rm H} & l_{\rm L} & \ell-1 \\ \frac{1}{2} & \frac{1}{2} & 1 \\ j_{\rm H} & j_{\rm L} & \ell \end{Bmatrix}^2$$

Here the final array in curly brackets is the Wigner 9j symbol. The bad news is that the 6j symbol does not allow any transitions of lowest multipole order to occur! Someone familiar with 6j symbols can immediately see that from the so-called triangle inequalities that the coefficients of 6j symbols must satisfy, {N.13}. Fortunately, it turns out that if you simply leave out the 6j symbol, you do seem to get the right values and selection rules.

The magnetic multipole matrix element also involves an angular momentum factor. This factor turns out to be relatively simple, {D.43.3}:

$$f_{\rm LH}^{{\rm mom},|\Delta j|} = \left(g_i - \frac{2}{1+\ell}\right)^2 \ell^2 \qquad \begin{matrix} l_{\rm min} = j_{\rm min} + \frac{1}{2} \\ l_{\rm max} = j_{\rm max} - \frac{1}{2} \end{matrix}$$

$$f_{\rm LH}^{{\rm mom},|\Delta j|+1} = \begin{cases} \left(g_i - \dfrac{2-4j_{\rm min}}{1+\ell}\right)^2 (j_{\rm max}+1)^2 & \begin{matrix} l_{\rm min} = j_{\rm min} - \frac{1}{2} \neq 0 \\ l_{\rm max} = j_{\rm max} - \frac{1}{2} \end{matrix} \\ \left(g_i - \dfrac{2+4(\ell+j_{\rm min})}{1+\ell}\right)^2 j_{\rm min}^2 & \begin{matrix} l_{\rm min} = j_{\rm min} + \frac{1}{2} \\ l_{\rm max} = j_{\rm max} + \frac{1}{2} \end{matrix} \end{cases} \tag{A.189}$$

Here "min" and "max" refer to whatever is the smaller, respectively larger, one of the initial and final values.

The stated values of the orbital angular momentum l are the only ones allowed by parity and the orbital angular momentum conservation condition (A.186). In particular, consider the first expression above, for the minimum multipole order $\ell = |\Delta j|$. According to this expression, the change in orbital angular momentum cannot exceed the change in net angular momentum. That forbids a lot of magnetic transitions in a shell model setting, transitions that seem perfectly fine if you only look at net angular momentum and parity. Add to that the earlier observation that M1 transitions cannot change the radial state at all. Magnetic transitions are quite handicapped according to the single-particle model used here.

Of course, a single-particle model is not exact for multiple-particle systems. In a more general setting, transitions that in the ideal model would violate the orbital angular momentum condition can occur. For example, consider the possibility that the true state picks up some uncertainty in orbital angular momentum.

Presumably such transitions would be unexpectedly slow compared to transitions that do not violate any approximate orbital angular momentum conditions. That makes estimating the magnetic transition rates much more tricky. After all, for nuclei the net angular momentum is usually known with some confidence, but the orbital angular momentum of individual nucleons is not.

Fortunately, for electric transitions orbital angular momentum conservation does not provide additional limitations. Here the orbital requirements are already satisfied if net angular momentum and parity are conserved.

The derived decay estimates are now used to define standard decay rates. It is assumed that the multipole order is minimal, $\ell = |\Delta j|$, and that the final angular momentum is $\frac{1}{2}$. As table A.3 shows, that makes the angular factor equal to 1. The standard electric decay rate is then

$$\boxed{\lambda^{\mathrm{E}\ell}_{\text{Weisskopf}} = \alpha\omega(kR)^{2\ell}\frac{2(\ell+1)}{\ell(2\ell+1)!!^2}\frac{9}{(\ell+3)^2}} \tag{A.190}$$

This decay rate is called the "Weisskopf unit" for electric multipole transitions. It is commonly indicated by W.u. Measured actual decay rates are compared to this unit to get an idea whether they are unusually high or low.

Note that the decay rates are typically orders of magnitude off the mark. That is due to effects that cannot be accounted for. Nucleons are not independent particles by far. And even if they were, their radial wave functions would not be constant. The used expression for the electric matrix element is probably no good, {N.14}. And especially higher multipole orders depend very sensitively on the nuclear radius, which is imprecisely defined.

The standard magnetic multipole decay rate becomes under the same assumptions:

$$\boxed{\lambda^{\mathrm{M}\ell}_{\text{Moszkowski}} = \alpha\omega(kR)^{2\ell}\frac{2(\ell+1)}{\ell(2\ell+1)!!^2}\left(\frac{\hbar}{2mcR}\right)^2\frac{9}{(\ell+2)^2}\left(g_i - \frac{2}{\ell+1}\right)^2\ell^2} \tag{A.191}$$

This decay rate is called the "Moszkowski unit" for magnetic multipole transitions.

Finally, it should be mentioned that it is customary to ballpark the final momentum factor in the Moszkowski unit by 40. That is because Jesus spent 40 days in the desert. Also, the factor $(\ell+2)^2$ is customarily replaced by $(\ell+3)^2$, [10, p. 9-49], [36, p. 676], [5, p. 242], because, hey, anything for a laugh. Other sources keep the $(\ell+2)^2$ factor just like it is, [11, p. 9-178], [31, p. 332], because,

hey, why not? Note that the Handbook of Physics does both, depending on the author you look at. Taking the most recent of the cited sources, as well as ⟦4⟧, as reference the new and improved magnetic transition rate may be:

$$\boxed{\lambda^{\mathrm{M}\ell}_{\mathrm{Weisskopf}} = \alpha\omega(kR)^{2\ell}\frac{2(\ell+1)}{\ell(2\ell+1)!!^2}\left(\frac{\hbar}{mcR}\right)^2\frac{90}{(\ell+3)^2}} \tag{A.192}$$

This is called the Weisskopf magnetic unit. Note that the humor factor has been greatly increased. Whether there is a 2 or 3 in the final fraction does not make a difference. All analysis is relative to the perception of the observer. Where one perceives a 2 another sees a 3. Everthing is relative, as Einstein supposedly said, and otherwise quantum mechanics definitely did.

Note that the Weisskopf magnetic unit looks exactly like the electric one, except for the addition of a zero and the additional fraction between parentheses. That makes it easier to remember, especially for those who can remember the electric unit. For them the savings in time is tremendous, because they do not have to look up the correct expression. That can save a lot of time because many standard references have the formulae wrong or in some weird system of units. All that time is much better spend trying to guess whether your source, or your editor, uses a 2 or a 3.

A.25.9 Errors in other sources

There is a notable amount of errors in descriptions of the Weisskopf and Moszkowski estimates found elsewhere. That does not even include not mentioning that the electric multipole rate is likely no good, {N.14}. Or randomly using $\ell+2$ or $\ell+3$ in the Weisskopf magnetic unit.

These errors are more basic. The first edition of the Handbook of Physics, [10, p. 9-49], gives both Weisskopf units wrong. Squares are missing on the $\ell+3$, and so is the fine structure constant. The other numerical factors are consistent between the two units, but not right. Probably a strangely normalized matrix element is given, rather than the stated decay rates λ, and in addition the square was forgotten.

The same Handbook, [10, p. 9-110], but a different author, uses $g_i/2$ instead of g_i in the Moszkowski estimate. (Even physicists themselves can get confused if sometimes you define g_p to be 5.6 and sometimes 2.8, which also happens to be the magnetic moment μ_p in nuclear magnetons, which is often used as a "nondimensional" unit where g_p is really needed, etcetera.) More seriously, this error is carried over to the given plot of the Moszkowski unit, which is therefore wrong. Which is in addition to the fact that the nuclear radius used in it is too large by modern standards, using 1.4 rather than 1.2 in (A.177).

The error is corrected in the second edition, [11, p. 9-178], but the Moszkowski plot has disappeared. In favor of the Weisskopf magnetic unit, of course.

Think of the scientific way in which the Weisskopf unit has been deduced! This same reference also gives the erroneous angular factor for magnetic transitions mentioned in the previous subsection. Of course an additional 6j symbol that sneaks in is easily overlooked.

No serious errors were observed in [33]. (There is a readily-fixed error in the conversion formula for when the initial and final states are swapped.) This source does not list the Weisskopf magnetic unit. (Which is certainly defensible in view of its nonsensical assumptions.) Unfortunately non-SI units are used.

The electric dipole matrix element in [36, p. 676] is missing a factor $1/2c$. The claim that this element can be found by "straightforward calculation" is ludicrous. Not only is the mathematics convoluted, it also involves the major assumption that the potentials depend only on position. A square is missing in the Moszkowski unit, and the table of corresponding widths are in eV instead of the stated 1/s.

All three units are given incorrectly in [31, p. 332]. There is a factor 4π in them that should not be there. And the magnetic rate is missing a factor ℓ^2. The constant in the numerical expression for M3 transitions should be 15, not 16. Of course, the difference is negligible compared to replacing the parenthetical expression by 40, or compared to the orders of magnitude that the estimate is commonly off anyway.

The Weisskopf units are listed correctly in [5, p. 242]. Unfortunately non-SI units are used. The Moszkowski unit is not mentioned. The nonsensical nature of the Weisskopf magnetic unit is not pointed out. Instead it is claimed that it is found by a similar calculation as the electric unit.

A.26 Fourier inversion theorem and Parseval

This note discusses Fourier series, Fourier integrals, and Parseval's identity.

Consider first one-dimensional Fourier series. They apply to functions $f(x)$ that are periodic with some given period ℓ:

$$f(x+\ell) = f(x) \quad \text{for all } x$$

Such functions can be written as a "Fourier series:"

$$\boxed{f(x) = \sum_{\text{all } k} f_k \frac{e^{\mathrm{i}kx}}{\sqrt{\ell}} \qquad f_k = \int_0^\ell f(x) \frac{e^{-\mathrm{i}kx}}{\sqrt{\ell}}\,\mathrm{d}x} \tag{A.193}$$

Here the k values are those for which the exponentials are periodic of period ℓ. According to the Euler formula (2.5), that means that $k\ell$ must be a whole multiple n of 2π, so

$$k = n\frac{2\pi}{\ell} \qquad n = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots \tag{A.194}$$

Note that notations for Fourier series can vary from one author to the next. The above form of the Fourier series is the prefered one for quantum mechanics. The reason is that the functions $e^{\mathrm{i}kx}/\sqrt{\ell}$ form an orthonormal set:

$$\int_0^\ell \frac{e^{-\mathrm{i}kx}}{\sqrt{\ell}} \frac{e^{\mathrm{i}\underline{k}x}}{\sqrt{\ell}} \,\mathrm{d}x = \begin{cases} 1 \text{ if } k = \underline{k} \\ 0 \text{ if } k \neq \underline{k} \end{cases} \tag{A.195}$$

Quantum mechanics just loves orthonormal sets of functions. In particular, note that the above functions are momentum eigenfunctions. Just apply the linear momentum operator $\widehat{p} = \hbar\mathrm{d}/\mathrm{i}\mathrm{d}x$ on them. That shows that their linear momentum is given by the de Broglie relation $p = \hbar k$. Here these momentum eigenfunctions are properly normalized. They would not be using different conventions.

That any (reasonable) periodic function $f(x)$ can be written as a Fourier series was already shown in {D.8}. That derivation took ℓ be the half-period. The formula for the coefficients f_k can also be derived directly: simply multiply the expression (A.193) for $f(x)$ with $e^{\mathrm{i}\underline{k}x}/\sqrt{\ell}$ for any arbitrary value of $\underline{k}$ and integrate over x. Because of the orthonormality (A.195), the integration produces zero for all k except if $k = \underline{k}$, and then it produces $f_{\underline{k}}$ as required.

Note from (A.193) that if you known $f(x)$ you can find all the f_k. Conversely, if you know all the f_k, you can find $f(x)$ at every position x. The formulae work both ways.

But the symmetry goes even deeper than that. Consider the inner product of a pair of functions $f(x)$ and $g(x)$:

$$\int_0^\ell f^*(x)g(x)\,\mathrm{d}x = \int_0^\ell \sum_{\text{all } k} f_k^* \frac{e^{-\mathrm{i}kx}}{\sqrt{\ell}} \sum_{\text{all } \underline{k}} g_{\underline{k}} \frac{e^{\mathrm{i}\underline{k}x}}{\sqrt{\ell}} \,\mathrm{d}x$$

Using the orthonormality property (A.195) that becomes

$$\int_0^\ell f^*(x)g(x)\,\mathrm{d}x = \sum_{\text{all } k} f_k^* g_k \tag{A.196}$$

Now note that if you look at the coefficients f_k and g_k as the coefficients of infinite-dimensional vectors, then the right hand side is just the inner product of these vectors. In short, Fourier series preserve inner products.

Therefore the equation above may be written more concisely as

$$\boxed{\langle f(x)|g(x)\rangle = \langle f_k|g_k\rangle} \tag{A.197}$$

This is the so-called "Parseval identity." Now transformations that preserve inner products are called "unitary" in mathematics. So the Parseval identity shows that the transformation from periodic functions to their Fourier coefficients is unitary.

That is quite important for quantum mechanics. For example, assume that $f(x)$ is a wave function of a particle stuck on a ring of circumference ℓ. Wave functions should be normalized, so:

$$\int_0^\ell f^*(x)f(x)\,\mathrm{d}x = 1 = \sum_{\text{all } k} f_k^* f_k$$

According to the Born interpretation, the left hand side says that the probability of finding the particle is 1, certainty, if you look at every position on the ring. But according to the orthodox interpretation of quantum mechanics, $f_k^* f_k$ in the right hand side gives the probability of finding the particle with momentum $p = \hbar k$. The fact that the total sum is 1 means physically that it is certain that the particle will be found with *some* momentum.

So far, only periodic functions have been covered. But functions in infinite space can be handled by taking the period ℓ infinite. To do that, note from (A.194) that the k values of the Fourier series are spaced apart over a distance

$$\Delta k = \frac{2\pi}{\ell}$$

In the limit $\ell \to \infty$, Δk becomes an infinitesimal increment $\mathrm{d}k$, and the sums become integrals. Now in quantum mechanics it is convenient to replace the coefficients f_k by a new function $f(k)$ that is defined so that

$$f_k = \sqrt{\Delta k} f(k) \quad \Longrightarrow \quad |f_k|^2 = |f(k)|^2 \Delta k$$

The reason that this is convenient is that $|f_k|^2$ gives the probability for wave number k. Then for a function $f(k)$ that is defined as above, $|f(k)|^2$ gives the probability *per unit k-range*.

If the above definition and $\sqrt{\ell} = \sqrt{2\pi}/\Delta k$ are substituted into the Fourier series expressions (A.193), in the limit $\ell \to \infty$ it gives the "Fourier integral" formulae:

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(k) e^{\mathrm{i}kx}\,\mathrm{d}k \qquad f(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-\mathrm{i}kx}\,\mathrm{d}x} \tag{A.198}$$

In books on mathematics you will usually find function $f(k)$ indicated as $\widehat{f}(k)$, to clarify that it is a completely different function than $f(x)$. Unfortunately, the hat is already used for something much more important in quantum mechanics. So in quantum mechanics you will have to look at the argument, x or k, to know which function it really is.

Of course, in quantum mechanics you are often more interested in the momentum than in the wave number. So it is often convenient to define a new function $f(p)$ so that $|f(p)|^2$ gives the probability per unit momentum range

rather than unit wave number range. Because $p = \hbar k$, the needed rescaling of $f(k)$ is by a factor $\sqrt{\hbar}$. That gives

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} f(p) e^{\mathrm{i}px/\hbar} \,\mathrm{d}p \qquad f(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} f(x) e^{-\mathrm{i}px/\hbar} \,\mathrm{d}x} \tag{A.199}$$

Using similar substitutions as for the Fourier series, the Parseval identity (A.197) becomes

$$\int_{-\infty}^{\infty} f^*(x) g(x) \,\mathrm{d}x = \int_{-\infty}^{\infty} f^*(k) g(k) \,\mathrm{d}k = \int_{-\infty}^{\infty} f^*(p) g(p) \,\mathrm{d}p$$

or in short

$$\boxed{\langle f(x) | g(x) \rangle = \langle f(k) | g(k) \rangle = \langle f(p) | g(p) \rangle} \tag{A.200}$$

This identity is sometimes called the "Plancherel theorem," after a later mathematician who generalized its applicability. The bottom line is that Fourier integral transforms too conserve inner products.

So far, this was all one-dimensional. However, the extension to three dimensions is straightforward. The first case to be considered is that there is periodicity in each Cartesian direction:

$$f(x+\ell_x, y, z) = f(x, y, z) \quad f(x, y+\ell_y, z) = f(x, y, z) \quad f(x, y, z+\ell_z) = f(x, y, z)$$

In quantum mechanics, this would typically correspond to the wave function of a particle stuck in a periodic box of dimensions $\ell_x \times \ell_y \times \ell_z$. When the particle leaves such a box through one side, it reenters it at the same time through the opposite side.

There are now wave numbers for each direction,

$$k_x = n_x \frac{2\pi}{\ell_x} \qquad k_y = n_y \frac{2\pi}{\ell_y} \qquad k_z = n_z \frac{2\pi}{\ell_z}$$

where n_x, n_y, and n_z are whole numbers. For brevity, vector notations may be used:

$$\vec{r} \equiv x\hat{\imath} + y\hat{\jmath} + z\hat{k} \qquad \vec{k} \equiv k_x\hat{\imath} + k_y\hat{\jmath} + k_z\hat{k} \qquad e^{\mathrm{i}k_x x} e^{\mathrm{i}k_y y} e^{\mathrm{i}k_z z} = e^{\mathrm{i}\vec{k}\cdot\vec{r}}$$

Here $\vec{k}$ is the "wave number vector."

The Fourier series for a three-dimensional periodic function is

$$\boxed{f(\vec{r}) = \sum_{\text{all } \vec{k}} f_{\vec{k}} \frac{e^{\mathrm{i}\vec{k}\cdot\vec{r}}}{\sqrt{\mathcal{V}}} \qquad f_{\vec{k}} = \int_{\mathcal{V}} f(\vec{r}) \frac{e^{-\mathrm{i}\vec{k}\cdot\vec{r}}}{\sqrt{\mathcal{V}}} \,\mathrm{d}^3\vec{r}} \tag{A.201}$$

Here $f(\vec{r})$ is shorthand for $f(x, y, z)$ and $\mathcal{V} = \ell_x \ell_y \ell_z$ is the volume of the periodic box.

The above expression for f may be derived by applying the one-dimensional transform in each direction in turn:

$$\begin{aligned} f(x,y,z) &= \sum_{\text{all } k_x} f_{k_x}(y,z)\frac{e^{\mathrm{i}k_x x}}{\sqrt{\ell_x}} \\ &= \sum_{\text{all } k_x}\sum_{\text{all } k_y} f_{k_x k_y}(z)\frac{e^{\mathrm{i}k_x x}}{\sqrt{\ell_x}}\frac{e^{\mathrm{i}k_y y}}{\sqrt{\ell_y}} \\ &= \sum_{\text{all } k_x}\sum_{\text{all } k_y}\sum_{\text{all } k_z} f_{k_x k_y k_z}\frac{e^{\mathrm{i}k_x x}}{\sqrt{\ell_x}}\frac{e^{\mathrm{i}k_y y}}{\sqrt{\ell_y}}\frac{e^{\mathrm{i}k_z z}}{\sqrt{\ell_z}} \end{aligned}$$

This is equivalent to what is given above, except for trivial changes in notation. The expression for the Fourier coefficients can be derived analogous to the one-dimensional case, integrating now over the entire periodic box.

The Parseval equality still applies

$$\boxed{\langle f(\vec{r})|g(\vec{r})\rangle = \langle f_{\vec{k}}|g_{\vec{k}}\rangle} \tag{A.202}$$

where the left inner product integration is over the periodic box.

For infinite space

$$\boxed{f(\vec{r}) = \frac{1}{\sqrt{2\pi}^3}\int_{\text{all } \vec{k}} f(\vec{k})e^{\mathrm{i}\vec{k}\cdot\vec{r}}\,\mathrm{d}^3\vec{k} \qquad f(\vec{k}) = \frac{1}{\sqrt{2\pi}^3}\int_{\text{all } \vec{r}} f(\vec{r})e^{-\mathrm{i}\vec{k}\cdot\vec{r}}\,\mathrm{d}^3\vec{r}} \tag{A.203}$$

$$\boxed{f(\vec{r}) = \frac{1}{\sqrt{2\pi\hbar}^3}\int_{\text{all } \vec{p}} f(\vec{p})e^{\mathrm{i}\vec{p}\cdot\vec{r}/\hbar}\,\mathrm{d}^3\vec{p} \qquad f(\vec{p}) = \frac{1}{\sqrt{2\pi\hbar}^3}\int_{\text{all } \vec{r}} f(\vec{r})e^{-\mathrm{i}\vec{p}\cdot\vec{r}/\hbar}\,\mathrm{d}^3\vec{r}} \tag{A.204}$$

$$\boxed{\langle f(\vec{r})|g(\vec{r})\rangle = \left\langle f(\vec{k})\middle|g(\vec{k})\right\rangle = \langle f(\vec{p})|g(\vec{p})\rangle} \tag{A.205}$$

These expressions are all obtained completely analogously to the one-dimensional case.

Often, the function is a vector rather than a scalar. That does not make a real difference since each component transforms the same way. Just put a vector symbol over f and g in the above formulae. The inner products are now defined as

$$\left\langle \vec{f}(\vec{r})|\vec{g}(\vec{r})\right\rangle \equiv \langle f_x(\vec{r})|g_x(\vec{r})\rangle + \langle f_y(\vec{r})|g_y(\vec{r})\rangle + \langle f_z(\vec{r})|g_z(\vec{r})\rangle$$

$$\left\langle \vec{f}_{\vec{k}}|\vec{g}_{\vec{k}}\right\rangle \equiv \langle f_{\vec{k}x}|g_{\vec{k}x}\rangle + \left\langle f_{\vec{k}y}\middle|g_{\vec{k}y}\right\rangle + \langle f_{\vec{k}z}|g_{\vec{k}z}\rangle$$

For the picky, converting Fourier series into Fourier integrals only works for well-behaved functions. But to show that it also works for nasty wave functions, you can set up a limiting process in which you approximate the nasty functions increasingly accurately using well-behaved ones. Now if the well-behaved functions are converging, then their Fourier transforms are too. The inner products of the differences in functions are the same according to Parseval. And according to the abstract Lebesgue variant of the theory of integration, that is enough to ensure that the transform of the nasty function exists. This works as long as the nasty wave function is square integrable. And wave functions need to be in quantum mechanics.

But being square integrable is not a strict requirement, as you may have been told elsewhere. A lot of functions that are not square integrable have meaningful, invertible Fourier transforms. For example, functions whose square magnitude integrals are infinite, but absolute value integrals are finite can still be meaningfully transformed. That is more or less the classical version of the inversion theorem, in fact. (See D.C. Champeney, A Handbook of Fourier Theorems, for more.)

A.27 Details of the animations

This note explains how the wave packet animations of chapter 7.11 and 7.12 were obtained. If you want a better understanding of unsteady solutions of the Schrödinger equation and their boundary conditions, this is a good place to start. In fact, deriving such solutions is a popular item in quantum mechanics books for physicists.

First consider the wave packet of the particle in free space, as shown in chapter 7.11.1. An energy eigenfunction with energy E in free space takes the general form

$$\psi_E = C_{\rm f} e^{{\rm i}px/\hbar} + C_{\rm b} e^{-{\rm i}px/\hbar} \qquad p = \sqrt{2mE}$$

where p is the momentum of the particle and $C_{\rm f}$ and $C_{\rm b}$ are constants.

To study a single wave packet coming in from the far left, the coefficient $C_{\rm b}$ has to be set to zero. The reason was worked out in chapter 7.10: combinations of exponentials of the form $C_{\rm b} e^{-{\rm i}px/\hbar}$ produce wave packets that propagate backwards in x, from right to left. Therefore, a nonzero value for $C_{\rm b}$ would add an unwanted second wave packet coming in from the far right.

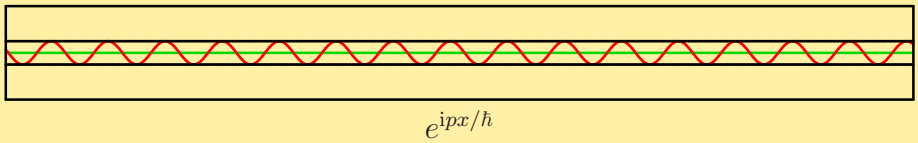


Figure A.9: Example energy eigenfunction for the particle in free space.

With only the coefficient $C_{\rm f}$ of the forward moving part left, you may as well scale the eigenfunction so that $C_{\rm f} = 1$, simplifying it to

$$\psi_E = e^{{\rm i}px/\hbar}$$

A typical example is shown in figure A.9. Plus and minus the magnitude of the eigenfunction are shown in black, and the real part is shown in red. This wave function is an eigenfunction of linear momentum, with p the linear momentum.

To produce a coherent wave packet, eigenfunctions with somewhat different energies E have to be combined together. Since the momentum is given by $p = \sqrt{2mE}$, different energy means different momentum p; therefore the wave packet can be written as

$$\Psi(x,t) = \int_{\text{all } p} c(p) e^{-{\rm i}Et/\hbar} \psi_E(x)\,{\rm d}p \tag{A.206}$$

where $c(p)$ is some function that is only nonzero in a relatively narrow range of momenta p around the nominal momentum. Except for that basic requirement, the choice of the function $c(p)$ is quite arbitrary. Choose some suitable function $c(p)$, then use a computer to numerically integrate the above integral at a large number of plot points and times. Dump the results into your favorite animation software and bingo, out comes the movie.

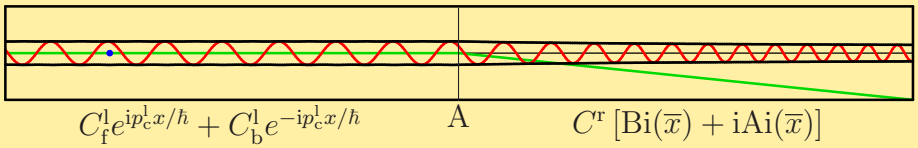


Figure A.10: Example energy eigenfunction for a particle entering a constant accelerating force field.

Next consider the animation of chapter 7.11.2, where the particle accelerates along a downward potential energy ramp starting from point A. A typical energy eigenfunction is shown in figure A.10. Since to the left of point A, the potential energy is still zero, in that region the energy eigenfunction is still of the form

$$\psi_E = C^{\rm l}_{\rm f} e^{{\rm i}p^{\rm l}_{\rm c}x/\hbar} + C^{\rm l}_{\rm b} e^{-{\rm i}p^{\rm l}_{\rm c}x/\hbar} \text{ for } x < x_{\rm A} \qquad p^{\rm l}_{\rm c} = \sqrt{2mE}$$

where $p^{\rm l}_{\rm c}$ is the momentum that a classical particle of energy E would have in the left region. (Quantum mechanics looks at the complete wave function, not just a single point of it, and would say that the momentum is uncertain.)

In this case, it can no longer be argued that the coefficient $C^{\rm l}_{\rm b}$ must be zero to avoid a packet entering from the far right. After all, the $C^{\rm l}_{\rm b} e^{-{\rm i}p^{\rm l}_{\rm c}x/\hbar}$ term does not extend to the far right anymore. To the right of point A, the potential changes linearly with position, and the exponentials are no longer valid.

In fact, it is known that the solution of the Hamiltonian eigenvalue problem in a region with a linearly varying potential is a combination of two weird functions Ai and Bi that are called the "Airy" functions. The bad news is that if you are interested in learning more about their properties, you will need an advanced mathematical handbook like [1] or at least look at addendum {A.29}. The good news is that free software to evaluate these functions and their first derivatives is readily available on the web. The general solution for a linearly varying potential is of the form

$$\psi_E = C_{\rm B}{\rm Bi}(\overline{x}) + C_{\rm A}{\rm Ai}(\overline{x}) \qquad \overline{x} = \sqrt[3]{\frac{2mV'}{\hbar^2}}\frac{V-E}{V'} \quad V' \equiv \frac{{\rm d}V}{{\rm d}x}$$

Note that $(V-E)/V'$ is the x-position measured from the point where $V=E$. Also note that the cube root is negative, so that $\overline{x}$ is.

It may be deduced from the approximate analysis of addendum {A.28} that to prevent a second wave packet coming in from the far right, Ai and Bi must appear together in the combination Bi + iAi as shown in figure A.10. The fact that no second packet comes in from the far right in the animation can be taken as an experimental confirmation of that result, so there seems little justification to go over the messy argument.

To complete the determination of the eigenfunction for a given value of E, the constants $C_{\rm f}^{\rm l}$, $C_{\rm b}^{\rm l}$ and $C^{\rm r}$ must still be determined. That goes as follows. For now, assume that $C^{\rm r}$ has the provisional value $c^{\rm r}=1$. Then provisional values $c_{\rm f}^{\rm l}$ and $c_{\rm b}^{\rm l}$ for the other two constants may be found from the requirements that the left and right regions give the same values for ψ_E and ${\rm d}\psi_E/{\rm d}x$ at the point A in figure A.10 where they meet:

$$c_{\rm f}^{\rm l} e^{{\rm i}p_{\rm c}^{\rm l}x_{\rm A}/\hbar} + c_{\rm b}^{\rm l} e^{-{\rm i}p_{\rm c}^{\rm l}x_{\rm A}/\hbar} = c^{\rm r}\left[{\rm Bi}(\overline{x}_{\rm A}) + {\rm iAi}(\overline{x}_{\rm A})\right]$$

$$c_{\rm f}^{\rm l}\frac{{\rm i}p_{\rm c}^{\rm l}}{\hbar} e^{{\rm i}p_{\rm c}^{\rm l}x_{\rm A}/\hbar} - c_{\rm b}^{\rm l}\frac{{\rm i}p_{\rm c}^{\rm l}}{\hbar} e^{-{\rm i}p_{\rm c}^{\rm l}x_{\rm A}/\hbar} = c^{\rm r}\left[{\rm Bi}'(\overline{x}_{\rm A}) + {\rm iAi}'(\overline{x}_{\rm A})\right]\frac{{\rm d}\overline{x}}{{\rm d}x}$$

That is equivalent to two equations for the two constants $c_{\rm f}^{\rm l}$ and $c_{\rm b}^{\rm l}$, since everything else can be evaluated, using the mentioned software. So $c_{\rm f}^{\rm l}$ and $c_{\rm b}^{\rm l}$ can be found from solving these two equations.

As the final step, it is desirable to normalize the eigenfunction ψ_E so that $C_{\rm f}^{\rm l}=1$. To do so, the entire provisional eigenfunction can be divided by $c_{\rm f}^{\rm l}$, giving $C_{\rm b}^{\rm l} = c_{\rm b}^{\rm l}/c_{\rm f}^{\rm l}$ and $C^{\rm r} = c^{\rm r}/c_{\rm f}^{\rm l}$. The energy eigenfunction has now been found. And since $C_{\rm f}^{\rm l}=1$, the $e^{{\rm i}p_{\rm c}^{\rm l}x/\hbar}$ term is exactly the same as the free space energy eigenfunction of the first example. That means that if the eigenfunctions ψ_E are combined into a wave packet in the same way as in the free space case, (A.206) with p replaced by $p_{\rm c}^{\rm l}$, the $e^{{\rm i}p_{\rm c}^{\rm l}x/\hbar}$ terms produce the exact same wave packet coming in from the far left as in the free space case.

For larger times, the $C_{\rm b}^{\rm l}e^{-{\rm i}p_{\rm c}^{\rm l}x/\hbar}$ terms produce a "reflected" wave packet that returns toward the far left. Note that $e^{-{\rm i}p_{\rm c}^{\rm l}x/\hbar}$ is the complex conjugate of $e^{{\rm i}p_{\rm c}^{\rm l}x/\hbar}$,

and it can be seen from the unsteady Schrödinger equation that if the complex conjugate of a wave function is taken, it produces a reversal of time. Wave packets coming in from the far left at large negative times become wave packets leaving toward the far left at large positive times. However, the constant $C_{\rm b}^{\rm l}$ turns out to be very small in this case, so there is little reflection.

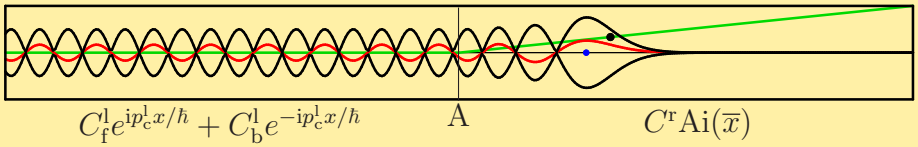


Figure A.11: Example energy eigenfunction for a particle entering a constant decelerating force field.

Next consider the animation of chapter 7.11.3, where the particle is turned back by an upward potential energy ramp. A typical energy eigenfunction for this case is shown in figure A.11. Unlike in the previous example, where the argument $\overline{x}$ of the Airy functions was negative at the far right, here it is positive. Table books that cover the Airy functions will tell you that the Airy function Bi blows up very strongly with increasing positive argument $\overline{x}$. Therefore, if the solution in the right hand region would involve any amount of Bi, it would locate the particle at infinite x for all times. For a particle not at infinity, the solution in the right hand region can only involve the Airy function Ai. That function decays rapidly with positive argument $\overline{x}$, as seen in figure A.11.

The further determination of the energy eigenfunctions proceeds along the same lines as in the previous example: give $C^{\rm r}$ a provisional value $c^{\rm r} = 1$, then compute $c_{\rm f}^{\rm l}$ and $c_{\rm b}^{\rm l}$ from the requirements that the left and right regions produce the same values for ψ and ${\rm d}\psi/{\rm d}x$ at the point A where they meet. Finally divide the eigenfunction by $c_{\rm f}^{\rm l}$. The big difference is that now $C_{\rm b}^{\rm l}$ is no longer small; $C_{\rm b}^{\rm l}$ turns out to be of unit magnitude just like $C_{\rm f}^{\rm l}$. It means that the incoming wave packet is reflected back completely.

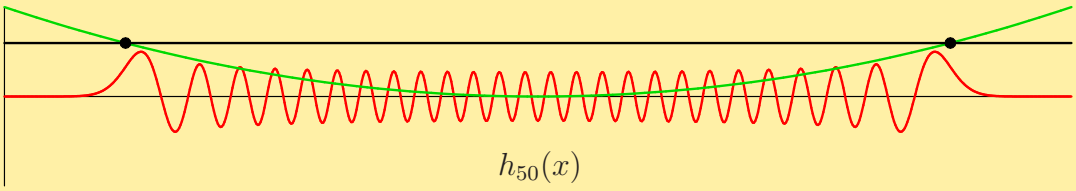


Figure A.12: Example energy eigenfunction for the harmonic oscillator.

For the harmonic oscillator of chapter 7.11.4, the analysis is somewhat different. In particular, chapter 4.1.2 showed that the energy levels of the one-dimensional harmonic oscillator are discrete,

$$E_n = \frac{2n+1}{2}\hbar\omega \text{ for } n = 0, 1, 2, \ldots$$

so that unlike the motions just discussed, the solution of the Schrödinger equation is a sum, rather than the integral (A.206),

$$\Psi(x,t) = \sum_{n=0}^{\infty} c_n e^{-\mathrm{i}E_n t/\hbar} h_n(x)$$

However, for large n the difference between summation and integration is small.

Also, while the energy eigenfunctions $h_n(x)$ are not exponentials as for the free particle, for large n they can be pairwise combined to approximate such exponentials. For example, eigenfunction h_{50}, shown in figure A.12, behaves near the center point much like a cosine if you scale it properly. Similarly, h_{51} behaves much like a sine. A cosine plus i times a sine gives an exponential, according to the Euler formula (2.5). Create similar exponential combinations of eigenfunctions with even and odd values of n for a range of n values, and there are the approximate exponentials that allow you to create a wave packet that is at the center point at time $t = 0$. In the animation, the range of n values was centered around $n = 50$, making the nominal energy hundred times the ground state energy. The exponentials degenerate over time, since their component eigenfunctions have slightly different energy, hence time evolution. That explains why after some time, the wave packet can return to the center point going the other way.

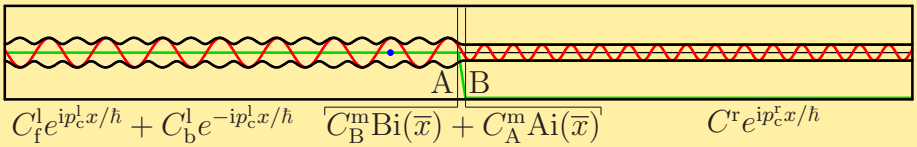


Figure A.13: Example energy eigenfunction for a particle encountering a brief accelerating force.

For the particle of chapter 7.12.1 that encounters a brief accelerating force, an example eigenfunction looks like figure A.13. In this case, the solution in the far right region is similar to the one in the far left region. However, there cannot be a term of the form $e^{-\mathrm{i}p^{\mathrm{r}}_{\mathrm{c}}x/\hbar}$ in the far right region, because when the eigenfunctions are combined, it would produce an unwanted wave packet coming in from the far right. In the middle region of linearly varying potential, the wave function is again a combination of the two Airy functions. The way to find the constants now has an additional step. First give the constant C^{r} of the far right exponential the provisional value $c^{\mathrm{r}} = 1$ and from that, compute provisional values $c^{\mathrm{m}}_{\mathrm{A}}$ and $c^{\mathrm{m}}_{\mathrm{B}}$ by demanding that the Airy functions give the same values for ψ and $\mathrm{d}\psi/\mathrm{d}x$ as the far-right exponential at point B, where they meet. Next compute provisional values $c^{\mathrm{l}}_{\mathrm{f}}$ and $c^{\mathrm{l}}_{\mathrm{b}}$ by demanding that the far-left exponentials give the same values for ψ and $\mathrm{d}\psi/\mathrm{d}x$ as the Airy functions

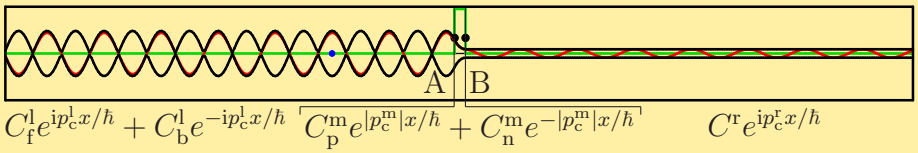


Figure A.14: Example energy eigenfunction for tunneling through a barrier.

at point A, where they meet. Finally, divide all the constants by $c_{\rm f}^{\rm l}$ to make $C_{\rm f}^{\rm l} = 1$.

For the tunneling particle of chapter 7.12.2, an example eigenfunction is as shown in figure A.14. In this case, the solution in the middle part is not a combination of Airy functions, but of real exponentials. It is essentially the same solution as in the left and right parts, but in the middle region the potential energy is greater than the total energy, making $p_{\rm c}^{\rm m} = \sqrt{2m(E - V_{\rm m})}$ an imaginary number. Therefore the arguments of the exponentials become real when written in terms of the absolute value of the momentum $|p_{\rm c}^{\rm m}| = \sqrt{2m(V_{\rm m} - E)}$. The rest of the analysis is similar to that of the previous example.

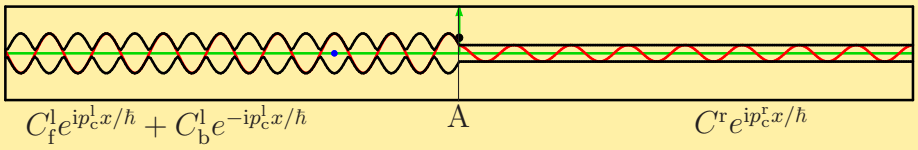


Figure A.15: Tunneling through a delta function barrier.

For the particle tunneling through the delta function potential in chapter 7.12.2, an example energy eigenfunction is shown in figure A.15. The potential energy in this case is $V = \nu\delta(x - x_{\rm A})$, where $\delta(x - x_{\rm A})$ is a spike at point A that integrates to one and the strength ν is a chosen constant. In the example, ν was chosen to be $\sqrt{2\hbar^2 E_{\rm nom}/m}$ with $E_{\rm nom}$ the nominal energy. For that strength, half the wave packet will pass through.

For a delta function potential, a modification must be made in the analysis as used so far. As figure A.15 illustrates, there are kinks in the energy eigenfunction at the location A of the delta function. The left and right expressions for the eigenfunction *do not* predict the same value for its derivative $\mathrm{d}\psi/\mathrm{d}x$ at point A. To find the difference, integrate the Hamiltonian eigenvalue problem from a point a very short distance ε before point A to a point the same very short distance behind it:

$$-\frac{\hbar^2}{2m}\int_{x=x_{\rm A}-\varepsilon}^{x_{\rm A}+\varepsilon} \frac{\mathrm{d}^2\psi}{\mathrm{d}x^2}\,\mathrm{d}x + \nu\int_{x=x_{\rm A}-\varepsilon}^{x_{\rm A}+\varepsilon} \delta(x - x_{\rm A})\psi\,\mathrm{d}x = \int_{x=x_{\rm A}-\varepsilon}^{x_{\rm A}+\varepsilon} E\psi\,\mathrm{d}x$$

The integral in the right hand side is zero because of the vanishingly small interval of integration. But the delta function spike in the left hand side integrates

to one regardless of the small integration range, so

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}\psi}{\mathrm{d}x}\Big|_{x_\mathrm{A}-\varepsilon}^{x_\mathrm{A}+\varepsilon} + \nu\psi(x_\mathrm{A}) = 0$$

For vanishingly small ε, $\mathrm{d}\psi/\mathrm{d}x$ at $x_\mathrm{A}+\varepsilon$ becomes what the right hand part of the eigenfunction gives for $\mathrm{d}\psi/\mathrm{d}x$ at x_A, while $\mathrm{d}\psi/\mathrm{d}x$ at $x_\mathrm{A}-\varepsilon$ becomes what the left hand part gives for it. As seen from the above equation, the difference is not zero, but $2m\nu\psi(x_\mathrm{A})/\hbar^2$.

So the correct equations for the provisional constants are in this case

$$c_\mathrm{f}^\mathrm{l} e^{\mathrm{i}p_\mathrm{c}^\mathrm{l} x_\mathrm{A}/\hbar} + c_\mathrm{b}^\mathrm{l} e^{-\mathrm{i}p_\mathrm{c}^\mathrm{l} x_\mathrm{A}/\hbar} = c^\mathrm{r} e^{\mathrm{i}p_\mathrm{c}^\mathrm{r} x_\mathrm{A}/\hbar}$$

$$\frac{\mathrm{i}p_\mathrm{c}^\mathrm{l}}{\hbar} c_\mathrm{f}^\mathrm{l} e^{\mathrm{i}p_\mathrm{c}^\mathrm{l} x_\mathrm{A}/\hbar} - \frac{\mathrm{i}p_\mathrm{c}^\mathrm{l}}{\hbar} c_\mathrm{b}^\mathrm{l} e^{-\mathrm{i}p_\mathrm{c}^\mathrm{l} x_\mathrm{A}/\hbar} = \frac{\mathrm{i}p_\mathrm{c}^\mathrm{r}}{\hbar} c^\mathrm{r} e^{\mathrm{i}p_\mathrm{c}^\mathrm{r} x_\mathrm{A}/\hbar} - \frac{2m\nu}{\hbar^2} c^\mathrm{r} e^{\mathrm{i}p_\mathrm{c}^\mathrm{r} x_\mathrm{A}/\hbar}$$

Compared to the analysis as used previously, the difference is the final term in the second equation that is added by the delta function.

The remainder of this note gives some technical details for if you are actually planning to do your own animations. It is a good idea to assume that the units of mass, length, and time are chosen such that $\hbar$ and the nominal energy are one, while the mass of the particle is one-half. That avoids having to guesstimate suitable values for all sorts of very small numbers. The Hamiltonian eigenvalue problem then simplifies to

$$-\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V\psi = E\psi$$

where the values of E of interest cluster around 1. The nominal momentum will be one too. In those units, the length of the plotted range was one hundred in all but the harmonic oscillator case.

It should be noted that to select a good function $c(p)$ in (A.206) is somewhat of an art. The simplest idea would be to choose $c(p)$ equal to one in some limited range around the nominal momentum, and zero elsewhere, as in

$$c(p) = 1 \quad \text{if } (1-r)p_\mathrm{nom} < p < (1+r)p_\mathrm{nom} \qquad c(p) = 0 \quad \text{otherwise}$$

where r is the relative deviation from the nominal momentum below which $c(p)$ is nonzero. However, it is know from Fourier analysis that the locations where $c(p)$ jumps from one to zero lead to lengthy wave packets when viewed in physical space. {D.44}. Functions $c(p)$ that do lead to nice compact wave packets are known to be of the form

$$c(p) = \exp\left(-\frac{(p-p_\mathrm{nom})^2}{r^2 p_\mathrm{nom}^2}\right)$$

And that is essentially the function $c(p)$ used in this study. The typical width of the momentum range was chosen to be $r = 0.15$, or 15%, by trial and error.

However, it is nice if $c(p)$ becomes not just very small, but exactly zero beyond some point, for one because it cuts down on the number of energy eigenfunctions that have to be evaluated numerically. Also, it is nice not to have to worry about the possibility of p being negative in writing energy eigenfunctions. Therefore, the final function used was

$$c(p) = \exp\left(-\frac{(p-p_{\rm nom})^2}{r^2[p_{\rm nom}^2-(p-p_{\rm nom})^2]}\right) \text{ for } 0<p<2p_{\rm nom} \quad c(p)=0 \text{ otherwise}$$

The actual difference in numerical values is small, but it does make $c(p)$ exactly zero for negative momenta and those greater than twice the nominal value. Strictly speaking, $c(p)$ should still be multiplied by a constant to make the total probability of finding the particle equal to one. But if you do not tell people what numbers for Ψ are on the vertical axes, you do not need to bother.

In doing the numerical integrations to find $\Psi(x,t)$, note that the mid point and trapezium rules of numerical integration are exponentially accurate under the given conditions, so there is probably not much motivation to try more advanced methods. The mid point rule was used.

The animations in this book used the numerical implementations `daie.f`, `dbie.f`, `daide.f`, and `dbide.f` from netlib.org for the Airy functions and their first derivatives. These offer some basic protection against underflow and overflow by splitting off an exponential for positive $\overline{x}$. It may be a good idea to check for underflow and overflow in general and use 64 bit precision. The examples here did.

For the harmonic oscillator, the larger the nominal energy is compared to the ground state energy, the more the wave packet can resemble a single point compared to the limits of motion. However, the computer program used to create the animation computed the eigenfunctions by evaluating the analytical expression given in derivation {D.12}, and explicitly evaluating the Hermite polynomials is very round-off sensitive. That limited it to a maximum of about hundred times the ground state energy when allowing for enough uncertainty to localize the wave packet. Round-off is a general problem for power series, not just for the Hermite polynomials. If you want to go to higher energies to get a smaller wave packet, you will want to use a finite difference or finite element method to find the eigenfunctions.

The plotting software used to produce the animations was a mixture of different programs. There are no doubt much simpler and better ways of doing it. In the animations presented here, first plots were created of Ψ versus x for a large number of closely spaced times covering the duration of the animation. These plots were converted to gifs using a mixture of personal software, netpbm, and ghostview. The gifs were then combined into a single movie using gifsicle.

A.28 WKB Theory of Nearly Classical Motion

WKB theory provides simple approximate solutions for the energy eigenfunctions when the conditions are almost classical, like for the wave packets of chapter 7.11. The approximation is named after Wentzel, Kramers, and Brillouin, who refined the ideas of Liouville and Green. The bandit scientist Jeffreys tried to rob WKB of their glory by doing the same thing two years earlier, and is justly denied all credit.

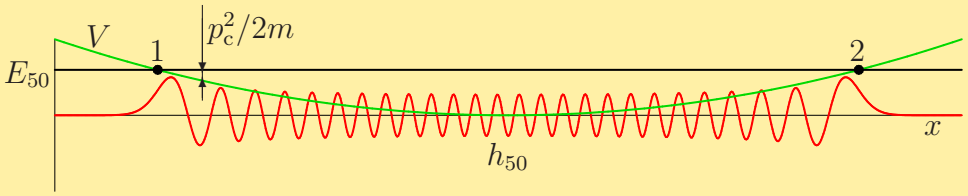


Figure A.16: Harmonic oscillator potential energy V, eigenfunction h_{50}, and its energy E_{50}.

The WKB approximation is based on the rapid spatial variation of energy eigenfunctions with almost macroscopic energies. As an example, figure A.16 shows the harmonic oscillator energy eigenfunction h_{50}. Its energy E_{50} is hundred times the ground state energy. That makes the kinetic energy $E-V$ quite large over most of the range, and that in turn makes the linear momentum large. In fact, the classical Newtonian linear momentum $p_c = mv$ is given by

$$\boxed{p_c \equiv \sqrt{2m(E-V)}} \tag{A.207}$$

In quantum mechanics, the large momentum implies rapid oscillation of the wave function: quantum mechanics associates the linear momentum with the operator $\hbar \mathrm{d}/\mathrm{i}\mathrm{d}x$ that denotes spatial variation.

The WKB approximation is most appealing in terms of the classical momentum p_c as defined above. To find its form, in the Hamiltonian eigenvalue problem

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V\psi = E\psi$$

take the $V\psi$ term to the other side and then rewrite $E-V$ in terms of the classical linear momentum. That produces

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = -\frac{p_c^2}{\hbar^2}\psi \tag{A.208}$$

Now under almost classical conditions, a single period of oscillation of the wave function is so short that normally p_c is almost constant over it. Then

by approximation the solution of the eigenvalue problem over a single period is simply an arbitrary combination of two exponentials,

$$\psi \sim c_{\rm f} e^{{\rm i}p_{\rm c}x/\hbar} + c_{\rm b} e^{-{\rm i}p_{\rm c}x/\hbar}$$

where the constants $c_{\rm f}$ and $c_{\rm b}$ are arbitrary. (The subscripts denote whether the wave speed of the corresponding term is forward or backward.)

It turns out that to make the above expression work over more than one period, it is necessary to replace $p_{\rm c}x$ by the antiderivative $\int p_{\rm c}\,{\rm d}x$. Furthermore, the "constants" $c_{\rm f}$ and $c_{\rm b}$ must be allowed to vary from period to period proportional to $1/\sqrt{p_{\rm c}}$.

In short, the WKB approximation of the wave function is, {D.46}:

$$\boxed{\text{classical WKB:} \qquad \psi \approx \frac{1}{\sqrt{p_{\rm c}}} \left[C_{\rm f} e^{{\rm i}\theta} + C_{\rm b} e^{-{\rm i}\theta} \right] \qquad \theta \equiv \frac{1}{\hbar} \int p_{\rm c}\,{\rm d}x} \tag{A.209}$$

where $C_{\rm f}$ and $C_{\rm b}$ are now true constants.

If you ever glanced at notes such as {D.12}, {D.14}, and {D.15}, in which the eigenfunctions for the harmonic oscillator and hydrogen atom were found, you recognize what a big simplification the WKB approximation is. Just do the integral for θ and that is it. No elaborate transformations and power series to grind down. And the WKB approximation can often be used where no exact solutions exist at all.

In many applications, it is more convenient to write the WKB approximation in terms of a sine and a cosine. That can be done by taking the exponentials apart using the Euler formula (2.5). It produces

$$\boxed{\text{rephrased WKB:} \qquad \psi \approx \frac{1}{\sqrt{p_{\rm c}}} \left[C_{\rm c} \cos\theta + C_{\rm s} \sin\theta \right] \qquad \theta \equiv \frac{1}{\hbar} \int p_{\rm c}\,{\rm d}x} \tag{A.210}$$

The constants $C_{\rm c}$ and $C_{\rm s}$ are related to the original constants $C_{\rm f}$ and $C_{\rm b}$ as

$$\boxed{C_{\rm c} = C_{\rm f} + C_{\rm b} \quad C_{\rm s} = {\rm i}C_{\rm f} - {\rm i}C_{\rm b} \qquad C_{\rm f} = \tfrac{1}{2}(C_{\rm c} - {\rm i}C_{\rm s}) \quad C_{\rm b} = \tfrac{1}{2}(C_{\rm c} + {\rm i}C_{\rm s})} \tag{A.211}$$

which allows you to convert back and forward between the two formulations as needed. Do note that either way, the constants depend on what you chose for the integration constant in the θ integral.

As an application, consider a particle stuck between two impenetrable walls at positions x_1 and x_2. An example would be the particle in a pipe that was studied way back in chapter 3.5. The wave function ψ must become zero at both x_1 and x_2, since there is zero possibility of finding the particle outside the impenetrable walls. It is now smart to chose the integration constant in θ so that $\theta_1 = 0$. In that case, $C_{\rm c}$ must be zero for ψ to be zero at x_1, (A.210). The

wave function must be just the sine term. Next, for ψ *also* to be zero at x_2, θ_2 must be a whole multiple n of π, because that are the only places where sines are zero. So $\theta_2 - \theta_1 = n\pi$, which means that

$$\boxed{\text{particle between impenetrable walls:} \quad \frac{1}{\hbar}\int_{\underline{x}=x_1}^{x_2} p_c(\underline{x})\,\mathrm{d}\underline{x} = n\pi} \tag{A.212}$$

Recall that p_c was $\sqrt{2m(E-V)}$, so this is just an equation for the energy eigenvalues. It is an equation involving just an integral; it does not even require you to find the corresponding eigenfunctions!

It does get a bit more tricky for a case like the harmonic oscillator where the particle is not caught between impenetrable walls, but merely prevented to escape by a gradually increasing potential. Classically, such a particle would still be rigorously constrained between the so called "turning points" where the potential energy V becomes equal to the total energy E, like the points 1 and 2 in figure A.16. But as the figure shows, in quantum mechanics the wave function does not become zero at the turning points; there is some chance for the particle to be found somewhat beyond the turning points.

A further complication arises since the WKB approximation becomes inaccurate in the immediate vicinity of the turning points. The problem is the requirement that the classical momentum can be approximated as a nonzero constant on a small scale. At the turning points the momentum becomes zero and that approximation fails.

However, it is possible to solve the Hamiltonian eigenvalue problem near the turning points assuming that the potential energy is not constant, but varies approximately linearly with position, {A.29}. Doing so and fixing up the WKB solution away from the turning points produces a simple result. The classical WKB approximation remains a sine, but at the turning points, $\sin\theta$ stays an angular amount $\pi/4$ short of becoming zero. (Or to be precise, it just seems to stay $\pi/4$ short, because the classical WKB approximation is no longer valid at the turning points.) Assuming that there are turning points with gradually increasing potential at both ends of the range, like for the harmonic oscillator, the total angular range will be short by an amount $\pi/2$.

Therefore, the expression for the energy eigenvalues becomes:

$$\boxed{\text{particle trapped between turning points:} \quad \frac{1}{\hbar}\int_{\underline{x}=x_1}^{x_2} p_c(\underline{x})\,\mathrm{d}\underline{x} = (n-\tfrac{1}{2})\pi} \tag{A.213}$$

The WKB approximation works fine in regions where the total energy E is less than the potential energy V. The classical momentum $p_c = \sqrt{2m(E-V)}$ is imaginary in such regions, reflecting the fact that classically the particle does not have enough energy to enter them. But, as the nonzero wave function beyond the turning points in figure A.16 shows, quantum mechanics does allow

some possibility for the particle to be found in regions where E is less than V. It is loosely said that the particle can "tunnel" through, after a popular way for criminals to escape from jail. To use the WKB approximation in these regions, just rewrite it in terms of the magnitude $|p_c| = \sqrt{2m(V-E)}$ of the classical momentum:

$$\boxed{\text{tunneling WKB:} \qquad \psi \approx \frac{1}{\sqrt{|p_c|}}\left[C_p e^{\gamma} + C_n e^{-\gamma}\right] \quad \gamma \equiv \frac{1}{\hbar}\int |p_c|\,\mathrm{d}x} \tag{A.214}$$

Note that γ is the equivalent of the angle θ in the classical approximation.

Key Points

- The WKB approximation applies to situations of almost macroscopic energy.
- The WKB solution is described in terms of the classical momentum $p_c \equiv \sqrt{2m(E-V)}$ and in particular its antiderivative $\theta = \int p_c\,\mathrm{d}x/\hbar$.
- The wave function can be written as (A.209) or (A.210), whatever is more convenient.
- For a particle stuck between impenetrable walls, the energy eigenvalues can be found from (A.212).
- For a particle stuck between a gradually increasing potential at both sides, the energy eigenvalues can be found from (A.213).
- The "tunneling" wave function in regions that classically the particle is forbidden to enter can be approximated as (A.214). It is in terms of the antiderivative $\gamma = \int |p_c|\,\mathrm{d}x/\hbar$.

A.28 Review Questions

1. Use the equation
$$\frac{1}{\hbar}\int_{\underline{x}=x_1}^{x_2} p_c(\underline{x})\,\mathrm{d}\underline{x} = n\pi$$
to find the WKB approximation for the energy levels of a particle stuck in a pipe of chapter 3.5.5. The potential V is zero inside the pipe, given by $0 \leqslant x \leqslant \ell_x$

   In this case, the WKB approximation produces the exact result, since the classical momentum really is constant. If there was a force field in the pipe, the solution would only be approximate.

   Solution wkb-a
2. Use the equation
$$\frac{1}{\hbar}\int_{\underline{x}=x_1}^{x_2} p_c(\underline{x})\,\mathrm{d}\underline{x} = (n-\tfrac{1}{2})\pi$$
to find the WKB approximation for the energy levels of the harmonic oscillator. The potential energy is $\frac{1}{2}m\omega x^2$ where the constant ω is the

classical natural frequency. So the total energy, expressed in terms of the turning points $x_2 = -x_1$ at which $E = V$, is $E = \frac{1}{2}m\omega {x_2}^2$.

In this case too, the WKB approximation produces the exact energy eigenvalues. That, however, is just a coincidence; the classical WKB wave functions are certainly not exact; they become infinite at the turning points. As the example h_{50} above shows, the true wave functions most definitely do not.

Solution wkb-b

A.29 WKB solution near the turning points

Both the classical and tunneling WKB approximations of addendum {A.28} fail near so-called "turning points" where the classical kinetic energy $E-V$ becomes zero. This note explains how the problem can be fixed.

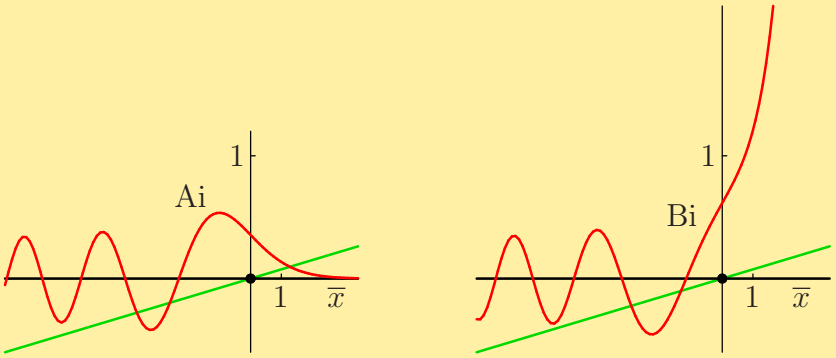


Figure A.17: The Airy Ai and Bi functions that solve the Hamiltonian eigenvalue problem for a linearly varying potential energy. Bi very quickly becomes too large to plot for positive values of its argument.

The trick is to use a different approximation near turning points. In a small vicinity of a turning point, it can normally be assumed that the x-derivative V' of the potential is about constant, so that the potential varies linearly with position. Under that condition, the exact solution of the Hamiltonian eigenvalue problem is known to be a combination of two special functions Ai and Bi that are called the "Airy" functions. These functions are shown in figure A.17. The general solution near a turning point is:

$$\psi = C_{\rm A}{\rm Ai}(\overline{x}) + C_{\rm B}{\rm Bi}(\overline{x}) \qquad \overline{x} = \sqrt[3]{\frac{2mV'}{\hbar^2}}\frac{V-E}{V'} \qquad V' \equiv \frac{{\rm d}V}{{\rm d}x}$$

Note that $(V-E)/V'$ is the x-position measured from the point where $V = E$, so that $\overline{x}$ is a local, stretched x-coordinate.

The second step is to relate this solution to the normal WKB approximations away from the turning point. Now from a macroscopic point of view, the WKB

approximation follows from the assumption that Planck's constant $\hbar$ is very small. That implies that the validity of the Airy functions normally extends to region where $|\overline{x}|$ is relatively large. For example, if you focus attention on a point where $V - E$ is a finite multiple of $\hbar^{1/3}$, $V - E$ is small, so the value of V' will deviate little from its value at the turning point: the assumption of linearly varying potential remains valid. Still, if $V - E$ is a finite multiple of $\hbar^{1/3}$, $|\overline{x}|$ will be proportional to $1/\hbar^{1/3}$, and that is large. Such regions of large, but not too large, $|\overline{x}|$ are called "matching regions," because in them *both* the Airy function solution and the WKB solution are valid. It is where the two meet and must agree.

It is graphically depicted in figures A.18 and A.19. Away from the turning points, the classical or tunneling WKB approximations apply, depending on whether the total energy is more than the potential energy or less. In the vicinity of the turning points, the solution is a combination of the Airy functions. If you look up in a mathematical handbook like [1] how the Airy functions can be approximated for large positive respectively negative $\overline{x}$, you find the expressions listed in the bottom lines of the figures. (After you rewrite what you find in table books in terms of useful quantities, that is!)

The expressions in the bottom lines must agree with what the classical, respectively tunneling WKB approximation say about the matching regions. At one side of the turning point, that relates the coefficients $C_{\rm p}$ and $C_{\rm n}$ of the tunneling approximation to the coefficients of $C_{\rm A}$ and $C_{\rm B}$ of the Airy functions. At the other side, it relates the coefficients $C_{\rm f}$ and $C_{\rm b}$ (or $C_{\rm c}$ and $C_{\rm s}$) of the classical WKB approximation to $C_{\rm A}$ and $C_{\rm B}$. The net effect of it all is to relate, "connect," the coefficients of the classical WKB approximation to those of the tunneling one. That is why the formulae in figures A.18 and A.19 are called the "connection formulae."

You may have noted the appearance of an additional constant c in figures A.18 and A.19. This nasty constant is defined as

$$c = \frac{\sqrt{\pi}}{(2m|V'|\hbar)^{1/6}} \tag{A.215}$$

and shows up uninvited when you approximate the Airy function solution for large $|\overline{x}|$. By cleverly absorbing it in a redefinition of the constants $C_{\rm A}$ and $C_{\rm B}$, figures A.18 and A.19 achieve that you do not have to worry about it unless you specifically need the actual solution at the turning points.

As an example of how the connection formulae are used, consider a right turning point for the harmonic oscillator or similar. Near such a turning point, the connection formulae of figure A.18 apply. In the tunneling region towards the right, the term $C_{\rm p}e^{\gamma}$ better be zero, because it blows up at large x, and that would put the particle at infinity for sure. So the constant $C_{\rm p}$ will have to be zero. Now the matching at the right side equates $C_{\rm p}$ to $C_{\rm B}e^{-\gamma_t}$ so $C_{\rm B}$ will have to be zero. That means that the solution in the vicinity of the turning

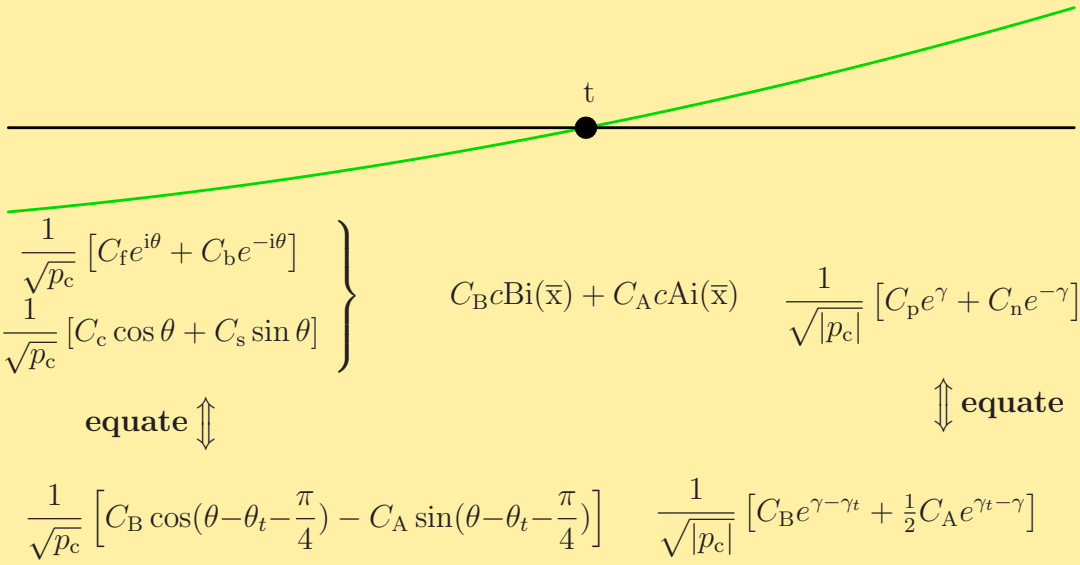


Figure A.18: Connection formulae for a turning point from normal motion to tunneling.

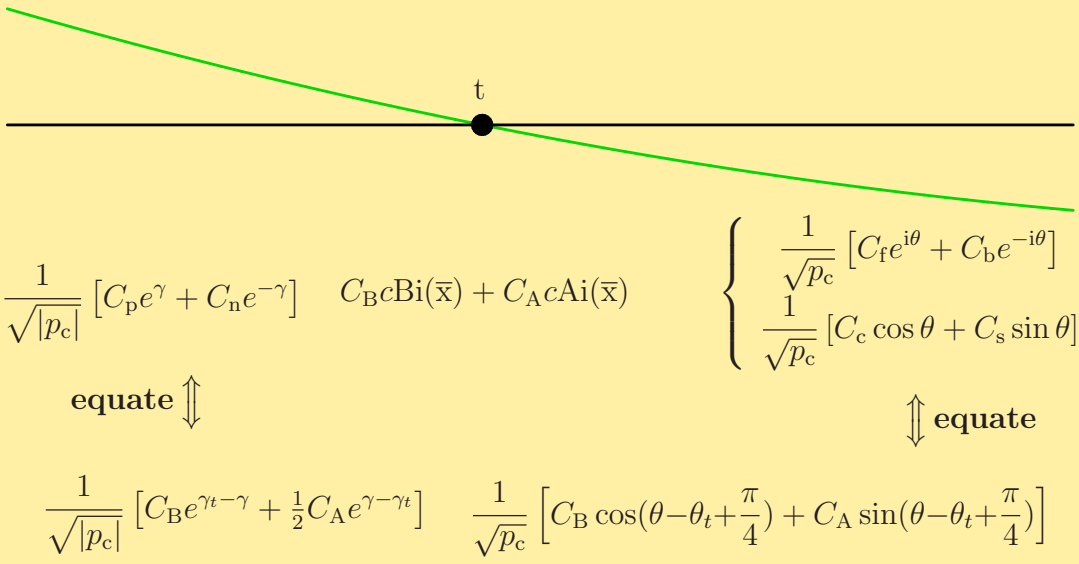


Figure A.19: Connection formulae for a turning point from tunneling to normal motion.

point will have to be a pure Ai function. Then the matching towards the left shows that the solution in the classical WKB region must take the form of a sine that, when extrapolated to the turning point $\theta = \theta_t$, stops short of reaching zero by an angular amount $\pi/4$. Hence the assertion in addendum {A.28} that the angular range of the classical WKB solution should be shortened by $\pi/4$ for each end at which the particle is trapped by a gradually increasing potential instead of an impenetrable wall.

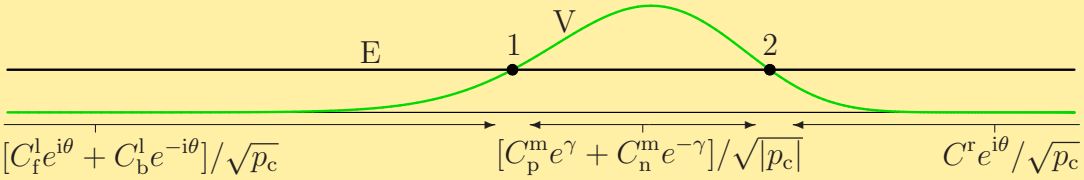


Figure A.20: WKB approximation of tunneling.

As another example, consider tunneling as discussed in chapter 7.12 and 7.13. Figure A.20 shows a sketch. The WKB approximation may be used if the barrier through which the particle tunnels is high and wide. In the far right region, the energy eigenfunction only involves a term $C^{\rm r} e^{{\rm i}\theta}$ with a forward wave speed. To simplify the analysis, the constant $C^{\rm r}$ can be taken to be one, because it does not make a difference how the wave function is normalized. Also, the integration constant in θ can be chosen such that $\theta = \pi/4$ at turning point 2; then the connection formulae of figure A.19 along with the Euler formula (2.5) show that the coefficients of the Airy functions at turning point 2 are $C_{\rm B} = 1$ and $C_{\rm A} = {\rm i}$. Next, the integration constant in γ can be taken such that $\gamma = 0$ at turning point 2; then the connection formulae of figure A.19 imply that $C^{\rm m}_{\rm p} = \frac{1}{2}{\rm i}$ and $C^{\rm m}_{\rm n} = 1$.

Next consider the connection formulae for turning point 1 in figure A.18. Note that $e^{-\gamma_1}$ can be written as $e^{\gamma_{12}}$, where $\gamma_{12} = \gamma_2 - \gamma_1$, because the integration constant in γ was chosen such that $\gamma_2 = 0$. The advantage of using $e^{\gamma_{12}}$ instead of $e^{-\gamma_1}$ is that it is independent of the choice of integration constant. Furthermore, under the typical conditions that the WKB approximation applies, for a high and wide barrier, $e^{\gamma_{12}}$ will be a very large number. It is then seen from figure A.18 that near turning point 1, $C_{\rm A} = 2e^{\gamma_{12}}$ which is large while $C_{\rm B}$ is small and will be ignored. And that then implies, using the Euler formula to convert Ai's sine into exponentials, that $|C^{\rm l}_{\rm f}| = e^{\gamma_{12}}$. As discussed in chapter 7.13, the transmission coefficient is given by

$$T = \frac{p^{\rm r}_{\rm c}}{p^{\rm l}_{\rm c}} \frac{|C^{\rm r}/\sqrt{p^{\rm r}_{\rm c}}|^2}{|C^{\rm l}_{\rm f}/\sqrt{p^{\rm l}_{\rm c}}|^2}$$

and plugging in $C^{\rm r} = 1$ and $|C^{\rm l}_{\rm f}| = e^{\gamma_{12}}$, the transmission coefficient is found to be $e^{-2\gamma_{12}}$.

A.30 Three-dimensional scattering

This note introduces some of the general concepts of three-dimensional scattering, in case you run into them. For more details and actual examples, a quantum mechanics text for physicists will need to be consulted; it is a big thing for them.

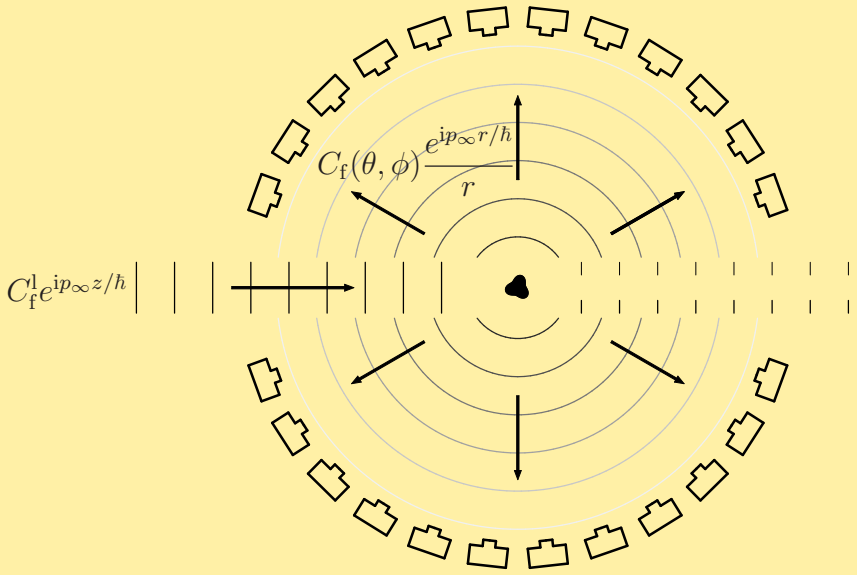


Figure A.21: Scattering of a beam off a target.

The basic idea is as sketched in figure A.21. A beam of particles is sent in from the far left towards a three-dimensional target. Part of the beam hits the target and is scattered, to be picked up by surrounding detection equipment.

It will be assumed that the collision with the target is elastic, and that the particles in the beam are sufficiently light that they scatter off the target without transferring kinetic energy to it. In that case, the target can be modeled as a steady potential energy field. And if the target and/or incoming particles are electrically neutral, it can also be assumed that the potential energy decays fairly quickly to zero away from the target. (In fact, a lot of results in this note turn out not apply to a slowly decaying potential like the Coulomb one.)

It is convenient to use a spherical coordinate system (r, θ, ϕ) with its origin at the scattering object and with its axis aligned with the direction of the incoming beam. Since the axis of a spherical coordinate system is usually called the z-axis, the horizontal coordinate will now be indicated as z, not x like in the one-dimensional analysis done earlier.

In the energy eigenfunctions, the incoming particle beam can be represented as a one-dimensional wave. However, unlike for the one-dimensional scattering of figure 7.22, now the wave is not just scattered to the left and right, but in all directions, in other words to all angles θ and ϕ. The far-field behavior of the

energy eigenfunctions is

$$\boxed{\psi_E \sim C_{\rm l} e^{{\rm i}p_\infty z/\hbar} + C_{\rm f}(\theta,\phi)\frac{e^{{\rm i}p_\infty r/\hbar}}{r} \quad \text{for} \quad r \to \infty \qquad p_\infty \equiv \sqrt{2mE}} \tag{A.216}$$

Here E is the kinetic energy of the incoming particles and m the mass. Therefore p_∞ is what classical physics would take to be the momentum of the particles at infinity. The first term in the far field behavior allows the incoming particles to be described, as well as the same particles going out again unperturbed. If some joker removes the target, that is all there is.

The second term describes the outgoing scattered particles. The constant $C_{\rm f}(\theta,\phi)$ is called the "scattering amplitude." The second term also contains a factor $e^{{\rm i}p_\infty r/\hbar}$ consistent with wave packets that move radially away from the target in the far field.

Finally, the second term contains a factor $1/r$. Therefore the magnitude of the second term decreases with the distance r from the target. This happens because the probability of finding a particle in a given detection area should decrease with distance. Indeed, the total detection area is $4\pi r^2$, where r is the distance at which the detectors are located. That increases proportional to r^2, and the total number of particles to detect per unit time is the same regardless of where the detectors are located. Therefore the probability of finding a particle per unit area should decrease proportional to $1/r^2$. Since the probability of finding a particle is proportional to the square of the wave function, the wave function itself must be proportional to $1/r$. The second term above makes it so.

Consider now the number of particles that is detected in a given small detection area ${\rm d}A$. The scattered stream of particles moving towards the detection area has a velocity $v = p_\infty/m$. Therefore in a time interval ${\rm d}t$, the detection area samples a volume of the scattered particle stream equal to ${\rm d}A \times v{\rm d}t$. The chances of finding particles are proportional to the square magnitude of the wave function times that volume. Using the asymptotic wave function above, the number of particles detected will be

$${\rm d}I = [\text{constant}]\ |C_{\rm f}(\theta,\phi)|^2 \frac{{\rm d}A}{r^2}{\rm d}t$$

The constant includes the factor p_∞/m. The constant must also account for the fact that the wave function is not normalized, and that there is a continuous stream of particles to be found, rather than just one particle.

According to the above expression, the number of particles detected in a given area ${\rm d}A$ is proportional to its three-dimensional angular extent

$${\rm d}\Omega \equiv \frac{{\rm d}A}{r^2}$$

This is the so-called "solid angle" occupied by the detection area element. It is the three-dimensional generalization of two-dimensional angles. In two dimensions, an element of a circle with arc length ${\rm d}s$ occupies an angle ${\rm d}s/r$ when

expressed in radians. Similarly, in three dimensions, an element of a sphere with area $\mathrm{d}A$ occupies a solid angle $\mathrm{d}A/r^2$ when expressed in "steradians."

In those terms, the number $\mathrm{d}I$ of particles detected in an infinitesimal solid angle $\mathrm{d}\Omega$ is

$$\mathrm{d}I = [\text{constant}]\ |C_\mathrm{f}(\theta,\phi)|^2 \mathrm{d}\Omega \mathrm{d}t$$

As noted, the constant of proportionality depends on the rate at which particles are sent at the target. The more particles are sent at the target, the more will be deflected. The number of particles in the incoming beam per unit beam cross-sectional area and per unit time is called the "luminosity" of the beam. It is related to the square of the wave function of the incoming beam through the relation

$$\mathrm{d}I_\mathrm{b} = [\text{constant}]\ |C_\mathrm{f}^\mathrm{I}|^2 \mathrm{d}A_\mathrm{b} \mathrm{d}t$$

Here $\mathrm{d}A_\mathrm{b}$ is a cross sectional area element of the incoming particle beam and $\mathrm{d}I_\mathrm{b}$ the number of particles passing through that area.

Physicist like to relate the scattered particle flow in a given infinitesimal solid angle $\mathrm{d}\Omega$ to an equivalent incoming beam area $\mathrm{d}A_b$ through which the same number of particles flow. Therefore they define the so-called "differential cross-section" as

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\omega} \equiv \frac{\mathrm{d}A_{\mathrm{b,equiv}}}{\mathrm{d}\Omega}} \tag{A.217}$$

The quantity $\mathrm{d}A_{\mathrm{b,equiv}}$ can be thought of as the infinitesimal area of the incoming beam that ends up in the infinitesimal solid angle $\mathrm{d}\Omega$. So the differential cross-section is a scattered particle density expressed in suitable terms.

Note how well chosen the term "differential cross-section" really is. If physicists had called it something like "scattered cross-section density," or even simply "scattered cross-section," nonexperts would probably have a pretty good guess what physicists were talking about. But "cross section" by itself can mean anything. There is nothing in the term to indicate that it is a measure for how many particles are scattered. And preceding it by "differential" is a stroke of genius because it is not a differential, it is a differential quotient. This will confuse mathematically literate nonexperts even more.

The differential cross section does not depend on how many particles are sent at the target, nor on wave function normalization. Following the expressions for the particle flows given above, the differential cross section is simply

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\omega} = \frac{|C_\mathrm{f}(\theta,\phi)|^2}{|C_\mathrm{f}^\mathrm{I}|^2}} \tag{A.218}$$

Moreover, the particle flow in an incoming beam area $\mathrm{d}A_\mathrm{b}$ may be measured using the same experimental techniques as are used to measure the deflected particle flow. Various systematic errors in the experimental method will then cancel in the ratio, giving more accurate values.

The total area of the incoming beam that gets scattered is called the "total cross-section" σ:

$$\boxed{\sigma \equiv A_{\text{b,total}}} \tag{A.219}$$

Of course, the name is quite below the normal standards of physicists, since it really is a total cross-section. Fortunately, physicist are clever enough not to say what cross section it is, and cross-section can mean many things. Also, by using the symbol σ instead of something logical like A_{b} for the differential cross-section, physicists do their best to reduce the damage as well as possible.

If you remain disappointed in physicists, take some comfort in the following term for scattering that can be described using classical mechanics: the "impact parameter." If you guess that it describes the local physics of the particle impact process, it is really hilarious to physicists. Instead, think "centerline offset;" it describes the location relative to the centerline of the incoming beam at which the particles come in; it has no direct relation whatsoever to what sort of impact (if any) these particles end up experiencing.

The total cross section can be found by integrating the differential cross section over all deflection angles:

$$\sigma = \int_{\text{all}} \frac{\mathrm{d}A_{\text{b,equiv}}}{\mathrm{d}\Omega}\,\mathrm{d}\Omega$$

In spherical coordinates this can be written out explicitly as

$$\boxed{\sigma = \int_{\theta=0^+}^{\pi} \int_{\phi=0}^{2\pi} \frac{|C_{\text{f}}(\theta,\phi)|^2}{|C_{\text{f}}^{\text{1}}|^2} \sin\theta\,\mathrm{d}\theta\mathrm{d}\phi} \tag{A.220}$$

A.30.1 Partial wave analysis

Jim Napolitano from RPI and Cornell notes:

The term "Partial Wave Analysis" is poorly defined and overused.

Gee, what a surprise! For one, they are component waves, not partial waves. But you already componently assumed that they might be.

This discussion will restrict itself to spherically symmetric scattering potentials. In that case, the analysis of the energy eigenfunctions can be done much like the analysis of the hydrogen atom of chapter 4.3. However, the boundary conditions at infinity will be quite different; the objective is not to describe bound particles, but particles that come in from infinity with positive kinetic energy and are scattered back to infinity. Also, the potential will of course not normally be a Coulomb one.

But just like for the hydrogen atom, the energy eigenfunctions can be taken to be radial functions times spherical harmonics Y_l^m:

$$\psi_{Elm}(r,\theta,\phi) = R_{El}(r)Y_l^m(\theta,\phi) \tag{A.221}$$

These energy eigenfunctions have definite angular momentum in the z-direction $m\hbar$, as well definite square angular momentum $l(l+1)\hbar^2$. The radial functions R_{El} will not be the same as the hydrogen R_{nl} ones.

The incoming plane wave $e^{\mathrm{i}p_\infty z/\hbar}$ has zero angular momentum in the z-direction. Unfortunately, it does *not* have definite square angular momentum. Instead, it can be written as a linear combination of free-space energy eigenfunctions with different values of l, hence with different square angular momentum:

$$e^{\mathrm{i}p_\infty z/\hbar} = \sum_{l=0}^{\infty} c_{\mathrm{w},l} j_l(p_\infty r/\hbar) Y_l^0(\theta) \qquad c_{\mathrm{w},l} = \mathrm{i}^l \sqrt{4\pi(2l+1)} \tag{A.222}$$

See {A.6} for a derivation and the precise form of the "spherical Bessel functions" j_l.

Now finding the complete energy eigenfunction corresponding to the incoming wave directly is typically awkward, especially analytically. Often it is easier to solve the problem for each term in the above sum separately and then add these solutions all together. That is where the name "partial wave analysis" comes from. Each term in the sum corresponds to a partial wave, if you use sufficiently lousy terminology.

The partial wave analysis requires that for each term in the sum, an energy eigenfunction is found of the form $\psi_{El} = R_{El}Y_l^0$. The required behavior of this eigenfunction in the far field is

$$\boxed{\psi_{El} \sim \left[c_{\mathrm{w},l} j_l(p_\infty r/\hbar) + c_{\mathrm{f},l} h_l^{(1)}(p_\infty r/\hbar)\right] Y_l^0(\theta) \qquad \text{for } r \to \infty} \tag{A.223}$$

Here the first term is the component of the incoming plane wave corresponding to spherical harmonic Y_l^0. The second term represents the outgoing deflected particles. The value of the coefficient $c_{\mathrm{f},l}$ is determined in the solution process.

Note that the above far field behavior is quite similar to that of the complete energy eigenfunction as given earlier in (A.216). However, here the coefficient $C_{\mathrm{f}}^{\mathrm{l}}$ was set to 1 for simplicity. Also, the radial part of the reflected wave function was written using a "Hankel function of the first kind" $h_l^{(1)}$. This Hankel function produces the same $e^{\mathrm{i}p_\infty r/\hbar}/r$ radial behavior as the second term in (A.216), {A.6} (A.25). However, the Hankel function has the advantage that it becomes exact as soon as the scattering potential becomes zero. It is not just valid at very large r like the bare exponential.

To be sure, for a slowly decaying potential like the Coulomb one, the Hankel function is no better than the exponential. However, the Hankel function is very

closely related to the Bessel function j_l, {A.6}, allowing various helpful results to be found in table books. If the potential energy is piecewise constant, it is even possible to solve the complete problem using Bessel and Hankel functions. These functions can be tied together at the jumps in potential in a way similar to addendum {A.27}.

In terms of the asymptotic behavior above, the differential cross section is

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\omega} = \frac{\hbar^2}{p_\infty^2} \sum_{l=0}^{\infty} \sum_{\underline{l}=0}^{\infty} \mathrm{i}^{l-\underline{l}} c_{\mathrm{f},l}^* c_{\mathrm{f},\underline{l}} Y_l^0(\theta) Y_{\underline{l}}^0(\theta)} \tag{A.224}$$

This can be verified using {A.6} (A.25), (A.216), and (A.218). The Bessel functions form the incoming wave and do not contribute. For the total cross-section, note that the spherical harmonics are orthonormal, so

$$\boxed{\sigma = \frac{\hbar^2}{p_\infty^2} \sum_{l=0}^{\infty} |c_{\mathrm{f},l}|^2}$$

One special case is worth mentioning. Consider particles of such low momentum that their typical quantum wave length, $2\pi\hbar/p_\infty$, is gigantic compared to the radial size of the scattering potential. Particles of such large wave lengths do not notice the fine details of the scattering potential at all. Conversely, normally the scattering potential only "notices" the incoming partial wave with $l = 0$. That is because the Bessel functions are proportional to

$$(p_\infty r/\hbar)^l$$

for small arguments. If the wave length is large compared to the typical radial size r of the scattering potential, this is negligible unless $l = 0$. Now $l = 0$ corresponds to a wave function that is the same in all directions; it is proportional to the constant spherical harmonic Y_0^0. If only the partial wave that is the same in all directions gets scattered, then the particles get scattered equally in all directions (if they get scattered at all.)

Coincidently, equal scattering in all directions also happens in another case: scattering of classical point particles from a hard elastic sphere. That is very much the opposite case, because negligible uncertainty in position requires high, not low, energy of the particles. In any case, the similarity between the two cases is is superficial. If a beam of classical particles is directed at a hard sphere, only an area of the beam equal to the frontal area of the sphere gets scattered. But if you work out the scattering of low-energy quantum particles from a hard sphere, you get a total scattering cross section that is 4 times bigger.

A.30.2 Partial wave amplitude

This subsection gives some further odds and ends on partial wave analysis, for the incurably curious.

Recall that a partial wave has an asymptotic behavior

$$\psi_{El} \sim \left[c_{\mathrm{w},l} j_l(p_\infty r/\hbar) + c_{\mathrm{f},l} h_l^{(1)}(p_\infty r/\hbar) \right] Y_l^0(\theta) \qquad \text{for } r \to \infty$$

The first term corresponds to the wave function of the incoming particles. The second term is the effect of the scattering potential.

Physicists like to write the coefficient of the scattered wave as

$$\boxed{c_{\mathrm{f},l} = \mathrm{i}k c_{\mathrm{w},l} a_l} \tag{A.225}$$

They call the so-defined constant a_l the "partial wave amplitude" because obviously it is not a partial wave amplitude. Confusing people is always funny.

Now every partial wave by itself is a solution to the Hamiltonian eigenvalue problem. That means that every partial wave must ensure that particles cannot just simply disappear. That restricts what the partial wave amplitude can be. It turns out that it can be written in terms of a real number δ_l:

$$\boxed{a_l = \frac{1}{k} e^{\mathrm{i}\delta_l} \sin\delta_l} \tag{A.226}$$

The real number δ_l is called the "phase shift."

Some physicist must have got confused here, because it really is a phase shift. To see that, consider the derivation of the above result. First the asymptotic behavior of the partial wave is rewritten in terms of exponentials using {A.6} (A.24) and (A.25). That gives

$$\psi_{El} \sim \ldots \left[e^{-\mathrm{i}p_\infty r/\hbar} + (-1)^{l+1} e^{\mathrm{i}p_\infty r/\hbar} (1 + 2\mathrm{i}k a_l) \right]$$

The dots stand for common factors that are not important for the discussion. Physically, the first term above describes spherical wave packets that move radially inwards toward the target. The second term describes wave packets that move radially outwards away from the target.

Now particles cannot just disappear. Wave packets that go in towards the target must come out again with the same amplitude. And that means that the two terms in the asymptotic behavior above must have the same magnitude. (This may also be shown mathematically using procedures like in {A.32}.)

Obviously the two terms do have the same magnitude in the absence of scattering, where a_l is zero. But in the presence of scattering, the final parenthetical factor will have to stay of magnitude one. And that means that it can be written in the form

$$1 + 2\mathrm{i}k a_l = e^{\mathrm{i}2\delta_l} \tag{A.227}$$

for some real number δ_l. (The factor 2 in the exponential is put in because physicists like to think of the wave being phase shifted twice, once on the way

in to the target and once on the way out.) Cleaning up the above expression using the Euler formula (2.5) gives the stated result.

If you add in the time dependent factor $e^{\mathrm{i}Et/\hbar}$ of the complete unsteady wave function, you can see that indeed the waves are shifted by a phase angle $2\delta_l$ compared to the unperturbed wave function. Without any doubt, the name of the physicist responsible for calling the phase angle a "phase angle" has been ostracized from physics. She will never be heard of again.

A.30.3 The Born approximation

The Born approximation assumes that the scattering potential is weak to derive approximate expressions for the scattering.

Consider first the case that the scattering potential is zero. In that case, the wave function is just that of the incoming particles:

$$\psi_E = e^{\mathrm{i}p_\infty z/\hbar} \qquad p_\infty = \sqrt{2mE}$$

where E is the energy of the particle and m its mass.

Born considered the case that the scattering potential V is not zero, but small. Then the wave function ψ_E will still be close to the incoming wave function, but no longer exactly the same. In that case an approximation to the wave function can be obtained from the so-called integral Schrödinger equation, {A.13} (A.42):

$$\psi_E(\vec{r}) = e^{\mathrm{i}p_\infty z/\hbar} - \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}\,'} \frac{e^{\mathrm{i}p_\infty |\vec{r}-\vec{r}\,'|/\hbar}}{|\vec{r}-\vec{r}\,'|} V(\vec{r}\,')\psi_E(\vec{r}\,')\,\mathrm{d}^3\vec{r}\,'$$

In particular, inside the integral the true wave function ψ_E can be replaced by the incoming wave function:

$$\psi_E(\vec{r}) \approx e^{\mathrm{i}p_\infty z/\hbar} - \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}\,'} \frac{e^{\mathrm{i}p_\infty |\vec{r}-\vec{r}\,'|/\hbar}}{|\vec{r}-\vec{r}\,'|} V(\vec{r}\,')e^{\mathrm{i}p_\infty z'/\hbar}\,\mathrm{d}^3\vec{r}\,' \tag{A.228}$$

It is not exact, but it is much better than just setting the integral to zero. The latter would make the wave function equal to the incoming wave. With the approximate integral, you get a valid approximation to the particle deflections.

To get the differential cross section, examine the behavior of (A.228) at given scattering angles θ and ϕ for large r. That produces, {D.47}:

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\omega}(\theta,\phi) \approx \left| \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}\,'} e^{-\mathrm{i}(\vec{p}_\infty - \vec{p}_\infty^{\,\mathrm{l}})\cdot\vec{r}\,'/\hbar}\, V(\vec{r}\,')\,\mathrm{d}^3\vec{r}\,' \right|^2 \qquad (V \text{ small})} \tag{A.229}$$

Here

$$\vec{p}_\infty^{\,\mathrm{l}} = p_\infty \hat{k} \qquad \vec{p}_\infty = p_\infty \hat{\imath}_r \qquad (\text{with } p_\infty = \sqrt{2mE}) \tag{A.230}$$

are the classical momentum *vectors* of the incoming and scattered particles. Note that the direction of $\vec{p}_\infty$ depends on the considered scattering angles θ and ϕ. And that apparently the momentum change of the particles is a key factor affecting the amount of scattering.

One additional approximation is worth mentioning. Consider particles of such low momentum that their quantum wave length, $2\pi\hbar/p_\infty$, is gigantic compared to the radial size of the scattering potential. Particles of such wave lengths do not notice the fine details of the scattering potential at all. Mathematically, p_∞ is so small that the argument of the exponential in the differential cross section above can be assumed zero. Then:

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\omega} \approx \left| \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}\,'} V(\vec{r}\,') \, \mathrm{d}^3\vec{r}\,' \right|^2 \qquad (V \text{ and } p_\infty \text{ small})} \tag{A.231}$$

The differential cross section no longer depends on the angular position. If particles get scattered at all, they get scattered equally in all directions.

Note that the integral is infinite for a Coulomb potential.

A.31 The Born series

The Born approximation is concerned with the problem of a particle of a given momentum that is slightly perturbed by a nonzero potential that it encounters. This note gives a description how this problem may be solved to high accuracy. The solution provides a model for the so-called "Feynman diagrams" of quantum electrodynamics.

It is assumed that in the absence of the perturbation, the wave function of the particle would be

$$\psi_0 = e^{\mathrm{i}kz}$$

In this state, the particle has a momentum $\hbar k$ that is purely in the z-direction. Note that the above state is not normalized, and cannot be. That reflects the Heisenberg uncertainty principle: since the particle has precise momentum, it has infinite uncertainty in position. For real particles, wave functions of the form above must be combined into wave packets, chapter 7.10. That is not important for the current discussion.

The *perturbed* wave function ψ can in principle be obtained from the so-called integral Schrödinger equation, {A.13} (A.42):

$$\psi(\vec{r}) = \psi_0(\vec{r}) - \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}\,'} \frac{e^{\mathrm{i}k|\vec{r}-\vec{r}\,'|}}{|\vec{r}-\vec{r}\,'|} V(\vec{r}\,')\psi(\vec{r}\,') \, \mathrm{d}^3\vec{r}\,'$$

Evaluating the right hand side in this equation would give ψ. Unfortunately, the right hand side cannot be evaluated because the integral contains the unknown wave function ψ still to be found. However, Born noted that if the perturbation

is small, then so is the difference between the true wave function ψ and the unperturbed one ψ_0. So a valid approximation to the integral can be obtained by replacing ψ in it by the known ψ_0. That is certainly much better than just leaving the integral away completely, which would give $\psi = \psi_0$.

And note that you can repeat the process. Since you now have an approximation for ψ that is better than ψ_0, you can put that approximation into the integral instead. Evaluating the right hand side then produces a still better approximation for ψ. Which can then be put into the integral. Etcetera.

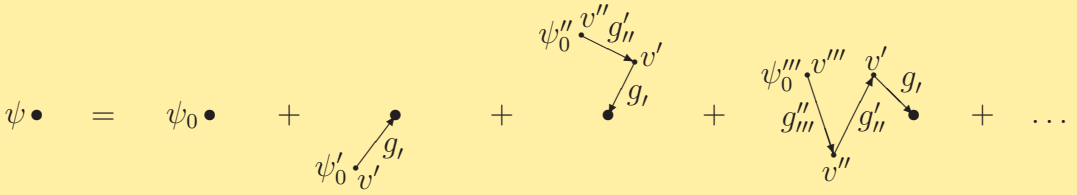


Figure A.22: Graphical interpretation of the Born series.

Graphically, the process is illustrated in figure A.22. The most inaccurate approximation is to take the perturbed wave function as the unperturbed wave function at the same position $\vec{r}$:

$$\psi \approx \psi_0$$

An improvement is to add the integral evaluated using the unperturbed wave function:

$$\psi(\vec{r}) = \psi_0(\vec{r}) - \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}\,'} \frac{e^{\mathrm{i}k|\vec{r}-\vec{r}\,'|}}{|\vec{r}-\vec{r}\,'|} V(\vec{r}\,')\psi_0(\vec{r}\,')\,\mathrm{d}^3\vec{r}\,'$$

To represent this concisely, it is convenient to introduce some shorthand notations:

$$\psi_0' \equiv \psi_0(\vec{r}\,') \qquad v' \equiv V(\vec{r}\,')\,\mathrm{d}^3\vec{r}\,' \qquad g_\prime = -\frac{m}{2\pi\hbar^2}\frac{e^{\mathrm{i}k|\vec{r}-\vec{r}\,'|}}{|\vec{r}-\vec{r}\,'|}$$

Using those notations the improved approximation to the wave function is

$$\psi \approx \psi_0 + \int \psi_0' v' g_\prime$$

Note what the second term does: it takes the unperturbed wave function at some different location $\vec{r}\,'$, multiplies it by a "vertex factor" v', and then adds it to the wave function at $\vec{r}$ multiplied by a "propagator" $g_\prime$. This is then summed over all locations $\vec{r}\,'$. The second term is illustrated in the second graph in the right hand side of figure A.22.

The next better approximation is obtained by putting the two-term approximation above in the integral:

$$\psi \approx \psi_0 + \int \left[\psi_0' + \int \psi_0'' v'' g_{\prime\prime}'\right] v' g_\prime$$

where

$$\psi_0'' \equiv \psi_0(\vec{r}\,'') \qquad v'' \equiv V(\vec{r}\,'')\,\mathrm{d}^3\vec{r}\,'' \qquad g_{\prime\prime}' = -\frac{m}{2\pi\hbar^2}\frac{e^{\mathrm{i}k|\vec{r}\,'-\vec{r}\,''|}}{|\vec{r}\,'-\vec{r}\,''|}$$

Note that it was necessary to change the notation for one integration variable to $\vec{r}\,''$ to avoid using the same symbol for two different things. Compared to the previous approximation, there is now a third term:

$$\psi = \psi_0 + \int \psi_0' v' g_{\prime} + \iint \psi_0'' v'' g_{\prime\prime}'\, v' g_{\prime}$$

This third term takes the unperturbed wave function at some position $\vec{r}\,''$, multiplies it by the local vertex factor v'', propagates that to a location $\vec{r}\,'$ using propagator $g_{\prime\prime}'$, multiplies it by the vertex factor v', and propagates it to the location $\vec{r}$ using propagator $g_{\prime}$. That is summed over all combinations of locations $\vec{r}\,''$ and $\vec{r}\,'$. The idea is shown in the third graph in the right hand side of figure A.22.

Continuing this process produces the Born series:

$$\psi = \psi_0 + \int \psi_0' v' g_{\prime} + \iint \psi_0'' v'' g_{\prime\prime}'\, v' g_{\prime} + \iiint \psi_0''' v''' g_{\prime\prime\prime}''\, v'' g_{\prime\prime}'\, v' g_{\prime} + \ldots$$

The Born series inspired Feynman to formulate relativistic quantum mechanics in terms of vertices connected together into "Feynman diagrams." Since there is a nontechnical, very readable discussion available from the master himself, [19], there does not seem much need to go into the details here.

A.32 The evolution of probability

This note looks at conservation of probability, and the resulting definitions of the reflection and transmission coefficients in scattering. It also explains the concept of the "probability current" that you may occasionally run into.

For the unsteady Schrödinger equation to provide a physically correct description of nonrelativistic quantum mechanics, particles should not be able to disappear into thin air. In particular, during the evolution of the wave function of a single particle, the total probability of finding the particle if you look everywhere should stay one at all times:

$$\int_{x=-\infty}^{\infty} |\Psi|^2\,\mathrm{d}x = 1 \text{ at all times}$$

Fortunately, the Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V\Psi$$

does indeed conserve this total probability, so all is well.

To verify this, note first that $|\Psi|^2 = \Psi^*\Psi$, where the star indicates the complex conjugate, so

$$\frac{\partial|\Psi|^2}{\partial t} = \Psi^*\frac{\partial\Psi}{\partial t} + \Psi\frac{\partial\Psi^*}{\partial t}$$

To get an expression for that, take the Schrödinger equation above times $\Psi^*/\mathrm{i}\hbar$ and add the complex conjugate of the Schrödinger equation,

$$-\mathrm{i}\hbar\frac{\partial\Psi^*}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi^*}{\partial x^2} + V\Psi^*,$$

times $-\Psi/\mathrm{i}\hbar$. The potential energy terms drop out, and what is left is

$$\frac{\partial|\Psi|^2}{\partial t} = \frac{\mathrm{i}\hbar}{2m}\left(\Psi^*\frac{\partial^2\Psi}{\partial x^2} - \Psi\frac{\partial^2\Psi^*}{\partial x^2}\right).$$

Now it can be verified by differentiating out that the right hand side can be rewritten as a derivative:

$$\frac{\partial|\Psi|^2}{\partial t} = -\frac{\partial J}{\partial x} \qquad \text{where } J = \frac{\mathrm{i}\hbar}{2m}\left(\Psi\frac{\partial\Psi^*}{\partial x} - \Psi^*\frac{\partial\Psi}{\partial x}\right) \tag{A.232}$$

For reasons that will become evident below, J is called the "probability current." Note that J, like Ψ, will be zero at infinite x for proper, normalized wave functions.

If (A.232) is integrated over all x, the desired result is obtained:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x=-\infty}^{\infty}|\Psi|^2\,\mathrm{d}x = -J\Big|_{x=-\infty}^{\infty} = 0.$$

Therefore, the total probability of finding the particle does not change with time. If a proper initial condition is provided to the Schrödinger equation in which the total probability of finding the particle is one, then it stays one for all time.

It gets a little more interesting to see what happens to the probability of finding the particle in some given finite region $a \leqslant x \leqslant b$. That probability is given by

$$\int_{x=a}^{b}|\Psi|^2\,\mathrm{d}x$$

and it can change with time. A wave packet might enter or leave the region. In particular, integration of (A.232) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{x=a}^{b}|\Psi|^2\,\mathrm{d}x = J_a - J_b$$

This can be understood as follows: J_a is the probability flowing out of the region $x < a$ into the interval $[a, b]$ through the end a. That increases the probability

within $[a, b]$. Similarly, J_b is the probability flowing out of $[a, b]$ at b into the region $x > b$; it decreases the probability within $[a, b]$. Now you see why J is called probability current; it is equivalent to a stream of probability in the positive x-direction.

The probability current can be generalized to more dimensions using vector calculus:

$$\vec{J} = \frac{\mathrm{i}\hbar}{2m} (\Psi \nabla \Psi^* - \Psi^* \nabla \Psi) \tag{A.233}$$

and the net probability flowing out of a region is given by

$$\int \vec{J} \cdot \vec{n} \, \mathrm{d}A \tag{A.234}$$

where A is the outside surface area of the region, and $\vec{n}$ is a unit vector normal to the surface. A surface integral like this can often be simplified using the divergence (Gauss or whatever) theorem of calculus.

Returning to the one-dimensional case, it is often desirable to relate conservation of probability to the energy eigenfunctions of the Hamiltonian,

$$-\frac{\hbar^2}{2m} \frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V\psi = E\psi$$

because the energy eigenfunctions are generic, not specific to one particular example wave function Ψ.

To do so, first an important quantity called the "Wronskian" must be introduced. Consider any two eigenfunctions ψ_1 and ψ_2 of the Hamiltonian:

$$-\frac{\hbar^2}{2m} \frac{\mathrm{d}^2\psi_1}{\mathrm{d}x^2} + V\psi_1 = E\psi_1$$

$$-\frac{\hbar^2}{2m} \frac{\mathrm{d}^2\psi_2}{\mathrm{d}x^2} + V\psi_2 = E\psi_2$$

If you multiply the first equation above by ψ_2, the second by ψ_1 and then subtract the two, you get

$$\frac{\hbar^2}{2m} \left(\psi_1 \frac{\mathrm{d}^2\psi_2}{\mathrm{d}x^2} - \psi_2 \frac{\mathrm{d}^2\psi_1}{\mathrm{d}x^2} \right) = 0$$

The constant $\hbar^2/2m$ can be divided out, and by differentiation it can be verified that the remainder can be written as

$$\frac{\mathrm{d}W}{\mathrm{d}x} = 0 \qquad \text{where } W = \psi_1 \frac{\mathrm{d}\psi_2}{\mathrm{d}x} - \psi_2 \frac{\mathrm{d}\psi_1}{\mathrm{d}x}$$

The quantity W is called the Wronskian. It is the same at all values of x.

As an application, consider the example potential of figure A.11 in addendum {A.27} that bounces a particle coming in from the far left back to where it came

from. In the left region, the potential V has a constant value V_l. In this region, an energy eigenfunction is of the form

$$\psi_E = C_\mathrm{f}^\mathrm{l} e^{\mathrm{i}p_\mathrm{c}^\mathrm{l} x/\hbar} + C_\mathrm{b}^\mathrm{l} e^{-\mathrm{i}p_\mathrm{c}^\mathrm{l} x/\hbar} \text{ for } x < x_\mathrm{A} \qquad \text{where } p_\mathrm{c}^\mathrm{l} = \sqrt{2m(E-V_\mathrm{l})}$$

At the far right, the potential grows without bound and the eigenfunction becomes zero rapidly. To make use of the Wronskian, take the first solution ψ_1 to be ψ_E itself, and ψ_2 to be its complex conjugate ψ_E^*. Since at the far right the eigenfunction becomes zero rapidly, the Wronskian is zero there. And since the Wronskian is constant, that means it must be zero everywhere. Next, if you plug the above expression for the eigenfunction in the left region into the definition of the Wronskian and clean up, you get

$$W = \frac{2\mathrm{i}p_\mathrm{c}^\mathrm{l}}{\hbar}\left(|C_\mathrm{b}^\mathrm{l}|^2 - |C_\mathrm{f}^\mathrm{l}|^2\right).$$

If that is zero, the magnitude of C_b^l must be the same as that of C_f^l.

This can be understood as follows: if a wave packet is created from eigenfunctions with approximately the same energy, then the terms $C_\mathrm{f}^\mathrm{l} e^{\mathrm{i}p_\mathrm{c}^\mathrm{l} x/\hbar}$ combine for large negative times into a wave packet coming in from the far left. The probability of finding the particle in that wave packet is proportional to the integrated square magnitude of the wave function, hence proportional to the square magnitude of C_f^l. For large positive times, the $C_\mathrm{b}^\mathrm{l} e^{-\mathrm{i}p_\mathrm{c}^\mathrm{l} x/\hbar}$ terms combine in a similar wave packet, but one that returns towards the far left. The probability of finding the particle in that departing wave packet must still be the same as that for the incoming packet, so the square magnitude of C_b^l must be the same as that of C_f^l.

Next consider a generic scattering potential like the one in figure 7.22. To the far left, the eigenfunction is again of the form

$$\psi_E = C_\mathrm{f}^\mathrm{l} e^{\mathrm{i}p_\mathrm{c}^\mathrm{l} x/\hbar} + C_\mathrm{b}^\mathrm{l} e^{-\mathrm{i}p_\mathrm{c}^\mathrm{l} x/\hbar} \text{ for } x << 0 \qquad \text{where } p_\mathrm{c}^\mathrm{l} = \sqrt{2m(E-V_\mathrm{l})}$$

while at the far right it is now of the form

$$\psi_E = C^\mathrm{r} e^{\mathrm{i}p_\mathrm{c}^\mathrm{r} x/\hbar} \text{ for } x >> 0 \qquad \text{where } p_\mathrm{c}^\mathrm{r} = \sqrt{2m(E-V_\mathrm{r})}$$

The Wronskian can be found the same way as before:

$$W = \frac{2\mathrm{i}p_\mathrm{c}^\mathrm{l}}{\hbar}\left(|C_\mathrm{b}^\mathrm{l}|^2 - |C_\mathrm{f}^\mathrm{l}|^2\right) = -\frac{2\mathrm{i}p_\mathrm{c}^\mathrm{r}}{\hbar}|C^\mathrm{r}|^2$$

The fraction of the incoming wave packet that ends up being reflected back towards the far left is called the "reflection coefficient" R. Following the same reasoning as above, it can be computed from the coefficients in the far left region of constant potential as:

$$R = \frac{|C_\mathrm{b}^\mathrm{l}|^2}{|C_\mathrm{f}^\mathrm{l}|^2}$$

The reflection coefficient gives the probability that the particle can be found to the left of the scattering region at large times.

Similarly, the fraction of the incoming wave packet that passes through the potential barrier towards the far right is called the "transmission coefficient" T. It gives the probability that the particle can be found to the right of the scattering region at large times. Because of conservation of probability, $T = 1 - R$.

Alternatively, because of the Wronskian expression above, the transmission coefficient can be explicitly computed from the coefficient of the eigenfunction in the far right region as

$$T = \frac{p_{\rm c}^{\rm r} |C^{\rm r}|^2}{p_{\rm c}^{\rm l} |C_{\rm f}^{\rm l}|^2} \qquad p_{\rm c}^{\rm l} = \sqrt{2m(E - V_{\rm l})} \quad p_{\rm c}^{\rm r} = \sqrt{2m(E - V_{\rm r})}$$

If the potential energy is the same at the far right and far left, the two classical momenta are the same, $p_{\rm c}^{\rm r} = p_{\rm c}^{\rm l}$. Otherwise, the reason that the ratio of classical momenta appears in the transmission coefficient is because the classical momenta in a wave packet have a different spacing with respect to energy if the potential energy is different. (The above expression for the transmission coefficient can also be derived explicitly using the Parseval equality of Fourier analysis, instead of inferred from conservation of probability and the constant Wronskian.)

A.33 Explanation of the London forces

To fully understand the details of the London forces, it helps to first understand the popular explanation of them, and why it is all wrong. To keep things simple, the example will be the London attraction between two neutral hydrogen atoms that are well apart. (This will also correct a small error that the earlier discussion of the hydrogen molecule made; that discussion implied incorrectly that there is no attraction between two neutral hydrogen atoms that are far apart. The truth is that there really is some Van der Waals attraction. It was ignored because it is small compared to the chemical bond that forms when the atoms are closer together and would distract from the real story.)

The popular explanation for the London force goes something like this: "Sure, there would not be any attraction between two distant hydrogen atoms if they were perfectly spherically symmetric. But according to quantum mechanics, nature is uncertain. So sometimes the electron clouds of the two atoms are somewhat to the left of the nuclei, like in figure A.23 (*b*). This polarization [dipole creation] of the atoms turns out to produce some electrostatic attraction between the atoms. At other times, the electron clouds are somewhat to the right of the nuclei like in figure A.23 (*c*); it is really the same thing seen in the mirror. In cases like figure A.23 (*a*), where the electron clouds move towards

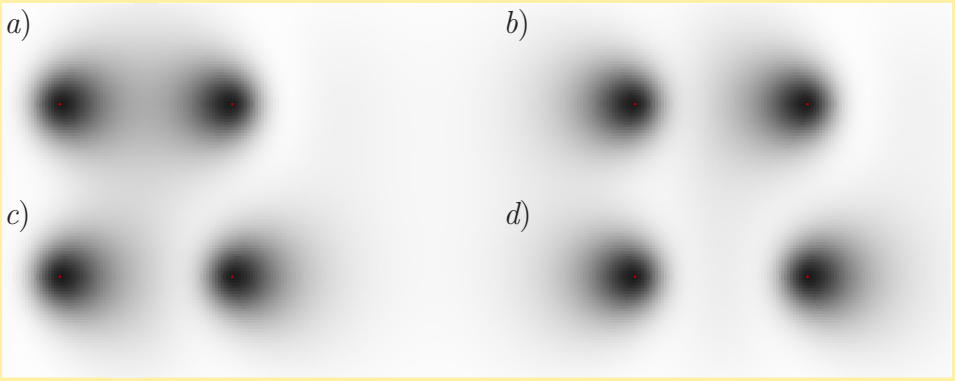


Figure A.23: Possible polarizations of a pair of hydrogen atoms.

each other, and (*b*), where they move away from each other, there is some repulsion between the atoms; however, the wave functions become correlated so that (*b*) and (*c*) are more likely than (*a*) and (*d*). Hence a net attraction results."

Before examining what is wrong with this explanation, first consider what is right. It is perfectly right that figure A.23 (*b*) and (*c*) produce some net attraction between the atoms, and that (*a*) and (*d*) produce some repulsion. This follows from the net Coulomb potential energy between the atoms for given positions of the electrons:

$$V_{\rm lr} = \frac{e^2}{4\pi\epsilon_0}\left(\frac{1}{d} - \frac{1}{r_{\rm l}} - \frac{1}{r_{\rm r}} + \frac{1}{r_{\rm lr}}\right)$$

where $e = 1.6\ 10^{-19}$ C is the magnitude of the charges of the protons and electrons, $\epsilon_0 = 8.85\ 10^{-12}$ C^2/J m is the permittivity of space, d is the distance between the nuclei, $r_{\rm l}$ is the distance between the left electron and the right nucleus, $r_{\rm r}$ the one between the right electron and the left nucleus, and $r_{\rm lr}$ is the distance between the two electrons. If the electrons charges are distributed over space according to densities $n_{\rm l}(\vec{r}_{\rm l})$ and $n_{\rm r}(\vec{r}_{\rm r})$, the classical potential energy is

$$V_{\rm lr} = \frac{e^2}{4\pi\epsilon_0}\int_{\text{all }\vec{r}_{\rm l}}\int_{\text{all }\vec{r}_{\rm r}}\left(\frac{1}{d} - \frac{1}{r_{\rm l}} - \frac{1}{r_{\rm r}} + \frac{1}{r_{\rm lr}}\right) n_{\rm l}(\vec{r}_{\rm l}) n_{\rm r}(\vec{r}_{\rm r})\, {\rm d}^3\vec{r}_{\rm l} {\rm d}^3\vec{r}_{\rm r}$$

(Since the first, $1/d$, term represents the repulsion between the nuclei, it may seem strange to integrate it against the *electron* charge distributions, but the charge distributions integrate to one, so they disappear. Similarly in the second and third term, the charge distribution of the uninvolved electron integrates away.)

Since it is assumed that the atoms are well apart, the integrand above can be simplified using Taylor series expansions to give:

$$V_{\rm lr} = \frac{e^2}{4\pi\epsilon_0}\int_{\text{all }\vec{r}_{\rm l}}\int_{\text{all }\vec{r}_{\rm r}}\frac{x_{\rm l}x_{\rm r} + y_{\rm l}y_{\rm r} - 2z_{\rm l}z_{\rm r}}{d^3} n_{\rm l}(\vec{r}_{\rm l}) n_{\rm r}(\vec{r}_{\rm r})\, {\rm d}^3\vec{r}_{\rm l} {\rm d}^3\vec{r}_{\rm r}$$

where the positions of the electrons are measured from their respective nuclei. Also, the two z axes are both taken horizontal and positive towards the left. For charge distributions as shown in figure A.23, the $x_\mathrm{l}x_\mathrm{r}$ and $y_\mathrm{l}y_\mathrm{r}$ terms integrate to zero because of odd symmetry. However, for a distribution like in figure A.23 (c), n_l and n_r are larger at positive z_l, respectively z_r, than at negative one, so the integral will integrate to a negative number. That means that the potential is lowered, there is attraction between the atoms. In a similar way, distribution (b) produces attraction, while (a) and (d) produce repulsion.

So there is nothing wrong with the claim that (b) and (c) produce attraction, while (a) and (d) produce repulsion. It is also perfectly right that the combined quantum wave function gives a higher probability to (b) and (c) than to (a) and (d).

So what is wrong? There are two major problems with the story.

1. *Energy eigenstates are stationary.* If the wave function oscillated in time like the story suggests, it would require uncertainty in energy, which would act to kill off the lowering of energy. True, states with the electrons at the same side of their nuclei are more likely to show up when you measure them, but to reap the benefits of this increased probability, you must *not* do such a measurement and just let the electron wave function sit there unchanging in time.
2. *The numbers are all wrong.* Suppose the wave functions in figures (b) and (c) shift (polarize) by a typical small amount ε. Then the attractive potential is of order ε^2/d^3. Since the distance d between the atoms is assumed large, the energy gained is a small amount times ε^2. But to shift atom energy eigenfunctions by an amount ε away from their ground state takes an amount of energy $C\varepsilon^2$ where C is some constant that is *not* small. So it would take more energy to shift the electron clouds than the dipole attraction could recover. In the ground state, the electron clouds should therefore stick to their original centered positions.

On to the correct quantum explanation. First the wave function is needed. If there were no Coulomb potentials linking the atoms, the combined ground-state electron wave function would simply take the form

$$\psi(\vec{r}_\mathrm{l},\vec{r}_\mathrm{r}) = \psi_{100}(\vec{r}_\mathrm{l})\psi_{100}(\vec{r}_\mathrm{r})$$

where ψ_{100} is the ground state wave function of a single hydrogen atom. To get a suitable correlated polarization of the atoms, throw in a bit of the ψ_{210} "$2\mathrm{p}_z$" states, as follows:

$$\psi(\vec{r}_\mathrm{l},\vec{r}_\mathrm{r}) = \sqrt{1-\varepsilon^2}\psi_{100}(\vec{r}_\mathrm{l})\psi_{100}(\vec{r}_\mathrm{r}) + \varepsilon\psi_{210}(\vec{r}_\mathrm{l})\psi_{210}(\vec{r}_\mathrm{r}).$$

For $\varepsilon > 0$, it produces the desired correlation between the wave functions: ψ_{100} is always positive, and ψ_{210} is positive if the electron is at the positive-z side of

its nucleus and negative otherwise. So if both electrons are at the same side of their nucleus, the product $\psi_{210}(\vec{r}_\mathrm{l})\psi_{210}(\vec{r}_\mathrm{r})$ is positive, and the wave function is increased, giving increased probability of such states. Conversely, if the electrons are at opposite sides of their nucleus, $\psi_{210}(\vec{r}_\mathrm{l})\psi_{210}(\vec{r}_\mathrm{r})$ is negative, and the wave function is reduced.

Now write the expectation value of the energy:

$$\langle E\rangle = \langle\sqrt{1-\varepsilon^2}\psi_{100}\psi_{100} + \varepsilon\psi_{210}\psi_{210}|H_\mathrm{l} + H_\mathrm{r} + V_\mathrm{lr}|\sqrt{1-\varepsilon^2}\psi_{100}\psi_{100} + \varepsilon\psi_{210}\psi_{210}\rangle$$

where H_l and H_r are the Hamiltonians of the individual electrons and

$$V_\mathrm{lr} = \frac{e^2}{4\pi\epsilon_0}\frac{x_\mathrm{l}x_\mathrm{r} + y_\mathrm{l}y_\mathrm{r} - 2z_\mathrm{l}z_\mathrm{r}}{d^3}$$

is again the potential between atoms. Working out the inner product, noting that the ψ_{100} and ψ_{210} are orthonormal eigenfunctions of the atom Hamiltonians H_l and H_r with eigenvalues E_1 and E_2, and that most V_lr integrals are zero on account of odd symmetry, you get

$$\langle E\rangle = 2E_1 + 2\varepsilon^2(E_2 - E_1) - 4\varepsilon\sqrt{1-\varepsilon^2}\frac{e^2}{4\pi\epsilon_0}\frac{1}{d^3}\langle\psi_{100}\psi_{100}|z_\mathrm{l}z_\mathrm{r}|\psi_{210}\psi_{210}\rangle.$$

The final term is the savior for deriving the London force. For small values of ε, for which the square root can be approximated as one, this energy-lowering term dominates the energy $2\varepsilon^2(E_2 - E_1)$ needed to distort the atom wave functions. The best approximation to the true ground state is then obtained when the quadratic in ε is minimal. That happens when the energy has been lowered by an amount

$$\frac{2}{E_2 - E_1}\left(\frac{e^2}{4\pi\epsilon_0}\langle\psi_{100}|z|\psi_{210}\rangle^2\right)^2\frac{1}{d^6}.$$

Since the assumed eigenfunction is not exact, this variational approximation will underestimate the actual London force. For example, it can be seen that the energy can also be lowered similar amounts by adding some of the $2\mathrm{p}_x$ and $2\mathrm{p}_y$ states; these cause the atom wave functions to move in opposite directions normal to the line between the nuclei.

So what is the physical meaning of the savior term? Consider the inner product that it represents:

$$\langle\psi_{100}\psi_{100}|V_\mathrm{lr}|\psi_{210}\psi_{210}\rangle.$$

That is the energy if both electrons are in the spherically symmetric ψ_{100} ground state if both electrons are in the antisymmetric $2\mathrm{p}_z$ state. The savior term is a twilight term, like the ones discussed earlier in chapter 5.3 for chemical bonds. It reflects nature's habit of doing business in terms of an unobservable wave function instead of observable probabilities.

A.34 Explanation of Hund's first rule

Hund's first rule of spin-alignment applies because electrons in atoms prefer to go into spatial states that are antisymmetric with respect to electron exchange. Spin alignment is then an unavoidable consequence of the weird antisymmetrization requirement.

To understand why electrons want to go into antisymmetric spatial states, the *interactions* between the electrons need to be considered. Sweeping them below the carpet as the discussion of atoms in chapter 5.9 did is not going to cut it.

To keep it as simple as possible, the case of the carbon atom will be considered. As the crude model of chapter 5.9 did correctly deduce, the carbon atom has two 1s electrons locked into a zero-spin singlet state, and similarly two 2s electrons also in a singlet state. Hund's rule is about the final two electrons that are in 2p states. As far as the simple model of chapter 5.9 was concerned, these electrons can do whatever they want within the 2p subshell.

To go one better than that, the correct interactions between the two 2p electrons will need to be considered. To keep the arguments manageable, it will still be assumed that the effects of the 1s and 2s electrons are independent of where the 2p electrons are.

Call the 2p electrons α and β. Under the stated conditions, their Hamiltonian takes the form

$$H_\alpha + H_\beta + V_{\alpha\beta}$$

where H_α and H_β are the single-electron Hamiltonians for the electrons α and β, consisting of their kinetic energy, their attraction to the nucleus, and the repulsion by the 1s and 2s electrons. Note that in the current analysis, it is not required that the 1s and 2s electrons are treated as located in the nucleus. Lack of shielding can be allowed now, but it must still be assumed that the 1s and 2s electrons are unaffected by where the 2p electrons are. In particular, H_α is assumed to be be independent of the position of electron β, and H_β independent of the position of electron α. The mutual repulsion of the two 2p electrons is given by $V_{\alpha\beta} = e^2/4\pi\epsilon_0|\vec{r}_\alpha - \vec{r}_\beta|$.

Now assume that electrons α and β appropriate two single-electron spatial 2p states for themselves, call them ψ_1 and ψ_2. For carbon, ψ_1 can be thought of as the $2p_z$ state and ψ_2 as the $2p_x$ state, The general spatial wave function describing the two electrons takes the generic form

$$a\psi_1(\vec{r}_1)\psi_2(\vec{r}_2) + b\psi_2(\vec{r}_1)\psi_1(\vec{r}_2).$$

The two states ψ_1 and ψ_2 will be taken to be orthonormal, like p_z and p_x are, and then the normalization requirement is that $|a|^2 + |b|^2 = 1$.

The expectation value of energy is

$$\langle a\psi_1\psi_2 + b\psi_2\psi_1|H_\alpha + H_\beta + V_{\alpha\beta}|a\psi_1\psi_2 + b\psi_2\psi_1\rangle.$$

That can be multiplied out and then simplified by noting that in the various inner product integrals involving the single-electron Hamiltonians, the integral over the coordinate unaffected by the Hamiltonian is either zero or one because of orthonormality. Also, the inner product integrals involving $V_{\alpha\beta}$ are pairwise the same, the difference being just a change of names of integration variables.

The simplified expectation energy is then:

$$E_{\psi_1} + E_{\psi_2} + \langle\psi_1\psi_2|V_{\alpha\beta}|\psi_1\psi_2\rangle + (a^*b + b^*a)\langle\psi_1\psi_2|V_{\alpha\beta}|\psi_2\psi_1\rangle.$$

The first two terms are the single-electron energies of states ψ_1 and ψ_2. The third term is the classical repulsion between between two electron charge distributions of strengths $|\psi_1|^2$ and $|\psi_2|^2$. The electrons minimize this third term by going into spatially separated states like the $2p_x$ and $2p_z$ ones, rather than into the same spatial state or into greatly overlapping ones.

The final one of the four terms is the interesting one for Hund's rule; it determines *how* the two electrons occupy the two states ψ_1 and ψ_2, symmetrically or antisymmetrically. Consider the detailed expression for the inner product integral appearing in the term:

$$\langle\psi_1\psi_2|V_{\alpha\beta}|\psi_2\psi_1\rangle = \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} V_{\alpha\beta} f(\vec{r}_1,\vec{r}_2) f^*(\vec{r}_2,\vec{r}_1)\, \mathrm{d}^3\vec{r}_1 \mathrm{d}^3\vec{r}_2$$

where $f(\vec{r}_1,\vec{r}_2) = \psi_2(\vec{r}_1)\psi_1(\vec{r}_2)$.

The sign of this inner product can be guesstimated. If $V_{\alpha\beta}$ would be the same for all electron separation distances, the integral would be zero because of orthonormality of ψ_1 and ψ_2. However, $V_{\alpha\beta}$ favors positions where $\vec{r}_1$ and $\vec{r}_2$ are close to each other; in fact $V_{\alpha\beta}$ is infinitely large if $\vec{r}_1 = \vec{r}_2$. At such a location $f(\vec{r}_1,\vec{r}_2)f^*(\vec{r}_2,\vec{r}_1)$ is a positive real number, so it tends to have a positive real part in regions it really counts. That means the inner product integral should have the same sign as $V_{\alpha\beta}$; it should be repulsive.

And since this integral is multiplied by a^*b+b^*a, the energy is smallest when that is most negative, which is for the antisymmetric spatial state $a = -b$. Since this state takes care of the sign change in the antisymmetrization requirement, the spin state must be unchanged under particle exchange; the spins must be aligned. More precisely, the spin state must be some linear combination of the three triplet states with net spin one. There you have Hund's rule, as an accidental byproduct of the Coulomb repulsion.

This leaves the philosophical question why for the two electrons of the hydrogen molecule in chapter 5.2 the symmetric state is energetically most favorable, while the antisymmetric state is the one for the 2p electrons. The real difference is in the kinetic energy. In both cases, the antisymmetric combination reduces the Coulomb repulsion energy between the electrons, and in the hydrogen molecule model, it also increases the nuclear attraction energy. But in the hydrogen molecule model, the symmetric state achieves a reduction in kinetic

energy that is more than enough to make up for it all. For the 2p electrons, the reduction in kinetic energy is nil. When the positive component wave functions of the hydrogen molecule model are combined into the symmetric state, they allow greater access to fringe areas farther away from the nuclei. Because of the uncertainty principle, less confined electrons tend to have less indeterminacy in momentum, hence less kinetic energy. On the other hand, the 2p states are half positive and half negative, and even their symmetric combination reduces spatial access for the electrons in half the locations.

A.35 The third law

In the simplest formulation, the third law of thermodynamics says that the entropy at absolute zero temperature is zero.

The original theorem is due to Nernst. A more recent formulation is

> "The contribution to the entropy of a system due to each component that is in internal equilibrium disappears at absolute zero." [D. Ter Haar (1966) *Elements of Thermostatistics*. Holt, Rinehart & Winston.]

A more readable version is

> "The entropy of every chemically simple, perfectly crystalline, body equals zero at the absolute zero of temperature." [G.H. Wannier (1966) *Statistical Physics*. Wiley.]

These formulations allow for the existence of meta-stable equilibria. The third law in its simple form assumes that strictly speaking every ground state is reasonably unique and that the system is in true thermal equilibrium. Experimentally however, many substances do not appear to approach zero entropy. Random mixtures as well as ice are examples. They may not be in true equilibrium, but if true equilibrium is not observed, it is academic.

The zero of entropy is important for mixtures, in which you need to add the entropies of the components together correctly. It also has implications for the behavior of various quantities at low temperatures. For example, it implies that the specific heats become zero at absolute zero. To see why, note that in a constant volume or constant pressure process the entropy changes are given by

$$\int \frac{C}{T}\,\mathrm{d}T$$

If the specific heat C would not become zero at $T = 0$, this integral would give an infinite entropy at that temperature instead of zero.

Another consequence of the third law is that it is not possible to bring a system to absolute zero temperature completely even in ideal processes. That

seems pretty self-evident from a classical point of view, but it is not so obvious in quantum terms. The third law also implies that isothermal processes become isentropic when absolute zero temperature is approached.

It may seem that the third law is a direct consequence of the quantum expression for the entropy,

$$S = -k_{\rm B} \sum P_q \ln(P_q)$$

At absolute zero temperature, the system is in the ground state. Assuming that the ground state is not degenerate, there is then only one nonzero probability $P_q = 1$ and for that probability $\ln(P_q)$ is zero. So the entropy is zero.

Even if the ground state is not unique, often it does not make much of a difference. For example, consider the case of a system of I noninteracting spin 1 bosons in a box. If you could really ignore the effect of all particle interactions on the energy, the I spin states would be arbitrary in the ground state. But even then there would be only about $\frac{1}{2}I^2$ different system states with the ground state energy, chapter 5.7. That produces an entropy of only about $-k_{\rm B} \ln(2/I^2)$. It would make the specific entropy proportional to $\ln(I)/I$, which is zero for a large-enough system.

On the other hand, if you ignore electromagnetic spin couplings of nuclei in a crystal, it becomes a different matter. Since the nuclear wave functions have no measurable overlap, to any conceivable accuracy the nuclei can assume independent spatial states. That gets rid of the (anti) symmetrization restrictions on their spin. And then the associated entropy can be nonzero. But of course, if the nuclear spin does not interact with anything, you may be able to ignore its existence altogether.

Even if a system has a unique ground state, the third law is not as trivial as it may seem. Thermodynamics deals not with finite systems but with idealized systems of infinite size. A very simple example illustrates why it makes a difference. Consider the possibility of a hypothetical system whose specific entropy depends on the number of particles I, temperature T, and pressure P as

$$s_{\rm h.s.}(I,T,P) = \frac{IT}{1+IT}$$

This system is consistent with the expression for the third law given above: for a given system size I, the entropy becomes zero at zero temperature. However, the idealized *infinite* system always has entropy 1; its entropy does *not* go to zero for zero temperature. The third law should be understood to say that this hypothetical system does not exist.

If infinite systems seem unphysical, translate it into real-life terms. Suppose your test tube has say $I = 10^{20}$ particles of the hypothetical system in it instead of infinitely many. Then to reduce the specific entropy from 1 to 0.5 would require the temperature to be reduced to a completely impossible 10^{-20} K. And

if you double the number of particles in the test tube, you would need another factor two reduction in temperature. In short, while formally the entropy for the finite hypothetical system goes to zero at absolute zero, the temperatures required to do so have no actual meaning.

A.36 Alternate Dirac equations

If you look in advanced books on quantum mechanics, you will likely find the Dirac equation written in a different form than given in chapter 12.12.

The Hamiltonian eigenvalue problem as given in that section was

$$\left(\alpha_0 mc^2 + \sum_i \alpha_i \widehat{p}_i c\right) \vec{\psi} = E\vec{\psi}$$

where $\vec{\psi}$ was a vector with four components.

Now assume for a moment that ψ is a state of definite momentum. Then the above equation can be rewritten in the form

$$\left(\gamma_0 \frac{E}{c} - \sum_i \gamma_i p_i\right) \vec{\psi} = mc\vec{\psi}$$

The motivation for doing so is that the coefficients of the γ matrices are the components of the relativistic momentum four-vector, chapter 1.3.1.

It is easy to check that the only difference between the α and γ matrices is that γ_1 through γ_3 get a minus sign in front of their bottom element. (Just multiply the original equation by α_0^{-1}/c and rearrange.)

The parenthetical expression above is essentially a four-vector dot product between the gamma matrices and the momentum four-vector. Especially if you give the dot product the wrong sign, as many physicists do. In particular, in the index notation of chapter 1.2.5, the parenthetical expression is then $\gamma^\mu p_\mu$. Feynman hit upon the bright idea to indicate dot products with γ matrices by a slash through the name. So you are likely to find the above equation as

$$\not{p}\vec{\psi} = mc\vec{\psi}$$

Isn't it beautifully concise? Isn't it completely incomprehensible?

Also consider the case that $\vec{\psi}$ is not an energy and momentum eigenfunction. In that case, the equation of interest is found from the usual quantum substitutions that E becomes $\mathrm{i}\hbar\partial/\partial t$ and $\widehat{\vec{p}}$ becomes $\hbar\partial/\mathrm{i}\partial t$. So the rewritten Dirac equation is then:

$$\mathrm{i}\hbar\left(\gamma_0 \frac{1}{c}\frac{\partial}{\partial t} + \sum_i \gamma_i \frac{\partial}{\partial x_i}\right) \vec{\psi} = mc\vec{\psi}$$

In index notation, the parenthetical expression reads $\gamma^\mu \partial_\mu$. So following Feynman

$$\mathrm{i}\hbar \not{\partial} \vec{\psi} = mc\vec{\psi}$$

Now all that the typical physics book wants to add to that is a suitable non-SI system of units. If you use the electron mass m as your unit of mass instead of the kg, c as unit of velocity instead of m/s, and $\hbar$ as your unit of angular momentum instead of kg m^2/s, you get

$$\mathrm{i}\not{\partial} \vec{\psi} = \vec{\psi}$$

No outsider will ever guess what that stands for!

A.37 Maxwell's wave equations

This note derives the wave equations satisfied by electromagnetic fields. The derivation will use standard formulae of vector analysis, as found in, for example, [41, 20.35-45].

The starting point is Maxwell's equations for the electromagnetic field in vacuum:

$$\nabla \cdot \vec{\mathcal{E}} = \frac{\rho}{\epsilon_0} \qquad (1) \qquad\qquad \nabla \cdot \vec{\mathcal{B}} = 0 \qquad (2)$$

$$\nabla \times \vec{\mathcal{E}} = -\frac{\partial \vec{\mathcal{B}}}{\partial t} \qquad (3) \qquad\qquad \nabla \times \vec{\mathcal{B}} = \frac{\vec{\jmath}}{\epsilon_0} + \frac{\partial \vec{\mathcal{E}}}{\partial t} \qquad (4)$$

Here $\vec{\mathcal{E}}$ is the electric field, $\vec{\mathcal{B}}$ the magnetic field, ρ the charge density, $\vec{\jmath}$ the current density, c the constant speed of light, and ϵ_0 is a constant called the permittivity of space. The charge and current densities are related by the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \vec{\jmath} = 0 \qquad (5)$$

To get a wave equation for the electric field, take the curl, $\nabla\times$, of (3) and apply the standard vector identity (D.1), (1) and (4) to get

$$\boxed{\frac{1}{c^2}\frac{\partial^2 \vec{\mathcal{E}}}{\partial t^2} - \nabla^2 \vec{\mathcal{E}} = -\frac{1}{\epsilon_0 c^2}\frac{\partial \vec{\jmath}}{\partial t} - \frac{1}{\epsilon_0}\nabla\rho} \qquad (A.235)$$

Similarly, for the magnetic field take the curl of (4) and use (2) and (3) to get

$$\boxed{\frac{1}{c^2}\frac{\partial^2 \vec{\mathcal{B}}}{\partial t^2} - \nabla^2 \vec{\mathcal{B}} = \frac{1}{\epsilon_0 c^2}\nabla \times \vec{\jmath}} \qquad (A.236)$$

These are uncoupled inhomogeneous wave equations for the components of $\vec{\mathcal{E}}$ and $\vec{\mathcal{B}}$, for given charge and current densities. According to the theory of partial differential equations, these equations imply that effects propagate no faster than the speed of light. You can also see the same thing pretty clearly from the fact that the homogeneous wave equation has solutions like

$$\sin\Big(k(y-ct)+\varphi\Big)$$

which are waves that travel with speed c in the y-direction.

The wave equations for the potentials φ and $\vec{A}$ are next. First note from (2) that the divergence of $\vec{\mathcal{B}}$ is zero. Then vector calculus says that it can be written as the curl of some vector. Call that vector $\vec{A}_0$.

$$\vec{\mathcal{B}} = \nabla \times \vec{A}_0 \tag{6a}$$

Next define

$$\vec{\mathcal{E}}_\varphi \equiv \vec{\mathcal{E}} + \frac{\partial \vec{A}_0}{\partial t}$$

Plug this into (3) to show that the curl of $\vec{\mathcal{E}}_\varphi$ is zero. Then vector calculus says that it can be written as minus the gradient of a scalar. Call this scalar φ_0. Plug that into the expression above to get

$$\vec{\mathcal{E}} = -\nabla\varphi_0 - \frac{\partial \vec{A}_0}{\partial t} \tag{7a}$$

Next, note that if you define modified versions $\vec{A}$ and φ of $\vec{A}_0$ and φ_0 by setting

$$\varphi = \varphi_0 - \frac{\partial \chi}{\partial t} \qquad \vec{A} = \vec{A}_0 + \nabla\chi$$

where χ is *any* arbitrary function of x, y, z, and t, then still

$$\vec{\mathcal{B}} = \nabla \times \vec{A} \tag{6}$$

since the curl of a gradient is always zero, and

$$\vec{\mathcal{E}} = -\nabla\varphi - \frac{\partial \vec{A}}{\partial t} \tag{7}$$

because the two χ terms drop out against each other.

The fact that $\vec{A}_0, \varphi_0$ and $\vec{A}, \varphi$ produce the same physical fields is the famous "gauge property" of the electromagnetic field.

Now you can select χ so that

$$\nabla \cdot \vec{A} + \frac{1}{c^2}\frac{\partial \varphi}{\partial t} = 0 \tag{8}$$

That is known as the "Lorenz condition." A corresponding gauge function is a "Lorenz gauge."

To find the gauge function χ that produces this condition, plug the definitions for $\vec{A}$ and φ in terms of $\vec{A}_0$ and φ_0 into the left hand side of the Lorentz condition. That produces, after a change of sign,

$$\frac{1}{c^2}\frac{\partial^2\chi}{\partial t^2} - \nabla^2\chi - \nabla\cdot\vec{A}_0 - \frac{1}{c^2}\frac{\partial\varphi_0}{\partial t} = 0$$

That is a second order inhomogeneous wave equation for χ.

Now plug the expressions (6) and (7) for $\vec{\mathcal{E}}$ and $\vec{\mathcal{B}}$ in terms of $\vec{A}$ and φ into the Maxwell's equations. Equations (2) and (3) are satisfied automatically. From (2), after using (8),

$$\boxed{\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} - \nabla^2\varphi = \frac{\rho}{\epsilon_0}} \tag{A.237}$$

From (4), after using (8),

$$\boxed{\frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2} - \nabla^2\vec{A} = \frac{\vec{\jmath}}{\epsilon_0 c^2}} \tag{A.238}$$

You can still select the two initial conditions for χ. The smart thing to do is select them so that φ and its time derivative are zero at time zero. In that case, if there is no charge density, φ will stay zero for all time. That is because its wave equation is then homogeneous. The Lorenz condition will then ensure that $\nabla\cdot\vec{A}$ is zero too.

Instead of the Lorenz condition, you could select χ to make $\nabla\cdot\vec{A}$ zero. That is called the "Coulomb gauge" or "transverse gauge" or "transverse gauge." It requires that χ satisfies the Poisson equation

$$-\nabla^2\chi = \nabla\cdot\vec{A}_0$$

Then the governing equations become

$$-\nabla^2\varphi = \frac{\rho}{\epsilon_0}$$

$$\frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2} - \nabla^2\vec{A} = \frac{\vec{\jmath}}{\epsilon_0 c^2} - \frac{1}{c^2}\nabla\frac{\partial\varphi}{\partial t}$$

Note that φ now satisfies a purely spatial Poisson equation.

A.38 Perturbation Theory

Most of the time in quantum mechanics, exact solution of the Hamiltonian eigenvalue problem of interest is not possible. To deal with that, approximations are made.

Perturbation theory can be used when the Hamiltonian H consists of two parts H_0 and H_1, where the problem for H_0 can be solved and where H_1 is small. The idea is then to adjust the found solutions for the "unperturbed Hamiltonian" H_0 so that they become approximately correct for $H_0 + H_1$.

This addendum explains how perturbation theory works. It also gives a few simple but important examples: the helium atom and the Zeeman and Stark effects. Addendum,{A.39} will use the approach to study relativistic effects on the hydrogen atom.

A.38.1 Basic perturbation theory

To use perturbation theory, the eigenfunctions and eigenvalues of the unperturbed Hamiltonian H_0 must be known. These eigenfunctions will here be indicated as $\psi_{\vec{n},0}$ and the corresponding eigenvalues by $E_{\vec{n},0}$. Note the use of the generic $\vec{n}$ to indicate the quantum numbers of the eigenfunctions. If the basic system is an hydrogen atom, as is often the case in textbook examples, and spin is unimportant, $\vec{n}$ would likely stand for the set of quantum numbers n, l, and m. But for a three-dimensional harmonic oscillator, $\vec{n}$ might stand for the quantum numbers n_x, n_y, and n_z. In a three-dimensional problem with one spinless particle, it takes three quantum numbers to describe an energy eigenfunction. However, which three depends on the problem and your approach to it. The additional subscript 0 in $\psi_{\vec{n},0}$ and $E_{\vec{n},0}$ indicates that they ignore the perturbation Hamiltonian H_1. They are called the unperturbed wave functions and energies.

The key to perturbation theory are the "Hamiltonian perturbation coefficients" defined as

$$\boxed{H_{\underline{\vec{n}}\vec{n},1} \equiv \langle \psi_{\underline{\vec{n}},0} | H_1 \psi_{\vec{n},0} \rangle} \tag{A.239}$$

If you can evaluate these for every pair of energy eigenfunctions, you should be OK. Note that evaluating inner products is just summation or integration; it is generally a lot simpler than trying to solve the eigenvalue problem $(H_0 + H_1)\,\psi = E\psi$.

In the application of perturbation theory, the idea is to pick one unperturbed eigenfunction $\psi_{\vec{n},0}$ of H_0 of interest and then correct it to account for H_1, and especially correct its energy $E_{\vec{n},0}$. Caution! If the energy $E_{\vec{n},0}$ is degenerate, i.e. there is more than one unperturbed eigenfunction $\psi_{\vec{n},0}$ of H_0 with that energy, you must use a "good" eigenfunction to correct the energy. How to do that will be discussed in subsection A.38.3.

For now just assume that the energy is not degenerate or that you picked a good eigenfunction $\psi_{\vec{n},0}$. Then a first correction to the energy $E_{\vec{n},0}$ to account for the perturbation H_1 is very simple, {D.79}; just add the corresponding Hamiltonian perturbation coefficient:

$$\boxed{E_{\vec{n}} = E_{\vec{n},0} + H_{\vec{n}\vec{n},1} + \ldots} \tag{A.240}$$

This is a quite user-friendly result, because it only involves the selected energy eigenfunction $\psi_{\vec{n},0}$. The other energy eigenfunctions are not involved. In a numerical solution, you might only have computed one state, say the ground state of H_0. Then you can use this result to correct the ground state energy for a perturbation even if you do not have data about any other energy states of H_0.

Unfortunately, it does happen quite a lot that the above correction $H_{\vec{n}\vec{n},1}$ is zero because of some symmetry or the other. Or it may simply not be accurate enough. In that case, to find the energy change you have to use what is called "second order perturbation theory:"

$$\boxed{E_{\vec{n}} = E_{\vec{n},0} + H_{\vec{n}\vec{n},1} - \sum_{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}} \frac{|H_{\underline{\vec{n}}\vec{n},1}|^2}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} + \ldots} \tag{A.241}$$

Now all eigenfunctions of H_0 will be needed, which makes second order theory a lot nastier. Then again, even if the "first order" correction $H_{\vec{n}\vec{n},1}$ to the energy is nonzero, the second order formula will likely give a much more accurate result.

Sometimes you may also be interested in what happens to the energy eigenfunctions, not just the energy eigenvalues. The corresponding formula is

$$\boxed{\psi_{\vec{n}} = \psi_{\vec{n},0} - \sum_{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}} \frac{H_{\underline{\vec{n}}\vec{n},1}}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} \psi_{\underline{\vec{n}},0} + \sum_{\substack{E_{\underline{\vec{n}},0} = E_{\vec{n},0} \\ \underline{\vec{n}} \neq \vec{n}}} c_{\underline{\vec{n}}} \psi_{\underline{\vec{n}},0} + \ldots} \tag{A.242}$$

That is the first order result. The second sum is zero if the problem is not degenerate. Otherwise its coefficients $c_{\underline{\vec{n}}}$ are determined by considerations found in derivation {D.79}.

In some cases, instead of using second order theory as above, it may be simpler to compute the first order wave function perturbation and the second order energy change from

$$\boxed{(H_0 - E_{\vec{n},0})\psi_{\vec{n},1} = -(H_1 - E_{\vec{n},1})\psi_{\vec{n},0} \qquad E_{\vec{n},2} = \langle \psi_{\vec{n},0} | (H_1 - E_{\vec{n},1})\psi_{\vec{n},1} \rangle} \tag{A.243}$$

Eigenfunction $\psi_{\vec{n},0}$ must be good. The good news is that this does not require all the unperturbed eigenfunctions. The bad news is that it requires solution of a nontrivial equation involving the unperturbed Hamiltonian instead of just

integration. It may be the best way to proceed for a perturbation of a numerical solution.

One application of perturbation theory is the "Hellmann-Feynman theorem." Here the perturbation Hamiltonian is an infinitesimal change ∂H in the unperturbed Hamiltonian caused by an infinitesimal change in some parameter that it depends on. If the parameter is called λ, perturbation theory says that the first order energy change is

$$\boxed{\frac{\partial E_{\vec{n}}}{\partial \lambda} = \left\langle \psi_{\vec{n},0} \middle| \frac{\partial H}{\partial \lambda} \psi_{\vec{n},0} \right\rangle} \tag{A.244}$$

when divided by the change in parameter $\partial\lambda$. If you can figure out the inner product, you can figure out the change in energy. But more important is the reverse: if you can find the derivative of the energy with respect to the parameter, you have the inner product. For example, the Hellmann-Feynman theorem is helpful for finding the expectation value of $1/r^2$ for the hydrogen atom, a nasty problem, {D.83}. Of course, always make sure the eigenfunction $\psi_{\vec{n},0}$ is a good one for the derivative of the Hamiltonian.

A.38.2 Ionization energy of helium

One prominent deficiency in the approximate analysis of the heavier atoms in chapter 5.9 was the poor ionization energy that it gave for helium. The purpose of this example is to derive a much more reasonable value using perturbation theory.

Exactly speaking, the ionization energy is the difference between the energy of the helium atom with both its electrons in the ground state and the helium ion with its second electron removed. Now the energy of the helium ion with electron 2 removed is easy; the Hamiltonian for the remaining electron 1 is

$$H_{\text{He ion}} = -\frac{\hbar^2}{2m_{\text{e}}}\nabla_1^2 - 2\frac{e^2}{4\pi\epsilon_0}\frac{1}{r_1}$$

where the first term represents the kinetic energy of the electron and the second its attraction to the two-proton nucleus. The helium nucleus normally also contains two neutrons, but they do not attract the electron.

This Hamiltonian is exactly the same as the one for the hydrogen atom in chapter 4.3, except that it has $2e^2$ where the hydrogen one, with just one proton in its nucleus, has e^2. So the solution for the helium ion is simple: just take the hydrogen solution, and everywhere where there is an e^2 in that solution, replace it by $2e^2$. In particular, the Bohr radius a for the helium ion is half the Bohr radius a_0 for hydrogen,

$$a = \frac{4\pi\epsilon_0\hbar^2}{m_{\text{e}}2e^2} = \tfrac{1}{2}a_0$$

and so its energy and wave function become

$$E_{\rm gs,ion} = -\frac{\hbar^2}{2m_{\rm e}a^2} = 4E_1 \qquad \psi_{\rm gs,ion}(\vec{r}) = \frac{1}{\sqrt{\pi a^3}} e^{-r/a}$$

where $E_1 = -13.6$ eV is the energy of the hydrogen atom.

It is interesting to see that the helium ion has four times the energy of the hydrogen atom. The reasons for this much higher energy are both that the nucleus is twice as strong, and that the electron is twice as close to it: the Bohr radius is half the size. More generally, in heavy atoms the electrons that are poorly shielded from the nucleus, which means the inner electrons, have energies that scale with the square of the nuclear strength. For such electrons, relativistic effects are much more important than they are for the electron in a hydrogen atom.

The neutral helium atom is not by far as easy to analyze as the ion. Its Hamiltonian is, from (5.34):

$$H_{\rm He} = -\frac{\hbar^2}{2m_{\rm e}}\nabla_1^2 - 2\frac{e^2}{4\pi\epsilon_0}\frac{1}{r_1} - \frac{\hbar^2}{2m_{\rm e}}\nabla_2^2 - 2\frac{e^2}{4\pi\epsilon_0}\frac{1}{r_2} + \frac{e^2}{4\pi\epsilon_0}\frac{1}{|\vec{r}_2 - \vec{r}_1|}$$

The first two terms are the kinetic energy and nuclear attraction of electron 1, and the next two the same for electron 2. The final term is the electron to electron repulsion, the curse of quantum mechanics. This final term is the reason that the ground state of helium cannot be found analytically.

Note however that the repulsion term is qualitatively similar to the nuclear attraction terms, except that there are four of these nuclear attraction terms versus a single repulsion term. So maybe then, it may work to treat the repulsion term as a small perturbation, call it H_1, to the Hamiltonian H_0 given by the first four terms? Of course, if you ask mathematicians whether 25% is a small amount, they are going to vehemently deny it; but then, so they would for any amount if there is no limit process involved, so just don't ask them, OK?

The solution of the eigenvalue problem $H_0\psi = E\psi$ is simple: since the electrons do not interact with this Hamiltonian, the ground state wave function is the product of the ground state wave functions for the individual electrons, and the energy is the sum of their energies. And the wave functions and energies for the separate electrons are given by the solution for the ion above, so

$$\psi_{\rm gs,0} = \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \qquad E_{\rm gs,0} = 8E_1$$

According to this result, the energy of the atom is $8E_1$ while the ion had $4E_1$, so the ionization energy would be $4|E_1|$, or 54.4 eV. Since the experimental value is 24.6 eV, this is no better than the 13.6 eV chapter 5.9 came up with.

To get a better ionization energy, try perturbation theory. According to first order perturbation theory, a better value for the energy of the hydrogen atom should be

$$E_{\rm gs} = E_{\rm gs,0} + \langle \psi_{\rm gs,0} | H_1 \psi_{\rm gs,0} \rangle$$

or substituting in from above,

$$E_{\rm gs} = 8E_1 + \frac{e^2}{4\pi\epsilon_0}\left\langle \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \right| \frac{1}{|\vec{r}_2 - \vec{r}_1|} \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \left. \right\rangle$$

The inner product of the final term can be written out as

$$\frac{e^2}{4\pi\epsilon_0} \frac{1}{\pi^2 a^6} \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} \frac{e^{-2(r_1+r_2)/a}}{|\vec{r}_2 - \vec{r}_1|} \, {\rm d}^3\vec{r}_1 {\rm d}^3\vec{r}_2$$

This integral can be done analytically. Try it, if you are so inclined; integrate ${\rm d}^3\vec{r}_1$ first, using spherical coordinates with $\vec{r}_2$ as their axis and doing the azimuthal and polar angles first. Be careful, $\sqrt{(r_1 - r_2)^2} = |r_1 - r_2|$, not $r_1 - r_2$, so you will have to integrate $r_1 < r_2$ and $r_1 > r_2$ separately in the final integration over ${\rm d}r_1$. Then integrate ${\rm d}^3\vec{r}_2$.

The result of the integration is

$$\frac{e^2}{4\pi\epsilon_0}\left\langle \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \right| \frac{1}{|\vec{r}_2 - \vec{r}_1|} \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \left. \right\rangle = \frac{e^2}{4\pi\epsilon_0} \frac{5}{8a} = \frac{5}{2}|E_1|$$

Therefore, the helium atom energy increases by $2.5|E_1|$ due to the electron repulsion, and with it, the ionization energy decreases to $1.5|E_1|$, or 20.4 eV. It is not 24.6 eV, but it is clearly much more reasonable than 54 or 13.6 eV were.

The second order perturbation result should give a much more accurate result still. However, if you did the integral above, you may feel little inclination to try the ones involving all possible products of hydrogen energy eigenfunctions.

Instead, the result can be improved using a variational approach, like the ones that were used earlier for the hydrogen molecule and molecular ion, and this requires almost no additional work. The idea is to accept the hint from perturbation theory that the wave function of helium can be approximated as $\psi_a(\vec{r}_1)\psi_a(\vec{r}_2)$ where ψ_a is the hydrogen ground state wave function using a modified Bohr radius a instead of a_0:

$$\psi_{\rm gs} = \psi_a(\vec{r}_1)\psi_a(\vec{r}_2) \qquad \psi_a(\vec{r}) \equiv \frac{1}{\sqrt{\pi a^3}} e^{-r/a}$$

However, instead of accepting the perturbation theory result that a should be half the normal Bohr radius a_0, let a be optimized to make the expectation energy for the ground state

$$E_{\rm gs} = \langle \psi_{\rm gs} | H_{\rm He} \psi_{\rm gs} \rangle$$

as small as possible. This will produce the most accurate ground state energy possible for a ground state wave function of this form, guaranteed no worse than assuming that $a = \frac{1}{2}a_0$, and probably better.

No new integrals need to be done to evaluate the inner product above. Instead, noting that for the hydrogen atom according to the virial theorem of chapter 7.2 the expectation kinetic energy equals $-E_1 = \hbar^2/2m_e a_0^2$ and the potential energy equals $2E_1$, two of the needed integrals can be inferred from the hydrogen solution: chapter 4.3,

$$\left\langle \psi_a \middle| -\frac{\hbar^2}{2m_e}\nabla^2\psi_a \right\rangle = \frac{\hbar^2}{2m_e a^2}$$

$$-\frac{e^2}{4\pi\epsilon_0}\left\langle \psi_a \middle| \frac{1}{r}\psi_a \right\rangle = -\frac{\hbar^2}{m_e a_0}\left\langle \psi_a \middle| \frac{1}{r}\psi_a \right\rangle = -\frac{\hbar^2}{m_e a_0}\frac{1}{a}$$

and this subsection added

$$\left\langle \psi_a\psi_a \middle| \frac{1}{|\vec{r}_2 - \vec{r}_1|}\psi_a\psi_a \right\rangle = \frac{5}{8a}$$

Using these results with the helium Hamiltonian, the expectation energy of the helium atom can be written out to be

$$\langle \psi_a\psi_a | H_{\rm He}\psi_a\psi_a \rangle = \frac{\hbar^2}{m_e a^2} - \frac{27}{8}\frac{\hbar^2}{m_e a_0 a}$$

Setting the derivative with respect to a to zero locates the minimum at $a = \frac{16}{27}a_0$, rather than $\frac{1}{2}a_0$. Then the corresponding expectation energy is $-3^6\hbar^2/2^8 m_e a_0^2$, or $3^6 E_1/2^7$. Putting in the numbers, the ionization energy is now found as 23.1 eV, in quite good agreement with the experimental 24.6 eV.

A.38.3 Degenerate perturbation theory

Energy eigenvalues are degenerate if there is more than one independent eigenfunction with that energy. Now, if you try to use perturbation theory to correct a degenerate eigenvalue of a Hamiltonian H_0 for a perturbation H_1, there may be a problem. Assume that there are $d > 1$ independent eigenfunctions with energy $E_{\vec{n},0}$ and that they are numbered as

$$\psi_{\vec{n}_1,0}, \psi_{\vec{n}_2,0}, \ldots, \psi_{\vec{n}_d,0}$$

Then as far as H_0 is concerned, any combination

$$\psi_{\vec{n},0} = c_1\psi_{\vec{n}_1,0} + c_2\psi_{\vec{n}_2,0} + \ldots + c_d\psi_{\vec{n}_d,0}$$

with arbitrary coefficients $c_1, c_2, \ldots, c_d$, (not all zero, of course), is just as good an eigenfunction with energy $E_{\vec{n},0}$ as any other.

Unfortunately, the full Hamiltonian $H_0 + H_1$ is not likely to agree with H_0 about that. As far as the full Hamiltonian is concerned, normally only very *specific* combinations are acceptable, the "good" eigenfunctions. It is said that

the perturbation H_1 "breaks the degeneracy" of the energy eigenvalue. The single energy eigenvalue splits into several eigenvalues of different energy. Only good combinations will show up these changed energies; the bad ones will pick up uncertainty in energy that hides the effect of the perturbation.

The various ways of ensuring good eigenfunctions are illustrated in the following subsections for example perturbations of the energy levels of the hydrogen atom. Recall that the unperturbed energy eigenfunctions of the hydrogen atom electron, as derived in chapter 4.3, and also including spin, are given as $\psi_{nlm}\uparrow$ and $\psi_{nlm}\downarrow$. They are highly degenerate: all the eigenfunctions with the same value of n have the same energy E_n, regardless of what is the value of the azimuthal quantum number $0 \leqslant l \leqslant n-1$ corresponding to the square orbital angular momentum $L^2 = l(l+1)\hbar^2$; regardless of what is the magnetic quantum number $|m| \leqslant l$ corresponding to the orbital angular momentum $L_z = m\hbar$ in the z-direction; and regardless of what is the spin quantum number $m_s = \pm\frac{1}{2}$ corresponding to the spin angular momentum $m_s\hbar$ in the z-direction. In particular, the ground state energy level E_1 is two-fold degenerate, it is the same for both $\psi_{100}\uparrow$, i.e. $m_s = \frac{1}{2}$ and $\psi_{100}\downarrow$, $m_s = -\frac{1}{2}$. The next energy level E_2 is eight-fold degenerate, it is the same for $\psi_{200}\updownarrow$, $\psi_{211}\updownarrow$, $\psi_{210}\updownarrow$, and $\psi_{21-1}\updownarrow$, and so on for higher values of n.

There are two important rules to identify the good eigenfunctions, {D.79}:

1. Look for good quantum numbers. The quantum numbers that make the energy eigenfunctions of the unperturbed Hamiltonian H_0 unique correspond to the eigenvalues of additional operators besides the Hamiltonian. If the perturbation Hamiltonian H_1 commutes with one of these additional operators, the corresponding quantum number is good. You do not need to combine eigenfunctions with different values of that quantum number.

   In particular, if the perturbation Hamiltonian commutes with all additional operators that make the eigenfunctions of H_0 unique, stop worrying: every eigenfunction is good already.

   For example, for the usual hydrogen energy eigenfunctions $\psi_{nlm}\updownarrow$, the quantum numbers l, m, and m_s make the eigenfunctions at a given unperturbed energy level n unique. They correspond to the operators $\widehat{L}^2$, $\widehat{L}_z$, and $\widehat{S}_z$. If the perturbation Hamiltonian H_1 commutes with any one of these operators, the corresponding quantum number is good. If the perturbation commutes with all three, all eigenfunctions are good already.

2. Even if some quantum numbers are bad because the perturbation does not commute with that operator, eigenfunctions are still good if there are no other eigenfunctions with the same unperturbed energy and the same good quantum numbers.

   Otherwise linear algebra is required. For each set of energy eigen-

functions

$$\psi_{\vec{n}_1,0}, \psi_{\vec{n}_2,0}, \ldots$$

with the same unperturbed energy and the same good quantum numbers, but different bad ones, form the matrix of Hamiltonian perturbation coefficients

$$\begin{pmatrix} \langle\psi_{\vec{n}_1,0}|H_1\psi_{\vec{n}_1,0}\rangle & \langle\psi_{\vec{n}_1,0}|H_1\psi_{\vec{n}_2,0}\rangle & \cdots \\ \langle\psi_{\vec{n}_2,0}|H_1\psi_{\vec{n}_1,0}\rangle & \langle\psi_{\vec{n}_2,0}|H_1\psi_{\vec{n}_2,0}\rangle & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

The eigenvalues of this matrix are the first order energy corrections. Also, the coefficients $c_1, c_2, \ldots$ of each good eigenfunction

$$c_1\psi_{\vec{n}_1,0} + c_2\psi_{\vec{n}_2,0} + \ldots$$

must be an eigenvector of the matrix.

Unfortunately, if the eigenvalues of this matrix are not all different, the eigenvectors are not unique, so you remain unsure about what are the good eigenfunctions. In that case, if the second order energy corrections are needed, the detailed analysis of derivation {D.79} will need to be followed.

If you are not familiar with linear algebra at all, in all cases mentioned here the matrices are just two by two, and you can find that solution spelled out in the notations under "eigenvector."

The following, related, practical observation can also be made:

Hamiltonian perturbation coefficients can only be nonzero if all the good quantum numbers are the same.

A.38.4 The Zeeman effect

If you put an hydrogen atom in an external magnetic field $\vec{\mathcal{B}}_{\rm ext}$, the energy levels of the electron change. That is called the "Zeeman effect."

If for simplicity a coordinate system is used with its z-axis aligned with the magnetic field, then according to chapter 13.4, the Hamiltonian of the hydrogen atom acquires an additional term

$$H_1 = \frac{e}{2m_{\rm e}}\mathcal{B}_{\rm ext}\left(\widehat{L}_z + 2\widehat{S}_z\right) \tag{A.245}$$

beyond the basic hydrogen atom Hamiltonian H_0 of chapter 4.3.1. Qualitatively, it expresses that a spinning charged particle is equivalent to a tiny electromagnet, and a magnet wants to align itself with a magnetic field, just like a compass needle aligns itself with the magnetic field of earth.

For this perturbation, the $\psi_{nml}\updownarrow$ energy eigenfunctions are already good ones, because H_1 commutes with all of $\widehat{L}^2$, $\widehat{L}_z$ and $\widehat{S}_z$. So, according to perturbation theory, the energy eigenvalues of an hydrogen atom in a magnetic field are approximately

$$E_n + \langle \psi_{nml}\updownarrow | H_1 | \psi_{nml}\updownarrow \rangle = E_n + \frac{e}{2m_{\rm e}} \mathcal{B}_{\rm ext} (m + 2m_s)\hbar$$

Actually, this is not approximate at all; it is the exact eigenvalue of $H_0 + H_1$ corresponding to the exact eigenfunction $\psi_{nml}\updownarrow$.

The Zeeman effect can be seen in an experimental spectrum. Consider first the ground state. If there is no electromagnetic field, the two ground states $\psi_{100}\uparrow$ and $\psi_{100}\downarrow$ would have exactly the same energy. Therefore, in an experimental spectrum, they would show up as a single line. But with the magnetic field, the two energy levels are different,

$$E_{100\downarrow} = E_1 - \frac{e\hbar}{2m_{\rm e}} \mathcal{B}_{\rm ext} \qquad E_{100\uparrow} = E_1 + \frac{e\hbar}{2m_{\rm e}} \mathcal{B}_{\rm ext} \qquad E_1 = -13.6 \text{ eV}$$

so the single line splits into two! Do note that the energy change due to even an extremely strong magnetic field of 100 Tesla is only 0.006 eV or so, chapter 13.4, so it is not like the spectrum would become unrecognizable. The single spectral line of the eight $\psi_{2lm}\updownarrow$ "L" shell states will similarly split in five closely spaced but separate lines, corresponding to the five possible values -2, -1, 0, 1 and 2 for the factor $m + 2m_s$ above.

Some disclaimers should be given here. First of all, the 2 in $m + 2m_s$ is only equal to 2 up to about 0.1% accuracy. More importantly, even in the absence of a magnetic field, the energy levels at a given value of n do not really form a single line in the spectrum if you look closely enough. There are small errors in the solution of chapter 4.3 due to relativistic effects, and so the theoretical lines are already split. That is discussed in addendum {A.39}. The description given above is a good one for the "strong" Zeeman effect, in which the magnetic field is strong enough to swamp the relativistic errors.

A.38.5 The Stark effect

If an hydrogen atom is placed in an external electric field $\vec{\mathcal{E}}_{\rm ext}$ instead of the magnetic one of the previous subsection, its energy levels will change too. That is called the "Stark effect." Of course a Zeeman, Dutch for sea-man, would be most interested in magnetic fields. A Stark, maybe in a spark? (Apologies.)

If the z-axis is taken in the direction of the electric field, the contribution of the electric field to the Hamiltonian is given by:

$$H_1 = e\mathcal{E}_{\rm ext} z \tag{A.246}$$

It is much like the potential energy mgh of gravity, with the electron charge e taking the place of the mass m, $\mathcal{E}_{\rm ext}$ that of the gravity strength g, and z that of the height h.

Since the typical magnitude of z is of the order of a Bohr radius a_0, you would expect that the energy levels will change due to the electric field by an amount of rough size $e\mathcal{E}_{\rm ext}a_0$. A strong laboratory electric field might have $e\mathcal{E}_{\rm ext}a_0$ of the order of 0.000 5 eV, [25, p. 339]. That is really small compared to the typical electron energy levels.

And additionally, it turns out that for many eigenfunctions, including the ground state, the first order correction to the energy is zero. To get the energy change in that case, you need to compute the second order term, which is a pain. And that term will be much smaller still than even $e\mathcal{E}_{\rm ext}a_0$ for reasonable field strengths.

Now first suppose that you ignore the warnings on good eigenfunctions, and just compute the energy changes using the inner product $\langle\psi_{nlm}\updownarrow|H_1\psi_{nlm}\updownarrow\rangle$. You will then find that this inner product is zero for whatever energy eigenfunction you take:

$$\langle\psi_{nlm}\updownarrow|e\mathcal{E}_{\rm ext}z\psi_{nlm}\updownarrow\rangle = 0 \text{ for all } n,\ l,\ m, \text{ and } m_s$$

The reason is that negative z values integrate away against positive ones. (The inner products are integrals of z times $|\psi_{nlm}|^2$, and $|\psi_{nlm}|^2$ is the same at opposite sides of the nucleus while z changes sign, so the contributions of opposite sides to the inner product pairwise cancel.)

So, since all first order energy changes that you compute are zero, you would naturally conclude that to first order approximation none of the energy levels of a hydrogen atom changes due to the electric field. But that conclusion is wrong for anything but the ground state energy. And the reason it is wrong is because the good eigenfunctions have not been used.

Consider the operators $\widehat{L}^2$, $\widehat{L}_z$, and S_z that make the energy eigenfunctions $\psi_{nlm}\updownarrow$ unique. If $H_1 = e\mathcal{E}_{\rm ext}z$ commuted with them all, the $\psi_{nlm}\updownarrow$ would be good eigenfunctions. Unfortunately, while z commutes with $\widehat{L}_z$ and S_z, it does not commute with $\widehat{L}^2$, see chapter 4.5.4. The quantum number l is bad.

Still, the two states $\psi_{100}\updownarrow$ with the ground state energy are good states, because there are no states with the same energy and a different value of the bad quantum number l. Really, spin has nothing to do with the Stark problem. If you want, you can find the purely spatial energy eigenfunctions first, then for every spatial eigenfunction, there will be one like that with spin up and one with spin down. In any case, since the two eigenfunctions $\psi_{100}\updownarrow$ are both good, the ground state energy does indeed not change to first order.

But now consider the eight-fold degenerate $n = 2$ energy level. Each of the four eigenfunctions $\psi_{211}\updownarrow$ and $\psi_{21-1}\updownarrow$ is a good one because for each of them, there is no other $n = 2$ eigenfunction with a different value of the bad quantum

number l. The energies corresponding to these good eigenfunctions too do indeed not change to first order.

However, the remaining two $n = 2$ spin-up states $\psi_{200}\uparrow$ and $\psi_{210}\uparrow$ have different values for the bad quantum number l, and they have the same values $m = 0$ and $m_s = \frac{1}{2}$ for the good quantum numbers of orbital and spin z-momentum. These eigenfunctions are bad and will have to be combined to produce good ones. And similarly the remaining two spin-down states $\psi_{200}\downarrow$ and $\psi_{210}\downarrow$ are bad and will have to be combined.

It suffices to just analyze the spin up states, because the spin down ones go exactly the same way. The coefficients c_1 and c_2 of the good combinations $c_1\psi_{200}\uparrow + c_2\psi_{210}\uparrow$ must be eigenvectors of the matrix

$$\begin{pmatrix} \langle\psi_{200}\uparrow|H_1\psi_{200}\uparrow\rangle & \langle\psi_{200}\uparrow|H_1\psi_{210}\uparrow\rangle \\ \langle\psi_{210}\uparrow|H_1\psi_{200}\uparrow\rangle & \langle\psi_{210}\uparrow|H_1\psi_{210}\uparrow\rangle \end{pmatrix} \qquad H_1 = e\mathcal{E}_{\rm ext}z$$

The "diagonal" elements of this matrix (top left corner and bottom right corner) are zero because of cancellation of negative and positive z values as discussed above. And the top right and bottom left elements are complex conjugates, (2.16), so only one of them needs to be actually computed. And the spin part of the inner product produces one and can therefore be ignored. What is left is a matter of finding the two spatial eigenfunctions involved according to (4.36), looking up the spherical harmonics in table 4.2 and the radial functions in table 4.4, and integrating it all against $e\mathcal{E}_{\rm ext}z$. The resulting matrix is

$$\begin{pmatrix} 0 & -3e\mathcal{E}_{\rm ext}a_0 \\ -3e\mathcal{E}_{\rm ext}a_0 & 0 \end{pmatrix}$$

The eigenvectors of this matrix are simple enough to guess; they have either equal or opposite coefficients c_1 and c_2:

$$\begin{pmatrix} 0 & -3e\mathcal{E}_{\rm ext}a_0 \\ -3e\mathcal{E}_{\rm ext}a_0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{pmatrix} = -3e\mathcal{E}_{\rm ext}a_0 \begin{pmatrix} \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{pmatrix}$$

$$\begin{pmatrix} 0 & -3e\mathcal{E}_{\rm ext}a_0 \\ -3e\mathcal{E}_{\rm ext}a_0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} \end{pmatrix} = 3e\mathcal{E}_{\rm ext}a_0 \begin{pmatrix} \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} \end{pmatrix}$$

If you want to check these expressions, note that the product of a matrix times a vector is found by taking dot products between the rows of the matrix and the vector. It follows that the good combination $\sqrt{\frac{1}{2}}\psi_{200}\uparrow + \sqrt{\frac{1}{2}}\psi_{210}\uparrow$ has a first order energy change $-3e\mathcal{E}_{\rm ext}a_0$, and the good combination $\sqrt{\frac{1}{2}}\psi_{200}\uparrow - \sqrt{\frac{1}{2}}\psi_{210}\uparrow$ has $+3e\mathcal{E}_{\rm ext}a_0$. The same applies for the spin down states. It follows that to first order the $n = 2$ level splits into three, with energies $E_2 - 3e\mathcal{E}_{\rm ext}a_0$, E_2, and

$E_2 + 3e\mathcal{E}_{\rm ext}a_0$, where the value E_2 applies to the eigenfunctions $\psi_{211}\updownarrow$ and $\psi_{21-1}\updownarrow$ that were already good. The conclusion, based on the wrong eigenfunctions, that the energy levels do not change was all wrong.

Remarkably, the good combinations of ψ_{200} and ψ_{210} are the "sp" hybrids of carbon fame, as described in chapter 5.11.4. Note from figure 5.13 in that section that these hybrids do *not* have the same magnitude at opposite sides of the nucleus. They have an intrinsic "electric dipole moment," with the charge shifted towards one side of the atom, and the electron then wants to align this dipole moment with the ambient electric field. That is much like in Zeeman splitting, where electron wants to align its orbital and spin magnetic dipole moments with the ambient magnetic field.

The crucial thing to take away from all this is: always, always, check whether the eigenfunction is good before applying perturbation theory.

It is obviously somewhat disappointing that perturbation theory did not give any information about the energy change of the ground state beyond the fact that it is second order, i.e. very small compared to $e\mathcal{E}_{\rm ext}a_0$. You would like to know approximately what it is, not just that it is very small. Of course, now that it is established that $\psi_{100}\uparrow$ is a good state with $m = 0$ and $m_s = \frac{1}{2}$, you could think about evaluating the second order energy change (A.241), by integrating $\langle\psi_{100}\uparrow|e\mathcal{E}_{\rm ext}z\psi_{nl0}\uparrow\rangle$ for all values of n and l. But after refreshing your memory about the analytical expression (D.8) for the ψ_{nlm}, you might think again.

It is however possible to find the perturbation in the wave function from the alternate approach (A.243), {D.80}. In that way the second order ground state energy is found to be

$$E_{100} = E_1 - \frac{3e\mathcal{E}_{\rm ext}a_0}{8|E_1|}3e\mathcal{E}_{\rm ext}a_0 \qquad E_1 = -13.6 \text{ eV}$$

Note that the atom likes an electric field: it lowers its ground state energy. Also note that the energy change is indeed second order; it is proportional to the square of the electric field strength. You can think of the attraction of the atom to the electric field as a two-stage process: first the electric field polarizes the atom by distorting its initially symmetric charge distribution. Then it interacts with this polarized atom in much the same way that it interacts with the sp hybrids. But since the polarization is now only proportional to the field strength, the net energy drop is proportional to the square of the field strength.

Finally, note that the typical value of 0.000 5 eV or so for $e\mathcal{E}_{\rm ext}a_0$ quoted earlier is very small compared to the about 100 eV for $8|E_1|$, making the fraction in the expression above very small. So, indeed the second order change in the ground state energy E_1 is much smaller than the first order energy changes $\pm 3e\mathcal{E}_{\rm ext}a_0$ in the E_2 energy level.

A weird prediction of quantum mechanics is that the electron will eventually escape from the atom, leaving it ionized. The reason is that the potential is linear in z, so if the electron goes out far enough in the z-direction, it will

eventually encounter potential energies that are lower than the one it has in the atom. Of course, to get at such large values of z, the electron must pass positions where the required energy far exceeds the -13.6 eV it has available, and that is impossible for a classical particle. However, in quantum mechanics the position of the electron is uncertain, and the electron does have some miniscule chance of "tunneling out" of the atom through the energy barrier, chapter 7.12.2. Realistically, though, for even strong experimental fields like the one mentioned above, the "life time" of the electron in the atom before it has a decent chance of being found outside it far exceeds the age of the universe.

A.39 The relativistic hydrogen atom

The description of the hydrogen atom given earlier in chapter 4.3 is very accurate by engineering standards. However, it is not exact. This addendum examines various relativistic effects that were ignored in the analysis.

The approach will be to take the results of chapter 4.3 as the starting point. Then corrections are applied to them using perturbation theory as described in addendum {A.38}.

A.39.1 Introduction

According to the description of the hydrogen atom given in chapter 4.3, all energy eigenfunctions $\psi_{nlm}\updownarrow$ with the same value of n have the same energy E_n. Therefore they should show up as a single line in an experimental line spectrum. But actually, when these spectra are examined very precisely, the E_n energy levels for a given value of n are found to consist of several closely spaced lines, rather than a single one. That is called the "hydrogen atom fine structure." It means that eigenfunctions that all should have exactly the same energy, don't.

To explain why, the solution of chapter 4.3 must be corrected for a variety of relativistic effects. Before doing so, it is helpful to express the nonrelativistic energy levels of that chapter in terms of the "rest mass energy" $m_e c^2$ of the electron, as follows:

$$\boxed{E_n = -\frac{\alpha^2}{2n^2} m_e c^2 \qquad \text{where } \alpha = \frac{e^2}{4\pi\epsilon_0 \hbar c} \approx \frac{1}{137}} \tag{A.247}$$

The constant α is called the "fine structure constant." It combines the constants $e^2/4\pi\epsilon_0$ from electromagnetism, $\hbar$ from quantum mechanics, and the speed of light c from relativity into one nondimensional number. It is without doubt the single most important number in all of physics, [19].

Nobody knows why it has the value that it has. Still, obviously it is a measurable value, so, following the stated ideas of quantum mechanics, maybe

the universe "measured" this value during its early formation by a process that we may never understand, (since we do not have other measured values for α to deduce any properties of that process from.) If you have a demonstrably better explanation, Sweden awaits you.

In any case, for engineering purposes it is a small number, less than 1%. That makes the hydrogen energy levels really small compared to the rest mass energy of the electron, because they are proportional to the square of α, which is as small as 0.005%. In simple terms, the electron in hydrogen stays well clear of the speed of light.

And that in turn means that the relativistic errors in the hydrogen energy levels are small. Still, even small errors can sometimes be very important. The required corrections are listed below in order of decreasing magnitude.

- *Fine structure.*

  The electron should really be described relativistically using the Dirac equation instead of classically. In classical terms, that will introduce three corrections to the energy levels:
  - Einstein's relativistic correction of the classical kinetic energy $p^2/2m_e$ of the electron.
  - "Spin-orbit interaction", due to the fact that the spin of the moving electron changes the energy levels. The spin of the electron makes it act like a little electromagnet. It can be seen from classical electrodynamics that a moving magnet will interact with the electric field of the nucleus, and that changes the energy levels. Note that the name spin-orbit interaction is a well chosen one, a rarity in physics.
  - There is a third correction for states of zero angular momentum, the Darwin term. It is a crude fix for the fundamental problem that the relativistic wave function is not just a modified classical one, but also involves interaction with the anti-particle of the electron, the positron.

  Fortunately, all three of these effects are very small; they are smaller than the uncorrected energy levels by a factor of order α^2, and the error they introduce is on the order of 0.001%. So the "exact" solution of chapter 4.3 is, by engineering standards, pretty exact after all.

- *Lamb shift.* Relativistically, the electron is affected by virtual photons and virtual electron-positron pairs. It adds a correction of relative magnitude α^3 to the energy levels, one or two orders of magnitude smaller still than the fine structure corrections. To understand the correction properly requires quantum electrodynamics.

- *Hyperfine splitting.* Like the electron, the proton acts as a little electromagnet too. Therefore the energy depends on how it aligns with the

magnetic field generated by the electron. This effect is a factor m_e/m_p smaller still than the fine structure corrections, making the associated energy changes about two orders of magnitude smaller.

Hyperfine splitting couples the spins of proton and electron, and in the ground state, they combine in the singlet state. A slightly higher energy level occurs when they are in a spin-one triplet state; transitions between these states radiate very low energy photons with a wave length of 21 cm. This is the source of the "21 centimeter line" or "hydrogen line" radiation that is of great importance in cosmology. For example, it has been used to analyze the spiral arms of the galaxy, and the hope at the time of this writing is that it can shed light on the so called "dark ages" that the universe went through. Since so little energy is released, the transition is very slow, chapter 7.6.1. It takes on the order of 10 million years, but that is a small time on the scale of the universe.

The message to take away from that is that even errors in the ground state energy of hydrogen that are two million times smaller than the energy itself can be of critical importance under the right conditions.

The following subsections discuss each correction in more detail.

A.39.2 Fine structure

From the Dirac equation, it can be seen that three terms need to be added to the nonrelativistic Hamiltonian of chapter 4.3 to correct the energy levels for relativistic effects. The three terms are worked out in derivation {D.81}. But that mathematics really provides very little insight. It is much more instructive to try to understand the corrections from a more physical point of view.

The first term is relatively easy to understand. Consider Einstein's famous relation $E = mc^2$, where E is energy, m mass, and c the speed of light. According to this relation, the kinetic energy of the electron is not $\frac{1}{2}m_e v^2$, with v the velocity, as Newtonian physics says. Instead it is the difference between the energy $m_{e,v}c^2$ based on the mass $m_{e,v}$ of the electron in motion and the energy $m_e c^2$ based on the mass m_e of the electron at rest. In terms of momentum $p = m_{e,v}v$, chapter 1.1.2,

$$T = m_e c^2 \sqrt{1 + \frac{p^2}{m_e^2 c^2}} - m_e c^2 \tag{A.248}$$

Since the speed of light is large compared to the typical speed of the electron, the square root can be expanded in a Taylor series, [41, 22.12], to give:

$$T \approx \frac{p^2}{2m_e} - \frac{p^4}{8m_e^3 c^2} + \ldots$$

The first term corresponds to the kinetic energy operator used in the nonrelativistic quantum solution of chapter 4.3. (It may be noted that the relativistic momentum $\vec{p}$ is based on the moving mass of the electron, not its rest mass. It is this relativistic momentum that corresponds to the operator $\widehat{\vec{p}} = \hbar\nabla/\mathrm{i}$. So the Hamiltonian used in chapter 4.3 was a bit relativistic already, because in replacing $\vec{p}$ by $\hbar\nabla/\mathrm{i}$, it used the relativistic expression.) The second term in the Taylor series expansion above is the first of the corrections needed to fix up the hydrogen energy levels for relativity. Rewritten in terms of the square of the classical kinetic energy operator, the Bohr ground state energy E_1 and the fine structure constant α, it is

$$H_{1,\text{Einstein}} = -\frac{\alpha^2}{4|E_1|}\left(\frac{\widehat{p}^2}{2m_\text{e}}\right)^2 \tag{A.249}$$

The second correction that must be added to the nonrelativistic Hamiltonian is the so-called "spin-orbit interaction." In classical terms, it is due to the spin of the electron, which makes it into a "magnetic dipole." Think of it as a magnet of infinitesimally small size, but with infinitely strong north and south poles to make up for it. The product of the infinitesimal vector from south to north pole times the infinite strength of the poles is finite, and defines the magnetic dipole moment $\vec{\mu}$. By itself, it is quite inconsequential since the magnetic dipole does not interact directly with the electric field of the nucleus. However, *moving* magnetic poles create an electric field just like the moving electric charges in an electromagnet create a magnetic field. The electric fields generated by the moving magnetic poles of the electron are opposite in strength, but not quite centered at the same position. Therefore they correspond to a motion-induced *electric* dipole. And an electric dipole does interact with the electric field of the nucleus; it wants to align itself with it. That is just like the magnetic dipole wanted to align itself with the external magnetic field in the Zeeman effect.

So how big is this effect? Well, the energy of an electric dipole $\vec{\wp}$ in an electric field $\vec{\mathcal{E}}$ is

$$E_{1,\text{spin-orbit}} = -\vec{\wp}\cdot\vec{\mathcal{E}}$$

As you might guess, the electric dipole generated by the magnetic poles of the moving electron is proportional to the speed of the electron $\vec{v}$ and its magnetic dipole moment $\vec{\mu}$. More precisely, the electric dipole moment $\vec{\wp}$ will be proportional to $\vec{v}\times\vec{\mu}$ because if the vector connecting the south and north poles is parallel to the motion, you do not have two neighboring currents of magnetic poles, but a single current of both negative and positive poles that completely cancel each other out. Also, the electric field $\vec{\mathcal{E}}$ of the nucleus is minus the gradient of its potential $e/4\pi\epsilon_0 r$, so

$$E_{1,\text{spin-orbit}} \propto (\vec{v}\times\vec{\mu})\cdot\frac{e}{4\pi\epsilon_0 r^3}\vec{r}$$

Now the order of the vectors in this triple product can be changed, and the dipole strength $\vec{\mu}$ of the electron equals its spin $\vec{S}$ times the charge per unit mass $-e/m_{\rm e}$, so

$$E_{1,\text{spin-orbit}} \propto \frac{e^2}{m_{\rm e} 4\pi\epsilon_0 r^3} (\vec{r} \times \vec{v}) \cdot \vec{S}$$

The expression between the parentheses is the angular momentum $\vec{L}$ save for the electron mass. The constant of proportionality is worked out in derivation {D.82}, giving the spin-orbit Hamiltonian as

$$H_{1,\text{spin-orbit}} = \alpha^2 |E_1| \left(\frac{a_0}{r}\right)^3 \frac{1}{\hbar^2} \widehat{\vec{L}} \cdot \widehat{\vec{S}} \tag{A.250}$$

The final correction that must be added to the nonrelativistic Hamiltonian is the so-called "Darwin term:"

$$H_{1,\text{Darwin}} = \alpha^2 |E_1| \, \pi a_0^3 \delta^3(\vec{r}) \tag{A.251}$$

According to its derivation in {D.81}, it is a crude fix-up for an interaction with a virtual positron that simply cannot be included correctly in a nonrelativistic analysis.

If that is not very satisfactory, the following much more detailed derivation can be found on the web. It *does* succeed in explaining the Darwin term fully within the nonrelativistic picture alone. First assume that the electric potential of the nucleus does not really become infinite as $1/r$ at $r = 0$, but is smoothed out over some finite nuclear size. Also assume that the electron does not "see" this potential sharply, but perceives of its features a bit vaguely, as diffused out symmetrically over a typical distance equal to the so-called Compton wave length $\hbar/m_{\rm e}c$. There are several plausible reasons why it might: (1) the electron has illegally picked up a chunk of a negative rest mass state, and it is trembling with fear that the uncertainty in energy will be noted, moving rapidly back and forwards over a Compton wave length in a so-called "Zitterbewegung"; (2) the electron has decided to move at the speed of light, which is quite possible nonrelativistically, so its uncertainty in position is of the order of the Compton wave length, and it just cannot figure out where the right potential is with all that uncertainty in position and light that fails to reach it; (3) the electron needs glasses. Further assume that the Compton wave length is much smaller than the size over which the nuclear potential is smoothed out. In that case, the potential within a Compton wave length can be approximated by a second order Taylor series, and the diffusion of it over the Compton wave length will produce an error proportional to the Laplacian of the potential (the only fully symmetric combination of derivatives in the second order Taylor series.). Now if the potential is smoothed over the nuclear region, its Laplacian, giving the charge density, is known to produce a nonzero spike only within that smoothed nuclear region, figure 13.7 or (13.30). Since the nuclear size is small compared

to the electron wave functions, that spike can then be approximated as a delta function. Tell all your friends you heard it here first.

The key question is now what are the changes in the hydrogen energy levels due to the three perturbations discussed above. That can be answered by perturbation theory as soon as the good eigenfunctions have been identified. Recall that the usual hydrogen energy eigenfunctions $\psi_{nlm}\updownarrow$ are made unique by the square angular momentum operator $\widehat{L}^2$, giving l, the z angular momentum operator $\widehat{L}_z$, giving m, and the spin angular momentum operator $\widehat{S}_z$ giving the spin quantum number $m_s = \pm\frac{1}{2}$ for spin up, respectively down. The decisive term whether these are good or not is the spin-orbit interaction. If the inner product in it is written out, it is

$$H_{1,\text{spin-orbit}} = \alpha^2 |E_1| \left(\frac{a_0}{r}\right)^3 \frac{1}{\hbar^2} \left(\widehat{L}_x \widehat{S}_x + \widehat{L}_y \widehat{S}_y + \widehat{L}_z \widehat{S}_z\right)$$

The radial factor is no problem; it commutes with every orbital angular momentum component, since these are purely angular derivatives, chapter 4.2.2. It also commutes with every component of spin because all spatial functions and operators do, chapter 5.5.3. As far as the dot product is concerned, it commutes with $\widehat{L}^2$ since all the components of $\widehat{\vec{L}}$ do, chapter 4.5.4, and since all the components of $\widehat{\vec{S}}$ commute with any spatial operator. But unfortunately, $\widehat{L}_x$ and $\widehat{L}_y$ do not commute with $\widehat{L}_z$, and $\widehat{S}_x$ and $\widehat{S}_y$ do not commute with $\widehat{S}_z$ (chapters 4.5.4 and 5.5.3):

$$[\widehat{L}_x, \widehat{L}_z] = -\mathrm{i}\hbar \widehat{L}_y \quad [\widehat{L}_y, \widehat{L}_z] = \mathrm{i}\hbar \widehat{L}_x \quad [\widehat{S}_x, \widehat{S}_z] = -\mathrm{i}\hbar \widehat{S}_y \quad [\widehat{S}_y, \widehat{S}_z] = \mathrm{i}\hbar \widehat{S}_x$$

The quantum numbers m and m_s are bad.

Fortunately, $\widehat{\vec{L}} \cdot \widehat{\vec{S}}$ does commute with the net z angular momentum $\widehat{J}_z$, defined as $\widehat{L}_z + \widehat{S}_z$. Indeed, using the commutators above and the rules of chapter 4.5.4 to take apart commutators:

$$[\widehat{L}_x \widehat{S}_x, \widehat{L}_z + \widehat{S}_z] = [\widehat{L}_x, \widehat{L}_z]\widehat{S}_x + \widehat{L}_x[\widehat{S}_x, \widehat{S}_z] = -\mathrm{i}\hbar \widehat{L}_y \widehat{S}_x - \mathrm{i}\hbar \widehat{L}_x \widehat{S}_y$$

$$[\widehat{L}_y \widehat{S}_y, \widehat{L}_z + \widehat{S}_z] = [\widehat{L}_y, \widehat{L}_z]\widehat{S}_y + \widehat{L}_y[\widehat{S}_y, \widehat{S}_z] = \mathrm{i}\hbar \widehat{L}_x \widehat{S}_y + \mathrm{i}\hbar \widehat{L}_y \widehat{S}_x$$

$$[\widehat{L}_z \widehat{S}_z, \widehat{L}_z + \widehat{S}_z] = [\widehat{L}_z, \widehat{L}_z]\widehat{S}_z + \widehat{L}_z[\widehat{S}_z, \widehat{S}_z] = 0$$

and adding it all up, you get $[\widehat{\vec{L}} \cdot \widehat{\vec{S}}, \widehat{J}_z] = 0$. The same way of course $\widehat{\vec{L}} \cdot \widehat{\vec{S}}$ commutes with the other components of net angular momentum $\widehat{\vec{J}}$, since the z-axis is arbitrary. And if $\widehat{\vec{L}} \cdot \widehat{\vec{S}}$ commutes with every component of $\widehat{\vec{J}}$, then it commutes with their sum of squares $\widehat{J}^2$. So, eigenfunctions of $\widehat{L}^2$, $\widehat{J}^2$, and $\widehat{J}_z$ *are* good eigenfunctions.

Such good eigenfunctions can be constructed from the $\psi_{nlm}\updownarrow$ by forming linear combinations of them that combine different m and m_s values. The

coefficients of these good combinations are called Clebsch-Gordan coefficients and are shown for $l = 1$ and $l = 2$ in figure 12.5. Note from this figure that the quantum number j of net square momentum can only equal $l + \frac{1}{2}$ or $l - \frac{1}{2}$. The half unit of electron spin is not big enough to change the quantum number of square orbital momentum by more than half a unit. For the rest, however, the detailed form of the good eigenfunctions is of no interest here. They will just be indicated in ket notation as $|nljm_j\rangle$, indicating that they have unperturbed energy E_n, square orbital angular momentum $l(l+1)\hbar^2$, square net (orbital plus spin) angular momentum $j(j+1)\hbar^2$, and net z angular momentum $m_j\hbar$.

As far as the other two contributions to the fine structure are concerned, according to chapter 4.3.1 $\widehat{\vec{p}}^{\,2}$ in the Einstein term consists of radial functions and radial derivatives plus $\widehat{L}^2$. These commute with the angular derivatives that make up the components of $\widehat{\vec{L}}$, and as spatial functions and operators, they commute with the components of spin. So the Einstein Hamiltonian commutes with all components of $\widehat{\vec{L}}$ and $\widehat{\vec{J}} = \widehat{\vec{L}} + \widehat{\vec{S}}$, hence with $\widehat{L}^2$, $\widehat{J}^2$, and $\widehat{J}_z$. And the delta function in the Darwin term can be assumed to be the limit of a purely radial function and commutes in the same way. The eigenfunctions $|nljm_j\rangle$ with given values of l, j, and m_j are good ones for the entire fine structure Hamiltonian.

To get the energy changes, the Hamiltonian perturbation coefficients

$$\langle m_j jln|H_{1,\text{Einstein}} + H_{1,\text{spin-orbit}} + H_{1,\text{Darwin}}|nljm_j\rangle$$

must be found. Starting with the Einstein term, it is

$$\langle m_j jln|H_{1,\text{Einstein}}|nljm_j\rangle = -\frac{\alpha^2}{4|E_1|}\langle m_j jln|\frac{\widehat{p}^4}{4m_{\text{e}}^2}|nljm_j\rangle$$

Unlike what you may have read elsewhere, $\widehat{p}^4$ is indeed a Hermitian operator, but $\widehat{p}^4|nljm_j\rangle$ may have a delta function at the origin, (13.30), so watch it with blindly applying mathematical manipulations to it. The trick is to take half of it to the other side of the inner product, and then use the fact that the eigenfunctions satisfy the nonrelativistic energy eigenvalue problem:

$$\begin{aligned}\langle m_j jln|\frac{\widehat{p}^2}{2m_{\text{e}}}\Big|\frac{\widehat{p}^2}{2m_{\text{e}}}|nljm_j\rangle &= \langle m_j jln|E_n - V\Big|E_n - V|nljm_j\rangle \\ &= \langle m_j jln|E_n^2 - 2VE_n + V^2|nljm_j\rangle\end{aligned}$$

Noting from chapter 4.3 that $E_n = E_1/n^2$, $V = 2E_1 a_0/r$ and that the expectation values of a_0/r and $(a_0/r)^2$ are given in derivation {D.83}, you find that

$$\langle m_j jln|H_{1,\text{Einstein}}|nljm_j\rangle = -\frac{\alpha^2}{4n^2}\left(\frac{4n}{l+\frac{1}{2}} - 3\right)|E_n|$$

The spin-orbit energy correction is

$$\langle m_j jln|H_{1,\text{spin-orbit}}|nljm_j\rangle = \alpha^2|E_1|\langle m_j jln|\left(\frac{a_0}{r}\right)^3 \frac{1}{\hbar^2}\widehat{\vec{L}}\cdot\widehat{\vec{S}}|nljm_j\rangle$$

For states with no orbital angular momentum, all components of $\widehat{\vec{L}}$ produce zero, so there is no contribution. Otherwise, the dot product $\widehat{\vec{L}}\cdot\widehat{\vec{S}}$ can be rewritten by expanding

$$\widehat{J}^2 = (\widehat{\vec{L}}+\widehat{\vec{S}})^2 = \widehat{L}^2 + \widehat{S}^2 + 2\widehat{\vec{L}}\cdot\widehat{\vec{S}}$$

to give

$$\begin{aligned}\widehat{\vec{L}}\cdot\widehat{\vec{S}}|nljm_j\rangle &= \tfrac{1}{2}\Big(\widehat{J}^2 - \widehat{L}^2 - \widehat{S}^2\Big)|nljm_j\rangle \\ &= \tfrac{1}{2}\hbar^2\Big(j(j+1) - l(l+1) - \tfrac{1}{2}(1+\tfrac{1}{2})\Big)|nljm_j\rangle\end{aligned}$$

That leaves only the expectation value of $(a_0/r)^3$ to be determined, and that can be found in derivation {D.83}. The net result is

$$\langle m_j jln|H_{1,\text{spin-orbit}}|nljm_j\rangle = \frac{\alpha^2}{4n^2}2n\frac{j(j+1)-l(l+1)-\frac{1}{2}(1+\frac{1}{2})}{l(l+\frac{1}{2})(l+1)}|E_n| \quad \text{if } l\neq 0$$

or zero if $l = 0$.

Finally the Darwin term,

$$\langle m_j jln|H_{1,\text{Darwin}}|nljm_j\rangle = \alpha^2|E_1|\,\pi a_0^3\langle m_j jln|\delta^3(\vec{r})|nljm_j\rangle$$

Now a delta function at the origin has the property to pick out the value at the origin of whatever function it is in an integral with, compare chapter 7.9.1. Derivation {D.15}, (D.9), implies that the value of the wave functions at the origin is zero unless $l = 0$, and then the value is given in (D.10). So the Darwin contribution becomes

$$\langle m_j jln|H_{1,\text{Darwin}}|nljm_j\rangle = \frac{\alpha^2}{4n^2}4n|E_n| \qquad \text{if } l = 0$$

To get the total energy change due to fine structure, the three contributions must be added together. For $l = 0$, add the Einstein and Darwin terms. For $l \neq 0$, add the Einstein and spin-orbit terms; you will need to do the two possibilities that $j = l + \frac{1}{2}$ and $j = l - \frac{1}{2}$ separately. All three produce the same final result, anyway:

$$\boxed{E_{nljm_j,1} = -\left(\frac{1}{n(j+\frac{1}{2})} - \frac{3}{4}\frac{1}{n^2}\right)\alpha^2|E_n|} \tag{A.252}$$

Since $j + \frac{1}{2}$ is at most n, the energy change due to fine structure is always negative. And it is the biggest fraction of E_n for $j = \frac{1}{2}$ and $n = 2$, where it is $-\frac{5}{16}\alpha^2|E_n|$, still no more than a sixth of a percent of a percent change in energy.

In the ground state j can only be one half, (the electron spin), so the ground state energy does not split into two due to fine structure. You would of course not expect so, because in empty space, both spin directions are equivalent. The ground state does show the largest absolute change in energy.

Woof.

A.39.3 Weak and intermediate Zeeman effect

The weak Zeeman effect is the effect of a magnetic field that is sufficiently weak that it leaves the fine structure energy eigenfunctions almost unchanged. The Zeeman effect is then a small perturbation on a problem in which the "unperturbed" (by the Zeeman effect) eigenfunctions $|nljm_j\rangle$ derived in the previous subsection are degenerate with respect to l and m_j.

The Zeeman Hamiltonian

$$H_1 = \frac{e}{2m_{\rm e}}\mathcal{B}_{\rm ext}\left(\widehat{L}_z + 2\widehat{S}_z\right)$$

commutes with both $\widehat{L}^2$ and $\widehat{J}_z = \widehat{S}_z + \widehat{L}_z$, so the eigenfunctions $|nljm_j\rangle$ are good. Therefore, the energy perturbations can be found as

$$\frac{e}{2m_{\rm e}}\mathcal{B}_{\rm ext}\langle m_j jln|\widehat{L}_z + 2\widehat{S}_z|nljm_j\rangle$$

To evaluate this rigorously would require that the $|nljm_j\rangle$ state be converted into the one or two $\psi_{nlm}{\uparrow\downarrow}$ states with $-l \leqslant m = m_j \pm \frac{1}{2} \leqslant l$ and $m_s = \mp\frac{1}{2}$ using the appropriate Clebsch-Gordan coefficients from figure 12.5.

However, the following simplistic derivation is usually given instead, including in this book. First get rid of L_z by replacing it by $\widehat{J}_z - \widehat{S}_z$. The inner product with $\widehat{J}_z$ can then be evaluated as being $m_j\hbar$, giving the energy change as

$$\frac{e}{2m_{\rm e}}\mathcal{B}_{\rm ext}\left[m_j\hbar + \langle m_j jln|\widehat{S}_z|nljm_j\rangle\right]$$

For the final inner product, make a semi-classical argument that only the component of $\widehat{\vec{S}}$ in the direction of $\vec{J}$ gives a contribution. Don't worry that $\vec{J}$ does not exist. Just note that the component in the direction of $\vec{J}$ is constrained by the requirement that $\widehat{\vec{L}}$ and $\widehat{\vec{S}}$ must add up to $\widehat{\vec{J}}$, but the component normal to $\vec{J}$ can be in any direction and presumably averages out to zero. Dismissing this component, the component in the direction of $\vec{J}$ is

$$\widehat{\vec{S}}_J = \frac{1}{J^2}(\widehat{\vec{S}}\cdot\widehat{\vec{J}})\widehat{\vec{J}}$$

and the dot product in it can be found from expanding

$$\widehat{L}^2 = \widehat{\vec{L}} \cdot \widehat{\vec{L}} = (\widehat{\vec{J}} - \widehat{\vec{S}}).(\widehat{\vec{J}} - \widehat{\vec{S}}) = J^2 - 2\widehat{\vec{J}} \cdot \widehat{\vec{S}} + S^2$$

to give

$$\widehat{\vec{S}}_J = \frac{J^2 - \widehat{L}^2 + S^2}{2J^2} \widehat{\vec{J}}$$

For a given eigenfunction $|nljm_j\rangle$, $J^2 = \hbar^2 j(j+1)$, $\widehat{L}^2 = \hbar^2 l(l+1)$, and $S^2 = \hbar^2 s(s+1)$ with $s = \frac{1}{2}$.

If the z-component of $\widehat{\vec{S}}_J$ is substituted for $\widehat{S}_z$ in the expression for the Hamiltonian perturbation coefficients, the energy changes are

$$\left[1 + \frac{j(j+1) - l(l+1) + s(s+1)}{2j(j+1)}\right] \frac{e\hbar}{2m_e} \mathcal{B}_{\text{ext}} m_j \tag{A.253}$$

(Rigorous analysis using figure 12.5, or more generally item 2 in chapter 12.8, produces the same results.) The factor within the brackets is called the "Landé g-factor." It is the factor by which the magnetic moment of the electron in the atom is larger than for a classical particle with the same charge and total angular momentum. It generalizes the g-factor of the electron in isolation to include the effect of orbital angular momentum. Note that it equals 2, the Dirac g-factor, if there is no orbital momentum, and 1, the classical value, if the orbital momentum is so large that the half unit of spin can be ignored.

In the intermediate Zeeman effect, the fine structure and Zeeman effects are comparable in size. The dominant perturbation Hamiltonian is now the combination of the fine structure and Zeeman ones. Since the Zeeman part does not commute with $\widehat{J}^2$, the eigenfunctions $|nljm_j\rangle$ are no longer good. Eigenfunctions with the same values of l and m_j, but different values of j must be combined into good combinations. For example, if you look at $n = 2$, the eigenfunctions $\left|21\frac{3}{2}\frac{1}{2}\right\rangle$ and $\left|21\frac{1}{2}\frac{1}{2}\right\rangle$ have the same unperturbed energy and good quantum numbers l and m_j. You will have to write a two by two matrix of Hamiltonian perturbation coefficients for them, as in addendum {A.38.3}, to find the good combinations and their energy changes. And the same for the $\left|21\frac{3}{2}{-}\frac{1}{2}\right\rangle$ and $\left|21\frac{1}{2}{-}\frac{1}{2}\right\rangle$ eigenfunctions. To obtain the matrix coefficients, use the Clebsch-Gordan coefficients from figure 12.5 to evaluate the effect of the Zeeman part. The fine structure contributions to the matrices are given by (A.252) when the j values are equal, and zero otherwise. This can be seen from the fact that the energy changes must be the fine structure ones when there is no magnetic field; note that j is a good quantum number for the fine structure part, so its perturbation coefficients involving different j values are zero.

A.39.4 Lamb shift

A famous experiment by Lamb & Retherford in 1947 showed that the hydrogen atom state $n = 2$, $l = 0$, $j = \frac{1}{2}$, also called the $2S_{1/2}$ state, has a somewhat

different energy than the state $n = 2$, $l = 1$, $j = \frac{1}{2}$, also called the $2P_{1/2}$ state. That was unexpected, because even allowing for the relativistic fine structure correction, states with the same principal quantum number n and same total angular momentum quantum number j should have the same energy. The difference in orbital angular momentum quantum number l should not affect the energy.

The cause of the unexpected energy difference is called Lamb shift. To explain why it occurs would require quantum electrodynamics, and that is well beyond the scope of this book. Roughly speaking, the effect is due to a variety of interactions with virtual photons and electron/positron pairs. A good qualitative discussion on a nontechnical level is given by Feynman [19].

Here it must suffice to list the approximate energy corrections involved. For states with zero orbital angular momentum, the energy change due to Lamb shift is

$$E_{\vec{n},1,\text{Lamb}} = -\frac{\alpha^3}{2n} k(n,0) E_n \qquad \text{if } l = 0 \tag{A.254}$$

where $k(n,0)$ is a numerical factor that varies a bit with n from about 12.7 to 13.2. For states with nonzero orbital angular momentum,

$$E_{\vec{n},1,\text{Lamb}} = -\frac{\alpha^3}{2n} \left[k(n,l) \pm \frac{1}{\pi(j+\frac{1}{2})(l+\frac{1}{2})} \right] E_n \qquad \text{if } l \neq 0 \text{ and } j = l \pm \tfrac{1}{2} \tag{A.255}$$

where $k(n,l)$ is less than 0.05 and varies somewhat with n and l.

It follows that the energy change is really small for states with nonzero orbital angular momentum, which includes the $2P_{1/2}$ state. The change is biggest for the $2S_{1/2}$ state, the other state in the Lamb & Retherford experiment. (True, the correction would be bigger still for the ground state $n = 1$, but since there are no states with nonzero angular momentum in the ground state, there is no splitting of spectral lines involved there.)

Qualitatively, the reason that the Lamb shift is small for states with nonzero angular momentum has to do with distance from the nucleus. The nontrivial effects of the cloud of virtual particles around the electron are most pronounced in the strong electric field very close to the nucleus. In states of nonzero angular momentum, the wave function is zero at the nucleus, (D.9). So in those states the electron is unlikely to be found very close to the nucleus. In states of zero angular momentum, the square magnitude of the wave function is $1/n^3\pi a_0^3$ at the nucleus, reflected in both the much larger Lamb shift as well as its approximate $1/n^3$ dependence on the principal quantum number n.

A.39.5 Hyperfine splitting

Hyperfine splitting of the hydrogen atom energy levels is due to the fact that the nucleus acts as a little magnet just like the electron. The single-proton nucleus

and electron have magnetic dipole moments due to their spin equal to

$$\vec{\mu}_{\rm p} = \frac{g_p e}{2m_{\rm p}} \widehat{\vec{S}}_{\rm p} \qquad \vec{\mu}_{\rm e} = -\frac{g_e e}{2m_{\rm e}} \widehat{\vec{S}}_{\rm e}$$

in which the g-factor of the proton is about 5.59 and that of the electron 2. The magnetic moment of the nucleus is much less than the one of the electron, since the much greater proton mass appears in the denominator. That makes the energy changes associated with hyperfine splitting really small compared to other effects such as fine structure.

This discussion will restrict itself to the ground state, which is by far the most important case. For the ground state, there is no orbital contribution to the magnetic field of the electron. There is only a "spin-spin coupling" between the magnetic moments of the electron and proton, The energy involved can be thought of most simply as the energy $-\vec{\mu}_{\rm e} \cdot \vec{\mathcal{B}}_p$ of the electron in the magnetic field $\vec{\mathcal{B}}_{\rm p}$ of the nucleus. If the nucleus is modelled as an infinitesimally small electromagnet, its magnetic field is that of an ideal current dipole as given in table 13.2. The perturbation Hamiltonian then becomes

$$H_{1,\text{spin-spin}} = \frac{g_{\rm p} g_{\rm e} e^2}{4 m_{\rm e} m_{\rm p} \epsilon_0 c^2} \left[\frac{3(\widehat{\vec{S}}_{\rm p} \cdot \vec{r})(\widehat{\vec{S}}_{\rm e} \cdot \vec{r}) - (\widehat{\vec{S}}_{\rm p} \cdot \widehat{\vec{S}}_{\rm e}) r^2}{4\pi r^5} + \frac{2(\widehat{\vec{S}}_{\rm p} \cdot \widehat{\vec{S}}_{\rm e})}{3} \delta^3(\vec{r}) \right]$$

The good states are not immediately self-evident, so the four unperturbed ground states will just be taken to be the ones which the electron and proton spins combine into the triplet or singlet states of chapter 5.5.6:

$$\text{triplet:} \quad \psi_{100}|1\ 1\rangle \quad \psi_{100}|1\ 0\rangle \quad \psi_{100}|1\ {-1}\rangle \qquad \text{singlet:} \quad \psi_{100}|0\ 0\rangle$$

or $\psi_{100}|s_{\rm net} m_{\rm net}\rangle$ for short, where $s_{\rm net}$ and $m_{\rm net}$ are the quantum numbers of net spin and its z-component. The next step is to evaluate the four by four matrix of Hamiltonian perturbation coefficients

$$\langle \underline{m}_{\rm net} \underline{s}_{\rm net} | \psi_{100} | H_{1,\text{spin-spin}} \psi_{100} | s_{\rm net} m_{\rm net} \rangle$$

using these states.

Now the first term in the spin-spin Hamiltonian does not produce a contribution to the perturbation coefficients. The reason is that the inner product of the perturbation coefficients written in spherical coordinates involves an integration over the surfaces of constant r. The ground state eigenfunction ψ_{100} is constant on these surfaces. So there will be terms like $3\widehat{S}_{{\rm p},x}\widehat{S}_{{\rm e},y} xy$ in the integration, and those are zero because x is just as much negative as positive on these spherical surfaces, (as is y). There will also be terms like $3\widehat{S}_{{\rm p},x}\widehat{S}_{{\rm e},x} x^2 - \widehat{S}_{{\rm p},x}\widehat{S}_{{\rm e},x} r^2$ in the integration. These will be zero too because by symmetry the averages of x^2, y^2, and z^2 are equal on the spherical surfaces, each equal to one third the average of r^2.

So only the second term in the Hamiltonian survives, and the Hamiltonian perturbation coefficients become

$$\frac{g_{\rm p} g_{\rm e} e^2}{6 m_{\rm e} m_{\rm p} \epsilon_0 c^2} \langle \underline{m}_{\rm net} \underline{s}_{\rm net} | \psi_{100} | (\widehat{\vec{S}}_{\rm p} \cdot \widehat{\vec{S}}_{\rm e}) \delta^3(\vec{r}) \psi_{100} | s_{\rm net} m_{\rm net} \rangle$$

The spatial integration in this inner product merely picks out the value $\psi_{100}^2(0) = 1/\pi a_0^3$ at the origin, as delta functions do. That leaves the sum over the spin states. According to addendum {A.10},

$$\text{triplet:} \quad \widehat{\vec{S}}_{\rm p} \cdot \widehat{\vec{S}}_{\rm e} \, |1 \, m_{\rm net}\rangle = \tfrac{1}{4}\hbar^2 |1 \, m_{\rm net}\rangle \qquad \text{singlet:} \quad \widehat{\vec{S}}_{\rm p} \cdot \widehat{\vec{S}}_{\rm e} \, |0 \, 0\rangle = -\tfrac{3}{4}\hbar^2 |0 \, 0\rangle$$

Since the triplet and singlet spin states are orthonormal, only the Hamiltonian perturbation coefficients for which $\underline{s}_{\rm net} = s_{\rm net}$ and $\underline{m}_{\rm net} = m_{\rm net}$ survive, and these then give the leading order changes in the energy.

Plugging it all in and rewriting in terms of the Bohr energy and fine structure constant, the energy changes are:

$$\text{triplet: } E_{1,\text{spin-spin}} = \tfrac{1}{3} g_{\rm p} g_{\rm e} \frac{m_{\rm e}}{m_{\rm p}} \alpha^2 |E_1| \qquad \text{singlet: } E_{1,\text{spin-spin}} = -g_{\rm p} g_{\rm e} \frac{m_{\rm e}}{m_{\rm p}} \alpha^2 |E_1| \tag{A.256}$$

The energy of the triplet states is raised and that of the singlet state is lowered. Therefore, in the true ground state, the electron and proton spins combine into the singlet state. If they somehow get kicked into a triplet state, they will eventually transition back to the ground state, say after 10 million years or so, and release a photon. Since the difference between the two energies is so tiny on account of the very small values of both α^2 and $m_{\rm e}/m_{\rm p}$, this will be a very low energy photon. Its wave length is as long as 0.21 m, producing the 21 cm hydrogen line.

A.40 Deuteron wave function

This addendum examines the form of the wave function of the deuteron. It assumes that the deuteron can be described as a two particle system; a proton and a neutron. In reality, both the proton and the neutron consist of three quarks. So the deuteron is really a six particle system. That will be ignored here.

Then the deuteron wave function is a function of the positions and spin angular momenta of the proton and neutron. That however can be simplified considerably. First of all, it helps if the center of gravity of the deuteron is taken as origin of the coordinate system. In that coordinate system, the individual positions of proton and neutron are no longer important. The only quantity that is important is the position vector going from neutron to proton, {A.5}:

$$\vec{r} \equiv \vec{r}_{\rm p} - \vec{r}_{\rm n}$$

That represents the relative position of the proton relative to the neutron.

Consider now the spin angular momenta of proton and neutron. The two have spin angular momenta of the same magnitude. The corresponding quantum number, called the "spin" for short, equals $s_{\rm p} = s_{\rm n} = \frac{1}{2}$. However, the proton and neutron can still have different spin angular momentum along whatever is chosen to be the z-axis. In particular, each can have a spin S_z along the z-axis that is either $\frac{1}{2}\hbar$ or $-\frac{1}{2}\hbar$.

All together it means that the deuteron wave function depends nontrivially on both the nucleon spacing and the spin components in the z-direction:

$$\psi = \psi\left(\vec{r}, S_{z,\rm p}, S_{z,\rm n}\right)$$

The square magnitude of this wave function gives the probability density to find the nucleons at a given spacing $\vec{r}$ and with given spin values along the z-axis.

It is solidly established by experiments that the wave function of the deuteron has net nuclear spin $j_{\rm N} = 1$ and even parity. The question to be examined now is what that means for the orbital angular momentum and the spins of the proton and neutron. To answer that, the wave function needs to be written in terms of states that have definite combined orbital angular momentum and definite combined spin.

The conditions for a state to have definite orbital angular momentum were discussed in chapter 4.2. The angular dependence of the state must be given by a spherical harmonic $Y_l^m(\theta,\phi)$. Here θ and ϕ are the angles that the vector $\vec{r}$ makes with the axes of the chosen spherical coordinate system. The azimuthal quantum number l describes the magnitude of the orbital angular momentum. In particular, the magnitude of the orbital momentum is $\sqrt{l(l+1)}\hbar$. The magnetic quantum number m describes the component of the orbital angular momentum along the chosen z-axis. In particular, that component equals $m\hbar$. Both $l \geqslant 0$ and $|m| \leqslant l$ must be integers.

As far as the combined spin angular momentum is concerned, the possibilities were discussed in chapter 5.5.6 and in more detail in chapter 12. First, the proton and neutron spins can cancel each other perfectly, producing a state of zero net spin. This state is called the "singlet" state. Zero net spin has a corresponding quantum number $s = 0$. And since the component of the angular momentum along any chosen z-axis can only be zero, so is the spin magnetic quantum number $m_s = 0$.

The second possibility is that the proton and neutron align their spins in parallel, crudely speaking. More precisely, the combined spin has a magnitude given by quantum number $s = \frac{1}{2}+\frac{1}{2} = 1$. The combined spin angular momentum along the chosen z direction is $m_s\hbar$ where m_s can be -1, 0, or 1.

The wave function of the deuteron can be written as a combination of the above states of orbital and spin angular momentum. It then takes the generic

form:

$$\psi = \sum_{nlmsm_s} c_{nlmsm_s} R_n(|\vec{r}|) Y_l^m(\theta,\phi)|s\, m_s\rangle \tag{A.257}$$

Here the c_{nlmsm_s} are constants. The functions R_n are not of particular interest here; any complete set of orthonormal radial functions will do. Note that the individual terms in the sum above are not supposed to be energy eigenfunctions. They are merely chosen states of definite orbital and spin angular momentum. The ket $|s\, m_s\rangle$ is a way of indicating the combined spin state of the two nucleons. It is defined in terms of the separate spins of the proton and neutron in chapter 5.5.6 (5.26).

The above expression for the wave function is quite generally valid for a system of two fermions. But it can be made much more specific based on the mentioned known properties of the deuteron.

The simplest is the fact that the parity of the deuteron is even. Spherical harmonics have odd parity if l is odd, and even if l is even, {D.14}. So there cannot be any odd values of l in the sum above. In other words, the constants c_{nlmsm_s} must be zero for odd l.

Physically, that means that the spatial wave function is symmetric with respect to replacing $\vec{r}$ by $-\vec{r}$. It may be noted that this spatial symmetry and the corresponding even parity are exactly what is expected theoretically. The reasons were explored earlier in {A.8} and {A.9}. The wave function for any given spin state should not change sign, and odd parity cannot meet that requirement. However, it should be noted that the arguments in {A.8} and {A.9} are not valid if the potential includes terms of second or higher order in the momentum. Some more advanced potentials that have been written down include such terms.

The spatial symmetry also means that the wave function is symmetric with respect to swapping the two nucleons. That is because $\vec{r}$ is the vector from neutron to proton, so swapping the two inverts the sign of $\vec{r}$. This does assume that the small difference in mass between the neutron and proton is ignored. Otherwise the swap would change the center of gravity. Recall that the (part of the) wave function considered here is relative to the center of gravity. In any case, the hypothetical wave functions for a bound state of two protons or one of two neutrons would be exactly symmetric under exchange of the positions of the two identical particles.

The condition that the nuclear spin $j_{\rm N} = 1$ is a bit more complex. First a brief review is needed into how angular momenta combine in quantum mechanics. (For a more complete description, see chapter 12.) A state with definite quantum numbers l and s has in general quantum uncertainty in the net nuclear spin $j_{\rm N}$. But the values of $j_{\rm N}$ cannot be completely arbitrary. The only values that can have nonzero probability are in the range

$$|l - s| \leqslant j_{\rm N} \leqslant l + s$$

The key is now that unless a state l, s has a nonzero probability for $j_N = 1$, it cannot appear in the deuteron wave function at all. To verify that, take an inner product of the state with the representation (A.257) of the deuteron wave function. In the left hand side, you get zero because the deuteron wave function has $j_N = 1$ and states of different j_N are orthogonal. In the right hand side, all terms except one drop out because the states in the sum are orthonormal. The one remaining term is the coefficient of the considered state. Then this coefficient must be zero since the left hand side is.

Using the above criterion, consider which states cannot appear in the deuteron wave function. First of all, states with $s = 0$ are according to the inequalities above states of nuclear spin $j_N = l$. That cannot be 1, since l had to be even because of parity. So states with $s = 0$ cannot appear in the deuteron wave function. It follows that the deuteron wave function has a combined nucleon spin $s = 1$ without quantum uncertainty.

Secondly, states with $l \geqslant 4$ have j_N at least equal to 3 according to the above inequalities. So these states cannot appear. That leaves only states with $l = 0$ or 2 and $s = 1$ as possibilities.

Now states with $l = 0$ and $s = 1$ are states with $j_N = 1$. Any such state can appear in the deuteron wave function. To what amount remains unknown. That would only be answerable if an exact solution to the proton-neutron deuteron would be available. But surely, based on arguments like those in {A.8} and {A.9}, it is to be expected that there is a significant $l = 0$ component.

States with $l = 2$ and $s = 1$ are also a possibility. But they cannot appear in arbitrary combinations. Any such state has multiple possible values for j_N in the range from 1 to 3. That uncertainty must be eliminated before the states are acceptable for the deuteron wave function. It turns out that pure $j_N = 1$ states can be obtained by taking specific combinations of states. In particular, groups of states that vary only in the quantum numbers m and m_s can be combined into states with $j_N = 1$. (For the curious, the specific combinations needed can be read off in figure 12.6).

The bottom line is that the deuteron wave function can have uncertainty in the orbital angular momentum. In particular, both orbital angular momentum numbers $l = 0$ and $l = 2$ can and do have a nonzero probability.

A.41 Deuteron model

A very simple model can be used to give some context to the data of the deuteron. This addendum describes that model. Then it discusses the various major problems of the model. Some possible fixes for these problems are indicated.

In all cases it is assumed that the deuteron is modelled as a two-particle system, a proton and a neutron. Furthermore, the proton and neutron are

assumed to have the same properties in the deuteron as they have in free space. These assumptions are not really true. For one, the proton and neutron are not elementary particles but combinations of quarks. However, ignoring that is a reasonable starting point in trying to understand the deuteron.

A.41.1 The model

The deuteron contains two nucleons, a proton and a neutron. The simple model assumes that the potential energy of the deuteron only depends on the distance r between the nucleons. More specifically, it assumes that the potential energy has some constant value $-V_0$ up to some spacing $r = d_0$. And it assumes that the potential is zero for spacings larger than d_0. Figure A.24 shows the idea.

This model is analytically solvable. First, the deuteron involves the motion of two particles, the proton and the neutron. However, the problem may be simplified to that of an imaginary single "reduced mass" encircling the center of gravity of the deuteron, addendum {A.5}.

The reduced mass in the simplified problem is half the mass of the proton or neutron. (That ignores the tiny difference in mass between the proton and neutron.) The potential for the reduced mass is $-V_0$ if the reduced mass is within a distance d_0 of the center of gravity and zero beyond that. A potential of this type is commonly called a [spherical] [square] well potential. Figure A.24 shows the potential in green.

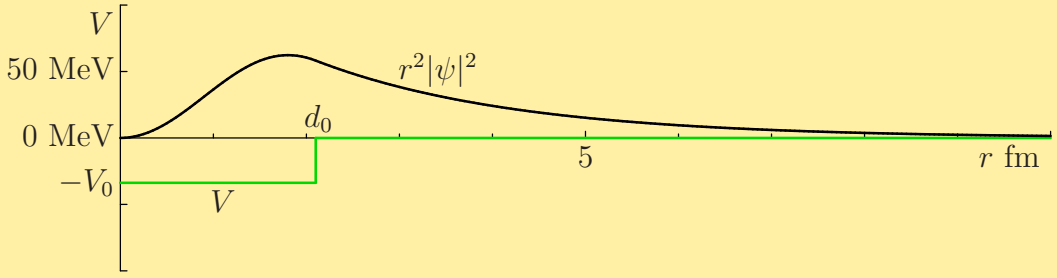


Figure A.24: Crude deuteron model. The potential is in green. The relative probability of finding the nucleons at a given spacing is in black.

The solution for the reduced mass problem may be worked out following addendum {A.6}. Note that the model involves two unknown parameters, the potential V_0 and the distance d_0. Two pieces of experimental information need to be used to fix values for these parameters.

First of all, the binding energy should match the experimental 2.2247 MeV. Also, the root-mean square radial position of the nucleons away from the center of the nucleus should be about 1.955 fm, [J.P. McTavish 1982 J. Phys. G 8 911; J.L. Friar *et al* 1984 Phys. Rev. C 30 1084]. (Based on electron scattering experiments, physicists are confident that the root-mean-square radial position

of the charge from the center of the deuteron is 2.14 fm. However, this "charge radius" is larger than the root mean square radial position of the nucleons. The main reason is that the proton has a finite size. For example, even in the hypothetical case that the distance of the two nucleons from the center of gravity would be zero, there would still be a positive charge radius; the one of the proton. The proton by itself has a significant charge radius, 0.88 fm.) The distance r of the reduced mass from the origin should match the distance between the nucleons; in other words it should be twice the 1.955 fm radius.

d_0	$V_{0,\text{min}}$	V_0	$\langle V \rangle$	$\langle T \rangle$	E	$r_{\text{rms}}/2$	$\langle T \rangle_{\text{min}}$	A_{S}
fm	MeV	MeV	MeV	MeV	MeV	fm	MeV	$1/\sqrt{\text{fm}}$
1.2	71.1	88.5	−21.0	18.8	−2.22	1.77	7.6	0.79
2.1	23.2	33.7	−12.5	10.3	−2.22	1.95	6.2	0.88
3.0	11.4	19.2	−9.2	7.0	−2.22	2.15	5.1	0.98
1.8	31.6	43.0	−13.7	11.7	−2.02	1.96	6.1	0.82
2.1	23.2	33.7	−12.5	10.3	−2.22	1.95	6.2	0.88
2.4	17.8	27.7	−11.7	9.3	−2.42	1.95	6.1	0.94

Table A.4: Deuteron model data. The top half of the table allows some deviation from the experimental nucleon root-mean-square radial position. The bottom half allows some deviation from the experimental energy.

Table A.4 shows that these two experimental constraints are met when the distance d_0 is 2.1 fm and the potential V_0 about 35 MeV. The fact that the distance d_0 matches the charge radius is just a coincidence.

There is some justification for this model. For one, it is well established that the nuclear force very quickly becomes negligible beyond some typical spacing between the nucleons. The above potential reflects that. Based on better models, (in particular the so-called OPEP potential), the typical range of the nuclear force is roughly 1.5 fm. The potential cut-off d_0 in the model is at 2.1 fm. Obviously that is in the ballpark, though it seems a bit big. (For a full-potential/zero-potential cut-off.)

The fact that both the model and exact potentials vanish at large nucleon spacings also reflects in the wave function. It means that the rate of decay of the wave function at large nucleon spacings is correctly represented. The rate of decay depends only on the binding energy E.

To be more precise, the model wave function is, {A.6},

$$\psi = \frac{A_{\text{S}}}{\sqrt{4\pi}} \frac{e^{-\sqrt{2m_{\text{red}}|E|}r/\hbar}}{r} \qquad \text{for} \qquad r > d_0$$

where A_S is some constant. Unlike the model, the experimental wave function has some angular variation. However, if this variation is averaged away, the experimental wave function decays just like the model for distances much larger than 1.5 fm. In addition, the model matches the experimental value for the constant A_S, 0.88, table A.4.

To be fair, this good agreement does not actually support the details of the potential as much as it may seem. As the difference between $-V_0$ and the expectation potential $\langle V \rangle$ in table A.4 shows, the nucleons are more likely to be found beyond the spacing d_0 than below it. And the root mean square separation of the nucleons depends mostly on the wave function at large values of r. As a consequence, if the model gets A_S right, then the root mean square separation of the nucleons cannot be much wrong either. That is true regardless of what exactly the potential for $r < d_0$ is. Still, the model does get the right value.

Another point in favor of the model is that the kinetic energy cannot be all wrong. In particular, the Heisenberg uncertainty relationship implies that the kinetic energy of the deuteron must be at least 6.2 MeV. The second-last column in the table shows the minimum kinetic energy that is possible for the root-mean-square radial nucleon position in the previous column. It follows that unavoidably the kinetic energy is significantly larger than the binding energy. That reflects the fact that the deuteron is only weakly bound. (For comparison, for the proton-electron hydrogen atom the kinetic energy and binding energy are equal.)

The model also supports the fact that there is only a single bound state for the deuteron. The second column in the table gives the smallest value of V_0 for which there is a bound state at all. Clearly, the estimated values of V_0 are comfortably above this minimum. But for a second bound state to exist, the value of V_0 needs to exceed the value in the second column by a factor 4. Obviously, the estimated values get nowhere close to that.

A final redeeming feature of the model is that the deduced potential V_0 is reasonable. In particular, 35 MeV is a typical potential for a nucleon inside a heavy nucleus. It is used as a ballpark in the computation of so-called alpha decay of nuclei, [31, p. 83, 252].

A.41.2 The repulsive core

While the model of the deuteron described in the previous subsection has several redeeming features, it also has some major problems. The problem to be addressed in this subsection is that the nuclear force becomes repulsive when the nucleons try to get too close together. The model does not reflect such a "repulsive core" at all.

A simple fix is to declare nucleon spacings below a certain value $d_{\min}$ to be off-limits. Typically, half a femtometer is used for $d_{\min}$. The potential is taken

to be infinite, rather than $-V_0$, for spacings below $d_{\min}$. That prevents the nucleons from getting closer than half a femtometer together.

d_0	$V_{0,\min}$	V_0	$\langle V\rangle$	$\langle T\rangle$	E	$r_{\rm rms}/2$	$\langle T\rangle_{\min}$	$A_{\rm S}$
fm	MeV	MeV	MeV	MeV	MeV	fm	MeV	$1/\sqrt{\text{fm}}$
0.7	2559.9	2657.3	−119.7	117.5	−2.22	1.75	7.6	0.78
1.7	71.1	88.5	−21.0	18.8	−2.22	1.96	6.1	0.88
2.6	23.2	33.8	−12.5	10.3	−2.22	2.16	5.0	0.99

Table A.5: Deuteron model data with a repulsive core of 0.5 fm.

The modifications needed to the mathematics to include this repulsive core are minor. Table A.5 summarizes the results. The value of V_0 for a second bound state would need to be about 160 MeV.

Note that the value of the potential cut-off distance d_0 has been reduced from 2.1 fm to 1.7 fm. As discussed in the previous subsection, that can be taken to be an improvement. Also, the expectation potential and kinetic energies seem much better. A much more advanced potential, the so-called Argonne v_{18}, gives 22 and 19.8 MeV for these expectation values.

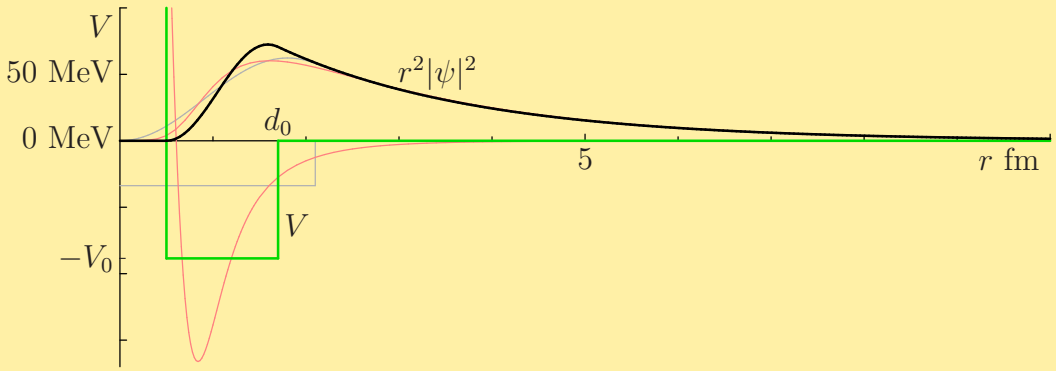


Figure A.25: Crude deuteron model with a 0.5 fm repulsive core. Thin grey lines are the model without the repulsive core. Thin red lines are more or less comparable results from the Argonne v_{18} potential.

Figure A.25 shows the potential and probability density. The previous results without repulsive core are shown as thin grey lines for easier comparison. Note that there are very distinctive differences between the wave functions with and without repulsive core. But astonishingly, the values for the root mean square nucleon separation $r_{\rm rms}$ are virtually identical. The value of $r_{\rm rms}$ is not at all a good quantity to gauge the accuracy of the model.

Figure A.25 also shows corresponding results according to the much more sophisticated Argonne v_{18} model, [51]. The top red line shows the probability density for finding the nucleons at that spacing. The lower curve shows an effective spherical potential. A note of caution is needed here; the true deuteron potential has *very* large deviations from spherical symmetry. So the comparison of potentials is fishy. What is really plotted in figure A.25 is the effective potential that integrated against the probability density produces the correct expectation potential energy.

It is interesting to see from figure A.25 how small the 2.2 MeV binding energy of the deuteron really is, as compared to the minimum value of the potential energy.

A.41.3 Spin dependence

An big problem with the model so far is that nucleon-nucleon interactions depend strongly on the nucleon spins. Such an effect also exists for the proton-electron hydrogen atom, {A.39.5}. However, there the effect is extremely small. For the deuteron, the effect of spin is dramatic.

The proton and neutron each have spin $\frac{1}{2}$. They must align these spins in parallel into a so-called triplet state of combined spin 1, chapter 5.5.6. If instead they align their spins in opposite directions in a singlet state of zero net spin, the deuteron will not bind. The model as given so far does not describe this.

One simple way to fix this up is to write two different potentials. One potential $V_{\rm t}(r)$ is taken to apply if the nucleons are in the triplet state. It can be modeled by the piecewise constant potential as discussed so far. A second potential $V_{\rm s}(r)$ is taken to apply if the nucleons are in the singlet state. A suitable form can be deduced from experiments in which nucleons are scattered off each other. This potential should not allow a bound state.

That leaves only the problem of how to write the complete potential. The complete potential should simplify to $V_{\rm t}$ for the triplet state and to $V_{\rm s}$ for the singlet state. A form that does that is

$$V = V_{\rm t}(r)\left[\frac{3}{4} + \frac{1}{\hbar^2}\widehat{\vec{S}}_{\rm p}\cdot\widehat{\vec{S}}_{\rm n}\right] + V_{\rm s}(r)\left[\frac{1}{4} - \frac{1}{\hbar^2}\widehat{\vec{S}}_{\rm p}\cdot\widehat{\vec{S}}_{\rm n}\right] \tag{A.258}$$

The reason that this works is because the dot product $\widehat{\vec{S}}_{\rm p}\cdot\widehat{\vec{S}}_{\rm n}$ between the proton and neutron spins is $\frac{1}{4}\hbar^2$ in the triplet state and $-\frac{3}{4}\hbar^2$ in the singlet state, {A.10}.

A.41.4 Noncentral force

So far it has been assumed that the potential in a given spin state only depends on the distance r between the nucleons. If true, that would imply that the orbital

angular momentum of the motion of the nucleons is conserved. In terms of classical physics, the forces between the particles would be along the connecting line between the particles. That does not produce a moment that can change orbital angular momentum.

In terms of quantum mechanics, it gets phrased a little differently. A potential that only depends on the distance between the particles commutes with the orbital angular momentum operators. Then so does the Hamiltonian. And that means that the energy states can also be taken to be states of definite orbital angular momentum.

In particular, in the ground state, the proton and neutron should then be in a state of zero orbital angular momentum. Such a state is spherically symmetric. Therefore the proton charge distribution should be spherically symmetric too. All that would be just like for the electron in the hydrogen atom. See chapters 4.2, 4.3, 4.5, and 7.3, addendum {A.39}, and derivations {A.8} and {A.9} for more details on these issues.

However, the fact is that the charge distribution of the deuteron is not quite spherically symmetric. Therefore, the potential cannot just depend on the distance r between proton and neutron. It must also depend on the direction of the vector $\vec{r}$ from neutron to proton. In particular, it must depend on how this vector aligns with the nucleon spins. There are no other directions to compare to in the deuteron besides the spins.

The orientation of the chosen coordinate system should not make a difference for the potential energy. From a classical point of view, there are three nuclear angles that are nontrivial. The first two are the angles that the vector $\vec{r}$ from neutron to proton makes with the neutron and proton spins. The third is the angle between the two spins. These three angles, plus the distance between the neutron and proton, fully determine the geometry of the nucleus.

To check that, imagine a coordinate system with origin at the neutron. Take the x-axis along the connecting line to the proton. Rotate the coordinate system around the x-axis until the neutron spin is in the xy-plane. What determines the geometry in this coordinate system is the angle in the xy-plane between the connecting line and the neutron spin. And the two angles that fix the direction of the proton spin; the one with the connecting line and the one with the neutron spin.

In quantum mechanics, angles involving angular momentum vectors are not well defined. That is due to angular momentum uncertainty, chapter 4.2. However, dot products between vectors can be used to substitute for angles between vectors, for given lengths of the vectors. Because the spin vectors have a given length, there are four parameters that fix the geometry:

$$r \qquad \widehat{\vec{S}}_{\rm n} \cdot \widehat{\vec{S}}_{\rm p} \qquad \widehat{\vec{S}}_{\rm n} \cdot \vec{r} \qquad \widehat{\vec{S}}_{\rm p} \cdot \vec{r}$$

The potential energy should depend on these four parameters. Note that the

effect of the first two parameters was already modelled in the previous subsections.

In order that orbital angular momentum is not conserved, the last two parameters should be involved. But not separately, because they change sign under a parity transformation or time reversal. It is known that to very good approximation, nuclei respect the parity and time-reversal symmetries. Terms quadratic in the last two parameters are needed. And in particular, the product of the last two parameters is needed. If you just square either parameter, you get a trivial multiple of r^2. That can be seen from writing the spins out in terms of the so-called Pauli matrices, as defined in chapter 12.

The bottom line is that the needed additional contribution to the potential is due to the product of the final two terms. This contribution is called the "tensor potential" for reasons that are not important. By convention, the tensor potential is written in the form

$$S_{12}V_{\rm T}(r) \quad \text{with} \quad S_{12} \equiv \frac{4}{\hbar^2}\left(\frac{3}{r^2}(\widehat{\vec{S}}_{\rm n}\cdot\vec{r})(\widehat{\vec{S}}_{\rm p}\cdot\vec{r}) - \widehat{\vec{S}}_{\rm n}\cdot\widehat{\vec{S}}_{\rm p}\right) \tag{A.259}$$

The choice of the symbol S_{12} is another one of those confusion-prone physics conventions. One source uses the same symbol only a few pages away for a matrix element. The division by r^2 is to make S_{12} independent of the distance between the nucleons. This distance is separately accounted for in the factor $V_{\rm T}$. Also the subtraction of the dot product between the spins makes the average of S_{12} over all directions of $\vec{r}$ zero. That means that ΔV only describes the angular variation of the potential. The angular average must be described by other potentials such as the ones written down earlier.

It turns out that S_{12} commutes with the operator of the square net nucleon spin but not with the operator of orbital angular momentum. That is consistent with the fact that the deuteron has definite net nucleon spin $s = 1$ but uncertain orbital angular momentum. Its quantum number of orbital angular momentum can be $l = 0$ or 2.

It should also be noted that S_{12} produces zero when applied on the singlet nucleon spin state. So the term has no effect on singlet states. These properties of S_{12} may be verified by crunching it out using the properties of the Pauli spin matrices, chapter 12.10, including that $\sigma_i^2 = I$ and $\sigma_i\sigma_{\bar{\imath}} = -\sigma_{\bar{\imath}}\sigma_i = \mathrm{i}\sigma_{\bar{\bar{\imath}}}$ with $i, \bar{\imath}, \bar{\bar{\imath}}$ successive numbers in the cyclic sequence $\ldots 1231231 \ldots$.

Figure A.26 illustrates the effects of the uncertainty in orbital angular momentum on the deuteron. The data are taken from the Argonne v_{18} potential, [51].

The black curve is the probability density for finding the nucleons at that spacing r. Most of this probability density is due to the spherically symmetric, $l = 0$, part of the wave function. This contribution is shown as the grey curve labelled 0. The contribution due to the $l = 2$ state is the grey curve labelled 2.

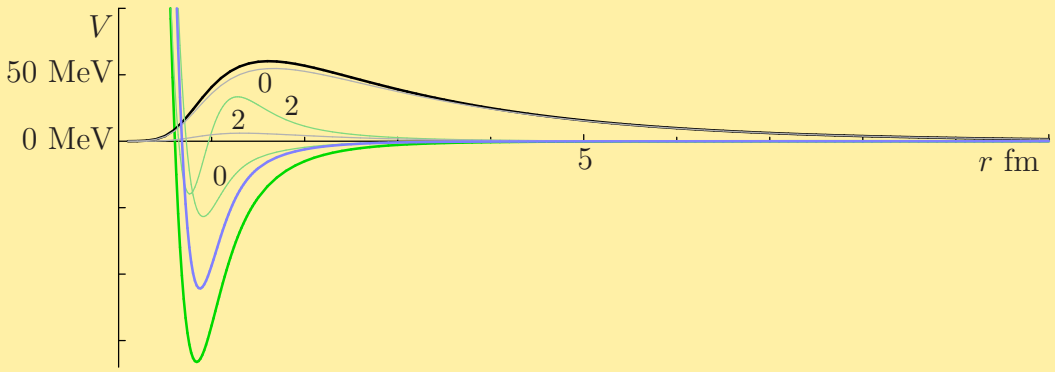


Figure A.26: Effects of uncertainty in orbital angular momentum.

The total probability of the $l = 2$ state is only 5.8% according to the Argonne v_{18} potential.

That might suggest that the effect of the $l = 2$ state on the deuteron binding could be ignored. But that is untrue. If the deuteron wave function was completely spherically symmetric, the potential would given by the thin green curve labeled 0. The binding energy from this potential is significantly less than that of the hypothetical dineutron, shown in blue. And the dineutron is not even bound. If the deuteron was in a pure $l = 2$ state, the binding would be less still according to the thin green line labelled 2. However, the deuteron is in a combination of the $l = 0$ and $l = 2$ states. The energy of the interaction of these two states lowers the potential energy greatly. It produces the combined potential energy curve shown as the thick green line.

In terms of chapter 5.3, the lowering of the potential energy is a twilight effect. Even for an $l = 2$ state probability of only 5.8%, a twilight effect can be quite large. That is because it is proportional to the square root of the 5.8% probability, which is a quarter. In addition, the factor $V_{\rm T}$ in the tensor potential turns out to be quite large.

A.41.5 Spin-orbit interaction

So far, the assumption has been that the potential of the deuteron only depends on its geometry. But scattering data suggests that the potential also depends on the motion of the nucleons. A similar effect, the "spin-orbit coupling" occurs for the proton-electron hydrogen atom, addendum {A.39}. However, there the effect is very small. Spin-orbit interaction is proportional to the dot product of net orbital angular momentum and net nucleon spin. In particular, it produces a term in the potential of the form

$$V_{\rm so}(r)\widehat{\vec{L}} \cdot \widehat{\vec{S}}$$

This term does not contribute to the uncertainty in orbital angular momentum. But it can explain how nucleon beams can get polarized with respect to spin in scattering experiments.

A.42 Nuclear forces

The purpose of this addendum is to examine the nature of nuclear forces somewhat closer. The forces will be modeled using the "meson exchange" idea. This idea illustrates one primary way that physicists cope with the fact that nuclei are too complex to describe exactly.

A.42.1 Basic Yukawa potential

As pointed out in chapter 7.5.2, the fundamental forces of nature between elementary particles are due to the exchange of bosons between these particles. In those terms, nuclei consist of quarks. The exchange of gluons between these quarks produces the so-called color force. It is that force that holds nuclei together. Unfortunately, describing that mathematically is not a practical proposition. Quantum chromedynamics is prohibitively difficult.

But you will never find quarks or gluons in isolation. Quarks and their gluons are always confined inside "colorless" combinations of two or three quarks. (To be painstakingly honest, there might be more exotic colorless combinations of quarks and gluons than that. But their energy should be too high to worry about here.) What is observed physically at the time of writing, 2012, are groups of three quarks, (baryons), three antiquarks, (antibaryons), and a quark and an antiquark (mesons). An easier description of nuclear forces can be based on these groups of quarks.

In this picture, nuclei can be taken to consist of nucleons. A nucleon consists of a group of three quarks, so it is a baryon. There are two types of nucleons: protons and neutrons. A proton contains two "up" quarks, at electric charge $\frac{2}{3}e$ each, and one "down" quark, at $-\frac{1}{3}e$. That makes the net charge of a proton $\frac{2}{3}e+\frac{2}{3}e-\frac{1}{3}e$ equal to e. A neutron has one up quark and two down ones, making its net charge $\frac{2}{3}e-\frac{1}{3}e-\frac{1}{3}e$ equal to zero.

For both protons and neutrons, the group of three quarks is in its ground state, much like a helium atom is normally in its ground state. Like single quarks, nucleons are fermions with spin equal to $\frac{1}{2}$. (Roughly speaking, two of the three quarks in nucleons align their spins in opposite directions, causing them to cancel each other.) Nucleons have positive intrinsic parity. That means that their mere presence does not produce a change in sign in the wave function when the coordinate system is inverted, chapter 7.3. (Actually, there is some ambiguity in the assignment of intrinsic parity to particles. But a fermion and

the corresponding antifermion must have opposite parity. Taking the parity of fermions positive makes that of the corresponding antifermions negative.)

Protons and neutrons combine together into nuclei. However, the protons in nuclei repel each other because of their electric charges. So there must be some compensating force that keeps the nucleons together anyway. This force is what is called the "nuclear force." The question in this addendum is how this nuclear force can be described. Its physical cause is still the force due to the exchange of gluons between quarks. But its mathematical description is going to be different. The reason is that by definition the nuclear force is a net force on nucleons, i.e. on *groups* of quarks. And it is assumed to depend on the average positions, and possibly momenta, of these *groups* of quarks.

Note that there is some approximation involved here. Exactly speaking, the nuclear forces should depend on the positions of the individual quarks in the nucleons, not just on their average position. That is a concern when two nucleons get very close together. For one, then the distinction between the two separate groups of quarks must blur. Nucleons do repel one another strongly at very close distances, much like atoms do due to Pauli repulsion, chapter 5.10. But still their quantum uncertainty in position creates a probability for them to be very close together. Fortunately, typical energy levels in normal nuclear physics are low enough that this is not believed to be a dominating effect. Indeed, the models discussed here are known to work very well at larger nucleon spacings. For smaller nucleon spacing however, they become much more complex, and their accuracy much more uncertain. And that happens well before the nucleons start intruding significantly on each others space. Little in life is ideal, isn't it?

In a particle exchange explanation of the nuclear force, roughly speaking nucleons have to "pop up" particles that other nucleons then absorb and vice-versa. The first question is what these particles would be. As already mentioned, only colorless combinations of quarks and their gluons are observed in isolation. Therefore only such colorless combinations can be expected to be able to readily bridge the gap between nucleons that are relatively far apart. The lowest energy of these colorless combinations are the easiest to pop up. And that are the pions; a pion is a meson consisting of a quark and antiquark pair in its ground state.

There are three types of pions. The π^+ pion consists of an up quark plus an antidown quark. Antiparticles have the opposite charge from the corresponding particles, so the antidown quark has charge $\frac{1}{3}e$. That makes the net charge of the π^+ pion $\frac{2}{3}e + \frac{1}{3}e$ equal to e, the same as that of the proton. The π^- pion consists of an antiup quark plus a down quark, producing a net charge $-\frac{2}{3}e - \frac{1}{3}e$ equal to $-e$. That is as it should be since self-evidently the π^- is the antiparticle of the π^+. The π^0 pion is a quantum superposition of an up-antiup pair and a down-antidown pair and is electrically neutral.

Pions are bosons of zero spin and negative intrinsic parity. The negative parity is due to the antiquark, and zero spin is due to the fact that in pions the

quark and antiquark align their spins in opposite directions in a singlet state, chapter 5.5.6.

These pions are the most important particles that protons and neutrons exchange. The first question is then of course where they come from. How is it possible that pions just appear out of nothing? Well, it is possible due to a mixture of special relativity and the uncertainty inherent in quantum mechanics.

The creation of particles out of energy is allowed by special relativity. As discussed in chapter 1.1.2, special relativity gives the energy E of a particle as:

$$E = \sqrt{\vec{p}^{\,2}c^2 + (mc^2)^2}$$

Here c is the speed of light, $\vec{p}$ the momentum of the particle, and m its mass (at rest). According to this expression, a particle at rest represents an amount of energy equal to mc^2. This is the rest mass energy. The charged π^+ and π^- pions have a rest mass energy of about 140 MeV, and the neutral π^0 135 MeV. So to create an actual pion requires at least 135 MeV of energy.

Quantum mechanics replaces the momentum $\vec{p}$ in the energy above by the operator $\hbar\nabla/\mathrm{i}$ in order to find the Hamiltonian. Then it applies that Hamiltonian to a pion wave function φ_π. But the square root in the above expression is a problem. Fortunately, for spinless bosons like pions an acceptable solution is easy: just square the energy. Or rather, apply the Hamiltonian twice. That produces the relativistic so-called Klein-Gordon eigenvalue problem

$$-\hbar^2c^2\nabla^2\varphi_\pi + \left(m_\pi c^2\right)^2\varphi_\pi = E^2\varphi_\pi \tag{A.260}$$

Now consider first a single nucleon located at the origin. Supposedly this nucleon can pop up a pion. But where would the nucleon get the 135 MeV or more of energy? Surely, if there was a probability of actually finding a 135 MeV pion well away from the nucleon, it would violate energy conservation. But remarkably, despite the positive pion rest mass energy, the Klein-Gordon equation has a simple solution where the total pion energy E appears to be zero:

$$\varphi_\pi = C\frac{e^{-r/R}}{r} \qquad R \equiv \frac{\hbar c}{m_\pi c^2} \approx 1.4 \text{ fm}$$

Here r is the distance from the nucleon and C an arbitrary constant. In effect, this solution has a big negative kinetic energy. You might say that a zero-energy pion "tunnels out" of the nucleon, chapter 7.12.2.

To check the above solution, just plug it in the Klein-Gordon equation (A.260) with $E = 0$, using the expression (N.5) for the Laplacian ∇^2 found in the notations. But to be true, this substitution is somewhat misleading. A more careful analysis shows that the left hand side in the Klein-Gordon equation does have a nonzero spike at $r = 0$, {D.2.2}. But there the pion will experience an interaction energy with the nucleon.

Now assume that the nucleon does indeed manage to create a pion field around itself. A field that acts as a potential φ that can produce forces on other nucleons. That would be much like a charged particle creates an electrostatic potential that can produce forces on other charged particles. Then it seems a plausible guess that the pion potential φ will vary with position just like the zero-energy wave function φ_π above. Look at electromagnetics. The photon of electromagnetics has zero rest mass. And for zero rest mass, the zero-energy wave function above becomes the correct C/r Coulomb potential of electromagnetics.

Actually, despite the fact that it works for electromagnetics, the zero-energy wave function above does not quite give the right form for a pion potential. But it does give the general idea. The correct potential is discussed in the next subsections. This subsection will stick with the form above as qualitatively OK.

Now consider a second nucleon. This nucleon will of course also create a pion potential. That is just like if it was all by itself. But in addition, it will interact with the pion potential created by the first nucleon. So there will be an energy of interaction between the nucleons. Taking another cue from electromagnetics, this energy should presumably be proportional to the potential that the first nucleon creates at the position of the second nucleon.

That idea then gives the interaction energy between two nucleons as

$$\boxed{V_{\text{Yukawa}} = -C_{\text{Y}} \frac{e^{-r/R}}{r} \qquad R \equiv \frac{\hbar c}{m_\pi c^2} \approx 1.4 \text{ fm}} \tag{A.261}$$

Here r is the distance between the two nucleons, and C_{Y} is some positive constant that must be determined experimentally. The above interaction energy is called the "Yukawa potential" after the Japanese physicist who first derived it. It is really a potential energy, rather than a potential. (At least in the terminology of this book for fields. In physics, pretty much everything is called a potential.)

The Yukawa potential is attractive. This is in contrast to the Coulomb potential, which is repulsive between like charges. The best physical explanation for the difference may be the analysis in {A.22}, in particular {A.22.5}. (There are many other "explanations" that derive the difference using an electromagnetic Hamiltonian or Lagrangian that already has the difference build in. But a derivation is not an explanation.)

Note the exponential in the Yukawa potential. It will make the potential negligibly small as soon as the distance r between the nucleons is significantly greater than R. With $\hbar c$ about 197 MeV fm and the average pion rest mass energy about 138 MeV, R is about 1.4 fm (femtometer). So unless the nucleons are within a distance not much greater than 1.4 fm from each other, they do not experience a nuclear force from each other. Yukawa had derived the typical range of the nuclear force.

Actually, at the time that Yukawa did his work, the pion was unknown. But the range of the nuclear force was fairly well established. So Yukawa really predicted the existence, as well as the mass of the pion, a then unknown particle! After a long and frustrating search, this particle was eventually discovered in cosmic rays.

The Yukawa potential also explained why heavy nuclei are unstable. Suppose that you keep stuffing nucleons, and in particular protons, into a nucleus. Because of the exponential in the Yukawa potential, the nuclear force is very short range. It is largely gone beyond distances of a couple of fm. So a proton gets pulled *into* the nucleus only by the nucleons in its immediate vicinity. But the Coulomb repulsion between protons does not have the exponential decay. So the same proton gets pushed *out* of the nucleus by protons from all over the nucleus. If the nucleus is big enough, the pushers simply have to win because of their much larger numbers.

Putting a lot of neutrons in the nucleus can help, because they produce nucleon attraction and no Coulomb repulsion. But neutrons by themselves are unstable. Put too many neutrons in a nucleus, and they will turn into protons by beta decay. Obviously, that defeats the purpose. As a result, beyond a certain size, the nucleus is going to fall apart whatever you do.

You can see why Yukawa would end up with the Nobel prize in physics.

A.42.2 OPEP potential

The Yukawa potential energy (A.261) described in the previous section is not quite right yet. It does not give the true nuclear force between two nucleons produced by pion exchange.

In a more careful analysis, the potential energy depends critically on the properties of the exchanged particle. See the next subsection for an explanation of that. For a pion, the relevant properties are that it has zero spin and negative parity. Taking that into account produces the so-called "one-pion exchange potential" energy or "OPEP" for short:

$$\boxed{V_{\rm OPEP} \sim \frac{g_\pi^2}{12}\left(\frac{m_\pi}{m_{\rm p}}\right)^2 m_\pi c^2 \;\vec{\tau}_1\cdot\vec{\tau}_2\Big[\vec{\sigma}_1\cdot\vec{\sigma}_2 + S_{12}V_{\rm T}\Big]\frac{e^{-r/R}}{r/R}} \tag{A.262}$$

$$S_{12} \equiv \frac{3}{r^2}(\vec{\sigma}_1\cdot\vec{r})(\vec{\sigma}_2\cdot\vec{r}) - \vec{\sigma}_1\cdot\vec{\sigma}_2 \qquad V_{\rm T} \equiv 1 + 3\frac{R}{r} + 3\frac{R^2}{r^2}$$

Here r is again the distance between nucleons 1 and 2, equal to the length of the vector $\vec{r}$ connecting the two nucleons, R is again the typical range of the nuclear force, m_π again the pion mass, while $m_{\rm p}$ is the nucleon mass:

$$r \equiv |\vec{r}| \equiv |\vec{r}_2 - \vec{r}_1| \quad R \equiv \frac{\hbar c}{m_\pi c^2} \approx 1.4 \text{ fm} \quad m_\pi c^2 \approx 138 \text{ MeV} \quad m_{\rm p} c^2 \approx 938 \text{ MeV}$$

Also $g_\pi^2 \approx 15$ is an empirical constant, [36, p. 135], [5, p. 85]. Further $\vec{\sigma}_1$ and $\vec{\sigma}_2$ are the nucleon spins $\widehat{\vec{S}}_1$ and $\widehat{\vec{S}}_2$, nondimensionalized by dividing by $\frac{1}{2}\hbar$.

Finally the dot product $\vec{\tau}_1 \cdot \vec{\tau}_2$ involves the so-called "isospin" of the nucleons. Isospin will be discussed in chapter 14.18. There it will be explained that it has nothing to do with spin. Instead isospin is related to nucleon type. In particular, if both nucleons involved are protons, or if both are neutrons, then $\vec{\tau}_1 \cdot \vec{\tau}_2 = 1$.

If one nucleon is a proton and the other a neutron, like in the deuteron, the value of $\vec{\tau}_1 \cdot \vec{\tau}_2$ can vary. But in any case, it is related to the symmetry of the spatial and spin states. In particular, compare also chapter 5.5.6 and {A.10},

$$\text{symmetric spatially:} \quad \vec{\sigma}_1 \cdot \vec{\sigma}_2 = \begin{cases} -3 \text{ (singlet)} & \Longrightarrow \quad \vec{\tau}_1 \cdot \vec{\tau}_2 = 1 \\ 1 \text{ (triplet)} & \Longrightarrow \quad \vec{\tau}_1 \cdot \vec{\tau}_2 = -3 \end{cases}$$

$$\text{antisymmetric spatially:} \quad \vec{\sigma}_1 \cdot \vec{\sigma}_2 = \begin{cases} -3 \text{ (singlet)} & \Longrightarrow \quad \vec{\tau}_1 \cdot \vec{\tau}_2 = -3 \\ 1 \text{ (triplet)} & \Longrightarrow \quad \vec{\tau}_1 \cdot \vec{\tau}_2 = 1 \end{cases}$$

For the deuteron, as well as for the hypothetical diproton and dineutron, the spatial state is symmetric under nucleon exchange. That is as you would expect for a ground state, {A.8} and {A.9}. It then follows from the above values that the first, $\vec{\sigma}_1 \cdot \vec{\sigma}_2$, term in the square brackets in the OPEP (A.262) produces a negative, attractive, potential for these nuclei. That is true regardless whether the spin state is singlet or triplet.

The second, $S_{12}V_{\rm T}$, term in the OPEP is called a tensor potential, {A.41.4}. This potential can create uncertainty in orbital angular momentum. As discussed in {A.41.4}, having a tensor potential is an essential part in getting the deuteron bound. But the tensor potential is zero for the singlet spin state. And the spin state must be the singlet one for the diproton and dineutron, to meet the antisymmetrization requirement for the two identical nucleons. So the diproton and dineutron are not bound.

The deuteron however can be in the triplet spin state. In that case the tensor potential is not zero. To be sure, the tensor potential does average out to zero over all directions of $\vec{r}$. But that does not prevent attractive niches to be found. And note how big the multiplying factor $V_{\rm T}$ is for R about 1.4 fm and nucleon spacings r down to say 2 fm. The tensor potential is big.

Of course, that also depends on S_{12}. But S_{12} is not small either. For example, if $\vec{r}$ is in the x-direction, then S_{12} times a triplet state is three times the opposite triplet state minus the original one.

A.42.3 Explanation of the OPEP potential

The purpose of this subsection is to explain the OPEP potential between nucleons as given in the previous subsection physically.

Note that the objective is not to give a rigorous derivation of the OPEP potential using advanced quantum field theory. Physicists presumably already

got the OPEP right. They better, because it is a standard part of current nuclear potentials. The explanations here will be based on simple physical assumptions. They follow the derivation of the Koulomb potential in {A.22.1}. That derivation was classical, although a simple quantum field version can be found in {A.22.3}. Note that the original Yukawa derivation was classical too. It was still worth a Nobel prize.

The arguments here are loosely based on [16, p. 282-288]. However, often the assumptions made in that reference seem quite arbitrary. To avoid that, the exposition below makes much more liberal use of quantum ideas. After all, in final analysis the classical field is just a reflection of underlying quantum mechanics. Hopefully the quantum arguments will show much more compellingly that things just have to be the way they are.

First of all, like in the first subsection it will be assumed that every nucleon can generate a pion potential. Other nucleons can observe that potential and interact with it, producing forces between the nucleons involved.

The net pion potential produced by all the nucleons will be called φ. It will be assumed that the energy in the observable pion field is given in terms of φ as

$$E_\varphi = \frac{\epsilon_1}{2} \int \left| \frac{1}{c} \frac{\partial \varphi}{\partial t} \right|^2 + |\nabla \varphi|^2 + \left| \frac{m_\pi c^2}{\hbar c} \varphi \right|^2 \, \mathrm{d}^3 \vec{r}$$

Here ϵ_1 is some empirical constant, m_π the pion mass, and the integral is over all space. There should be *some* expression for the energy in the observable field, and the integral above is what the Klein-Gordon equation for free pions preserves, {D.32}. So it seems the likely expression. Also, the above integral gives the correct energy in an electrostatic field, chapter 13.2 (13.11), taking into account that the photon has no mass.

(Do note that there are some qualifications to the statement that the above integral gives the correct energy in an electrostatic field. The electromagnetic field is quite tricky because, unlike the pion, the photon wave function is a relativistic four-vector. See {A.22} for more. But at the very least, the integral above gives the correct expression for the *effective* energy in the electrostatic field.)

Finally it will be assumed that there is an interaction energy between the observable pion field and the nucleons. But the precise expression for that interaction energy is not yet obvious. Only a generic expression can reasonably be postulated at this stage. In particular, it will be postulated that the interaction energy of the pion field with an arbitrary nucleon numbered i takes the form:

$$E_{\varphi i} = - \int \varphi f_i \, \mathrm{d}^3 \vec{r}$$

The minus sign is inserted since the interaction will presumably lower the energy. If it did not, there should be no pion field at all in the ground state. The factor f_i will be called the interaction factor of nucleon i.

It still needs to be figured out what is the appropriate form of this interaction factor. But it will be assumed that it involves the wave function Ψ_i of nucleon i in some way. In particular, in regions where the wave function is zero, f_i will be zero too. That means that where there is no probability of finding the nucleon, there is no interaction of the nucleon with the field either. In other words, the interaction is "local," rather than long range; it occurs at the location of the nucleon. One motivation for this assumption is that long-range interactions are just bound to produce problems with special relativity.

It will further be assumed that the wave function of each nucleon i is slighly spread out around some nominal position $\vec{r}_i$. After all, if you want a potential in terms of nucleon positions, then nucleons should at least approximately have positions. One immediate consequence is then that the interaction factor f_i is zero except close to the nominal position $\vec{r}_i$ of the nucleon.

The ground state is the state in which the combined pion field and interaction energy is minimal. To find the properties of that state requires variational calculus. This is worked out in considerable detail in {A.22.1} and {A.2}. (While those derivations do not include the m_π term, its inclusion is trivial.) The analysis shows that the observable potential must satisfy

$$-\nabla^2\varphi + \left(\frac{m_\pi c^2}{\hbar c}\right)^2 \varphi = \frac{1}{\epsilon_1}\sum_i f_i \tag{A.263}$$

As noted above, the interaction factors f_i in the right hand side are zero away from the nucleons. And that means that away from the nucleons the potential satisfies the Klein-Gordon eigenvalue problem with zero energy. That was a good guess, in the first subsection! But now the complete potential can be figured out, given the interaction factors f_i.

The variational analysis further shows that the energy of interaction between a nucleon numbered i and one numbered j is:

$$V_{ij} = -\int \varphi_i(\vec{r}) f_j(\vec{r})\, \mathrm{d}^3\vec{r} \tag{A.264}$$

Here φ_i is the potential caused by nucleon i. In other words, φ_i is the solution of (A.263) if only a single interaction factor f_i in the sum in the right hand side is included.

The big question remains, what exactly is the interaction factor f_i between the pion field and a nucleon i? The first guess would be that the interaction energy at a given position is proportional to the probability of finding the nucleon at that position. In short,

$$f_{i,\mathrm{fg}} = g|\Psi_i|^2 \quad ?$$

where "fg" stands for "first-guess" and g is some constant. This reflects that the probability of finding the nucleon is given by its square wave function $|\Psi_i|^2$. The above interaction factor is essentially what you would have in electrostatics.

There g would be the electric charge, so for pions you could call it the "mesic charge." (Note that the square wave function integrates to 1 so the integrated interaction factor above is g.)

Given the above first-guess interaction factor, according to (A.263) a nucleon i would create a first-guess potential, {D.2.2},

$$\varphi_{i,\text{fg}} = \frac{g}{4\pi\epsilon_1}\frac{e^{-r/R}}{r} \qquad \text{(except vanishingly close to the nucleon } i\text{)} \qquad \text{(A.265)}$$

Here r is the distance from the nucleon. If you assume for simplicity that the nucleon is at the origin, r is the distance from the origin.

The above potential is spherically symmetric; it is the same in all directions. (That is true even if the nucleon wave function is not spherically symmetric. The wave function is only nonzero very close to $\vec{r}_i$, so it looks like a single point away from the immediate vicinity of the nucleon.)

The interaction energy with a second nucleon j may now be found using (A.264). In particular, because the wave function of nucleon j is only nonzero very close to its nominal position $\vec{r}_j$, you can approximate $\varphi_i(\vec{r})$ in (A.264) as $\varphi_i(\vec{r}_j)$. Then you can take it out of the integral. So the interaction energy is proportional to $\varphi_i(\vec{r}_j)$. That is the potential caused by nucleon i evaluated at the position of nucleon j. That was another good guess, in the first subsection! More precisely, you get

$$V_{ij,\text{fg}} = -\frac{g^2}{4\pi\epsilon_1}\frac{e^{-r_{ij}/R}}{r_{ij}}$$

where $r_{ij} = |\vec{r}_j - \vec{r}_i|$ is the distance between the nucleons. This first guess potential energy is the Yukawa potential of the first subsection.

The Yukawa potential would be appropriate for a field of spinless pions with positive intrinsic parity. And except for the sign problem mentioned in the first subsection, it also gives the correct Coulomb potential energy in electrostatics.

Unfortunately, as noted in the first subsection, the pion has negative intrinsic parity, not positive. And that is a problem. Imagine for a second that a nucleon pops up a pion. The nucleon has positive parity. However, the pion that pops up has negative intrinsic parity. And parity is preserved, chapter 7.3. If the intrinsic parity of the pion is negative, its orbital parity must be negative too to maintain a positive combined system parity, chapter 7.4.2. Negative orbital parity means that the pion wave function φ_π must have opposite values at $\vec{r}$ and $-\vec{r}$. But as mentioned, the first-guess potential is spherically symmetric; the values at $\vec{r}$ and $-\vec{r}$ are the same.

(Note that this argument blurs the distinction between a pion wave function φ_π and an observable pion potential φ. But you would expect them to be closely related, {A.22.3}. In particularly, reasonably speaking you would expect that spherically symmetric wave functions correspond to spherically symmetric observable potentials, as well as vice-versa.)

(You might also, correctly, object to the inaccurate picture that the nucleon pops up a pion. The ground state of the nucleon-pions system is a state of definite energy. Energy states are stationary, chapter 7.1.4. However, in energy states the complete nucleon-pions system should have definite angular momentum and parity, chapter 7.3. That is just like nuclei in energy states have definite angular momentum and parity, chapter 14.1. The term in the nucleon-pions system wave function in which there is just the nucleon, with no pions, already sets the angular momentum and parity. A different term in the system wave function, in particular one in which there is a pion in a state of definite angular momentum and parity, cannot have different angular momentum or parity. Otherwise angular momentum and parity would have uncertainty.)

So how to fix this? Suppose that you differentiate the first-guess potential (A.265) with respect to, say, x. The differentiation will bring in a factor x in the potential,

$$\frac{\partial \varphi_{i,\mathrm{fg}}}{\partial x} = \frac{\partial \varphi_{i,\mathrm{fg}}}{\partial r}\frac{x}{r}$$

And that factor x will produce an opposite sign at $-\vec{r}$ compared to $\vec{r}$. That means that the parity is now negative as it should be.

According to (A.263), the first guess potential satisfies

$$-\nabla^2\varphi_{i,\mathrm{fg}} + \left(\frac{m_\pi c^2}{\hbar c}\right)^2 \varphi_{i,\mathrm{fg}} = \frac{1}{\epsilon_1} g|\Psi_i|^2$$

Differentiating both sides with respect to x, you get for its x-derivative, the second-guess potential $\varphi_{i,\mathrm{sg}}$:

$$-\nabla^2\varphi_{i,\mathrm{sg}} + \left(\frac{m_\pi c^2}{\hbar c}\right)^2 \varphi_{i,\mathrm{sg}} = \frac{1}{\epsilon_1}\sum_i g\frac{\partial}{\partial x}|\Psi_i|^2 \qquad \varphi_{i,\mathrm{sg}} = \frac{\partial \varphi_{i,\mathrm{fg}}}{\partial x}$$

So apparently, if you put a x-derivative on the square nucleon wave function in the first-guess interaction factor $f_{i,\mathrm{fg}}$ you get a pion potential consistent with parity conservation.

There are a couple of new problems. First of all, this potential now has orbital angular momentum. If you check out the spherical harmonics in table 4.3, you see that a spherically symmetric wave function has no orbital angular momentum. But the factor x produces a wave function of the form

$$c(r)Y_1^1 - c(r)Y_1^{-1}$$

where $c(r)$ is some spherically symmetric function. The first term above has angular momentum $\hbar$ in the z-direction. The second term has angular momentum $-\hbar$ in the z-direction. So there is uncertainty in angular momentum, but it is not zero. The azimuthal quantum number of square orbital angular momentum, call it l_π, is 1 with no uncertainty.

So where does this angular momentum come from? Angular momentum should be preserved. The pion itself has no spin. So its orbital angular momentum will have to come from the half unit of nucleon spin. Indeed it is possible for half a unit of nucleon spin, $s_i = \frac{1}{2}$, and one unit of pion orbital angular momentum, $l_\pi = 1$, to combine into still only half a unit of net angular momentum $j = \frac{1}{2}$, 7.4.2.

But consider also the angular momentum in the z-direction. If the pion is given $\hbar$ in the z-direction, then that must come from the fact that the nucleon spin changes from $\frac{1}{2}\hbar$ in the z-direction to $-\frac{1}{2}\hbar$. Conversely, if the pion has $-\hbar$, then the nucleon must change from $-\frac{1}{2}\hbar$ to $\frac{1}{2}\hbar$. Either way, the nucleon spin in the z-direction must flip over.

In quantum terms, how does that happen? Consider the scaled nucleon z spin operator σ_z for a second. If you apply this operator on the "spin-up" state with z spin $\frac{1}{2}\hbar$, you get a multiple of the same state back. (Actually, because of the scaling, you get the exact same state back.) The spin-up state is an eigenstate of the operator σ_z as it should. But the spin-up state is *not* an eigenstate of the operators σ_x and σ_y. These operators do not commute with σ_z. So if you apply σ_x or σ_y on the spin-up state, you will also get some of the $-\frac{1}{2}\hbar$ spin-down state. In fact, if you look a bit closer at angular momentum, chapter 12.10, you see that you get *only* a spin-down state. So both σ_x and σ_y do exactly what is needed; they flip spin-up over to spin-down. Similarly, they flip spin-down over to spin-up.

The second problem has to do with the original notion of differentiating the spherically symmetric potential with respect to x. Why not y or z or some oblique direction? The pion field should not depend on how you have oriented your mathematical axes system. But the x-derivative does depend on it. A similar problem exists of course with arbitrarily choosing one of the operators σ_x or σ_y above.

Now dot products are the same regardless of how the coordinate system is oriented. That then suggests how both problems above can be solved at the same time. In the first-guess interaction factor, add the dot product between the scaled nucleon spin $\vec{\sigma}_i$ and the spatial differentiation operator ∇_i. That gives the third-guess interaction factor as

$$f_{i,\mathrm{tg}} = gR\vec{\sigma}_i \cdot \nabla_i |\Psi_i|^2 = gR \left[\sigma_x \frac{\partial}{\partial x_i} + \sigma_y \frac{\partial}{\partial y_i} + \sigma_z \frac{\partial}{\partial z_i} \right] |\Psi_i|^2$$

The factor R has been added to keep the units of the first guess intact.

Time for a reality check. Consider a nucleon in the spin-up state. If the "mesic charge" g would be zero, there would be no pion field. There would just be this bare nucleon with half a unit of spin-up and positive parity. Next assume that g is not zero, but still small. Then the bare nucleon term should still dictate the spin and intrinsic parity. There will now also be terms with pions in the complete system wave function, but they must obey the same spin and

parity. You can work out the detailed effect of the third guess interaction factor above using table 4.3 and chapter 12.10. If you do, you see that it associates the spin-up nucleon with a state

$$\sqrt{4\pi}\frac{\partial\varphi_{i,\mathrm{fg}}}{\partial r}\left(\sqrt{\tfrac{1}{3}}Y_1^0\uparrow - \sqrt{\tfrac{2}{3}}Y_1^1\downarrow\right)$$

where Y_1^0 and Y_1^1, table 4.3, describe the spatial pion potential and $\uparrow$ and $\downarrow$ nucleon spin-up, respectively spin-down. Loosely associating the pion potential with a pion wave function, you can check from the Clebsch-Gordon tables 12.5 that the state in parentheses obeys the spin and parity of the original bare nucleon.

So the third guess seems pretty good. But there is one more thing. Recall that there are three different pions, with different charges, So you would expect that there are really three different functions f_i, one for each pion field. Alternatively, the function f_i should be three-dimensional vector. But what sort of vector?

Note that charge is preserved. If a proton pops up a positively charged π^+ pion, it must itself change into a uncharged neutron. And if a neighboring neutron absorbs that π^+, it acquires its positive charge and turns into a proton. The same thing happens when a neutron emits a negatively charged π^- that a proton absorbs. Whenever a charged particle is exchanged between a proton and a neutron, both change type. (Charged particles cannot be exchanged between nucleons of the same type because there are no nucleons with negative charge or with two units of positive charge.)

So, it is necessary to describe change of nucleon type. Physicists do that in a very weird way; they pattern the mathematics on that of spin, chapter 14.18. First a completely abstract "123" coordinate system is introduced. If a nucleon is a proton, then it is said that the nucleon has a component $\frac{1}{2}$ along the abstract 3-axis. If a nucleon is a neutron, it is said that it has a component $-\frac{1}{2}$ along the 3-axis.

Compare that with spin. If a nucleon is spin-up, it has a spin component $\frac{1}{2}\hbar$ along the physical z-axis. If it is spin-down, it has a spin component $-\frac{1}{2}\hbar$ along the z-axis. The idea is very similar.

Now recall from above that the operators σ_x and σ_y flip over the spin in the z-direction. In 123-space, physicist define abstract operators τ_1 and τ_2 that do a similar thing: they flip over the value along the 3-axis. And that means that these operators change protons into neutrons or vice-versa. So they do exactly what is needed in exchanges of charged pions. Physicist also define an operator τ_3, analogous to σ_z, which does not change the 3-component.

Of course, all this may seem an extremely round-about way of doing something simple: define operators that flip over nucleon type. And normally it really would be. But if it is assumed that nuclear forces are charge-independent, (which is a reasonable approximation), things change. In that case it turns out

that the physics must remain the same under rotations of this abstract 123-coordinate system. And that requirement can again be met by forming a dot product, this time between $\vec{\tau} = (\tau_1, \tau_2, \tau_3)$ vectors.

That idea then gives the final expression for the functions f_i:

$$\vec{f_i} = gR\vec{\tau_i} \;\; \vec{\sigma_i} \cdot \nabla_i |\Psi_i|^2 = gR\vec{\tau_i} \left[\sigma_x \frac{\partial}{\partial x_i} + \sigma_y \frac{\partial}{\partial y_i} + \sigma_z \frac{\partial}{\partial z_i} \right] |\Psi_i|^2$$

Note that $\vec{f_i}$ is now a three-dimensional vector because $\vec{\tau_i}$ is. In the final potential, $\vec{\tau_i}$ gets into a dot product with $\vec{\tau_j}$ of the other nucleon. That makes the complete potential the same regardless of rotation of the abstract 123-coordinate system as it should.

Now it is just a matter of working out the final potential. Do one thing at a time. Recall first the effect of the x-derivative on the nucleon wave function. It produces a potential that is the x-derivative of the spherically symmetric first-guess potential (A.265). That works out to

$$\frac{gR}{4\pi\epsilon_1} \frac{\mathrm{d}e^{-r/R}/r}{\mathrm{d}r} \frac{\partial r}{\partial x} = -\frac{gR}{4\pi\epsilon_1} \left[\frac{1}{Rr^2} + \frac{1}{r^3} \right] e^{-r/R} x$$

Of course, there ares similar expressions for the derivatives in the other two directions. So the potential produced by nucleon i at the origin is

$$\varphi_i(\vec{r}) = -\frac{gR}{4\pi\epsilon_1} \vec{\tau_i} \left[\frac{1}{Rr^2} + \frac{1}{r^3} \right] e^{-r/R} \vec{r} \cdot \vec{\sigma}_i$$

Now the interaction potential with another nucleon follows from (A.264). But here you need to be careful. The integral will involve terms like

$$[\text{some constant}] \int \varphi_i(\vec{r}) \frac{\partial |\Psi_j|^2}{\partial x} \, \mathrm{d}^3\vec{r}$$

In this case, you cannot just approximate $\vec{r}$ in $\varphi_i(\vec{r})$ as the nominal position $\vec{r}_j$ of nucleon j. That is not accurate. Since the x-derivative works on a very concentrated wave function, it will produce large negative and positive values, and errors will accumulate. The solution is to perform an integration by parts in the x-direction. That puts the derivative on the potential instead of the wave function and adds a minus sign. Then you can evaluate this negative derivative of the potential at the nominal position of nucleon $\vec{r}_j$.

Differentating the potential is a bit of a mess, but straightforward. Then the potential becomes

$$V_{ij} \sim \frac{g^2}{12\pi\epsilon_1 R} \;\; \vec{\tau_i} \cdot \vec{\tau_j} \Big[\vec{\sigma_i} \cdot \vec{\sigma_j} + S_{ij} V_{\mathrm{T}} \Big] \frac{e^{-r/R}}{r/R}$$

If you define the constant g_π appropriately, this gives the OPEP potential (A.262).

A.42.4 Multiple pion exchange and such

Unfortunately, the nuclear force is not just a matter of the exchange of single pions. While the OPEP works very well at nucleon distances above 3 fm, at shorter ranges other processes become important.

The most important range is the one of the primary nucleon attractions. Conventionally, this range is ballparked as nucleon distances in the range $1 < r < 2$ fm, [5, p. 91], [[3]]. (References vary about the actual range however, [31, p. 111], [36, pp. 177].) In this range, two-pion exchanges dominate. In such exchanges two pions appear during the course of the interaction. Since this requires double the uncertainty in energy, the typical range is correspondingly smaller than for one-pion exchanges.

Two-pion exchanges are much more difficult to crunch out than one-pion ones. In addition, it turns out that straightforward two-pion exchanges are not enough, [[3]]. The interactions also have to include various so-called "resonances."

Resonances are extremely short-lived excited states of baryons and mesons. They decay through the strong force, which typically takes on the order of 10^{-23} s. A particle moving near the speed of light will only travel a distance of the order of a femtometer during such a time. Therefore resonances are not observed directly. Instead they are deduced from experiments in which particles are scattered off each other. Excited states of nucleons can be deduced from preferred scattering of particular frequencies. More or less bound states of pions can be deduced from collision dynamics effects. Collisions involving three particles are quite different if two of the three particles stick together, even briefly, than if all three go off in separate directions.

The lowest energy excited state for nucleons is a set of resonances called the "delta particles," Δ^{++}, Δ^{+}, Δ^{0}, and Δ^{-}. In the deltas, the three constituent quarks of the nucleons align their spins in the same direction for a net spin of $\frac{3}{2}$. The state further has enough antisymmetry to allow three quarks to be equal. That explains the nucleon charge $2e$ of the Δ^{++}, consisting of three up quarks at $\frac{2}{3}e$ each, and the charge $-e$ of the Δ^{-}, consisting of three down quarks at $-\frac{1}{3}e$ each. The delta resonances are often indicated by $\Delta(1232)$, where the quantity between parentheses is the nominal rest mass energy in MeV. That allows excited states of higher energy to be accommodated. If the excited states allow no more than two quarks to be equal, like the normal nucleons, they are indicated by N instead of Δ. In those terms, the normal proton and neutron are $N(939)$ states. (The rest mass energies are nominal because resonances have a tremendous uncertainty in energy. That is to be expected from their short life time on account of the energy-time uncertainty relationship. The "width" of the delta energy is over 100 MeV.)

Pion resonances of interest involve the 775 MeV rho (ρ), and the 783 MeV omega (ω) resonances. Both of these states have spin 1 and odd parity. The

550 MeV eta (η) particle is also of importance. This particle has spin 0 and odd parity like the pions. The eta is not really a resonance, based on its relatively long 0.5 10^{-18} s life time.

Older references like [36] picture the resonances as correlated multi-pion states. However, quantum chromedynamics has been associating actual particles with them. Take the rho, for example. In [36] it is pictured as a two-pion correlated state. (A true bound state of two 140 MeV pions should have an energy less than 280 MeV, not 775 MeV.) However, quantum chromedynamics identifies a rho as a single excited pion with a 775 MeV rest mass. It does decay almost instantly into two pions. The omega, pictured as a three-pion correlated state, is according to quantum chromedynamics a quantum superposition of half an up-antiup and half a down-antidown quark pair, not unlike the neutral rho. It usually decays into three pions. Quantum chromedynamics describes the η as a meson having a strange-antistrange quark component.

The rho and omega resonances appear to be important for the nucleon repulsions at short range. And 3 and 4 pion exchanges have about the same range as the ω. So if the omega is included, as it normally is, it seems that multi-pion exchanges should be included too. Crunching out complete multi-pion pion exchanges, with the additional complications of the mentioned resonances, is a formidable task.

One-meson exchanges are much easier to analyze than multi-meson ones. Therefore physicists may model the multi-pion processes as the exchange of one combined boson, rather than of multiple pions. That produces so-called "one-boson exchange potentials," or "OBEP"s for short. They work surprisingly well.

The precise Yukawa potential that is produced depends on the spin and parity of the exchanged boson, [36, pp. 176ff], [[3]]. The pion has zero spin and negative parity. Such a particle is often called "pseudoscalar." Scalar means that its wave function at each point is a just a number. However, normal numbers, like say a mass, do not change sign if the directions of the axes are inverted. The eta is a 0^- pseudoscalar like the pion. It produces a similar potential as the OPEP.

However, the rho and omega are 1^- bosons. Such bosons are often called ""vector particles." Their wave function at each point is a three-dimensional vector, {A.20}. And normal vectors do change sign if the directions of the axes are inverted, so the rho and omega are not pseudovectors. Vector bosons generate a repulsive potential, among various other effects. That can take care of the needed repulsive short range forces quite nicely.

Unfortunately, to describe the attractive forces in the intermediate range, OBEP models need a roughly 600 MeV 0^+ "scalar" boson. In fact, many OBEP models use both a 500 MeV and a 700 MeV scalar boson. The existence of such scalar resonances has never been accepted. While an older reference like [36, pp. 172] may point to a perceived very wide resonance at 700 MeV, how

convincing can a 700 MeV resonance with a width of at least 600 MeV be? This lack of physical justification does detract from the OBEP potentials.

And of course, they are approximations in any case. There are important issues like multi-nucleon interactions and electromagnetic properties that probably only a comprehensive description of the actual exchange processes can correctly describe, [[3]]. Despite much work, nuclear potentials remain an active research area. One author already thinks in terms of millennia, [32].

A.43 Classical vibrating drop

The simplest collective description for a nucleus models it as a vibrating drop of a macroscopic liquid. To represent the nuclear Coulomb repulsions the liquid can be assumed to be positively charged. This section gives a condensed derivation of small vibrations of such a liquid drop according to classical mechanics. It will be a pretty elaborate affair, condensed or not.

A.43.1 Basic definitions

The drop is assumed to be a sphere of radius R_0 when it is not vibrating. For a nucleus, R_0 can be identified as the nuclear radius,

$$R_0 = R_A A^{1/3}$$

When vibrating, the radial position of the surface will be indicated by R. This radial position will depend on the spherical angular coordinates θ and ϕ, figure N.3, and time.

The mass density, (mass per unit volume), of the liquid will be indicated by ρ_m. Ignoring the difference between proton and neutron mass, for a nucleus the mass density can be identified as

$$\rho_m = \frac{A m_{\rm p}}{\frac{4}{3}\pi R_0^3}$$

The mass density is assumed constant throughout the drop. This implies that the liquid is assumed to be incompressible, meaning that the volume of any chunk of liquid is unchangeable.

The charge density is defined analogously to the mass density:

$$\rho_c = \frac{Ze}{\frac{4}{3}\pi R_0^3}$$

It too is assumed constant.

The surface tension σ can be identified as

$$\sigma = \frac{C_s A^{2/3}}{4\pi R_0^2} = \frac{C_s}{4\pi R_A^2}$$

A.43.2 Kinetic energy

The possible frequencies of vibration can be figured out from the kinetic and potential energy of the droplet. The kinetic energy is easiest to find and will be done first.

As noted, the liquid will be assumed to be incompressible. To see what that means for the motion, consider an arbitrary chunk of liquid. An elementary element of surface area $\mathrm{d}S$ of that chunk gobbles up an amount of volume while it moves given by $\vec{v}\cdot\vec{n}\,\mathrm{d}S$, where v is the liquid velocity and $\vec{n}$ is a unit vector normal to the surface. But for a given chunk of an incompressible liquid, the total volume cannot change. Therefore:

$$\int_S \vec{v}\cdot\vec{n}\,\mathrm{d}S = 0$$

Using the Gauss-Ostrogradsky, or divergence theorem, this means that $\nabla\cdot\vec{v}$ must integrate to zero over the interior of the chunk of liquid. And if it must integrate to zero for whatever you take the chunk to be, it must be zero uniformly:

$$\nabla\cdot\vec{v} = 0$$

This is the famous "continuity equation" for incompressible flow. But it is really no different from Maxwell's continuity equation for the flow of charges if the charge density remains constant, chapter 13.2.

To describe the dynamics of the drop, the independent variables will be taken to be time and the *unperturbed* positions $\vec{r}_0$ of the infinitesimal volume elements $\mathrm{d}^3\vec{r}$ of liquid. The velocity field inside the drop is governed by Newton's second law. On a unit volume basis, this law takes the form

$$\rho_m\frac{\partial\vec{v}}{\partial t} = -\rho_c\nabla\varphi - \nabla p$$

where φ is the electrostatic potential and p is the pressure. It will be assumed that the motion is slow enough that electrostatics may be used for the electromagnetic force. As far as the pressure force is concerned, it is one of the insights obtained in classical fluid mechanics that a constant pressure acting equally from all directions on a volume element of liquid does not produce a net force. To get a net force on an element of liquid, the pressure force on the front of the element pushing it back must be different from the one on the rear pushing it forward. So there must be variations in pressure to get a net force on elements of liquid. Using that idea, it can be shown that for an infinitesimal element of liquid, the net force per unit volume is minus the gradient of pressure. For a real classical liquid, there may also be viscous internal forces in addition to pressure forces. However, viscosity is a macroscopic effect that is not relevant to the nuclear quantum system of interest here. (That changes in a two-liquid description, [40, p. 187].)

Note that the gradients of the potential and pressure should normally be evaluated with respect to the perturbed position coordinates $\vec{r}$. But if the amplitude of vibrations is infinitesimally small, it is justified to evaluate ∇ using the unperturbed position coordinates $\vec{r}_0$ instead. Similarly, ∇ in the continuity equation can be taken to be with respect to the unperturbed coordinates.

If you take the divergence of Newton's equation. i.e. multiply with $\nabla\cdot$, the left hand side vanishes because of continuity, and so the sum of potential and pressure satisfies the so-called "Laplace equation:"

$$\nabla^2(\rho_c\varphi + p) = 0$$

The solution can be derived in spherical coordinates r_0, θ_0, and ϕ_0 using similar, but simpler, techniques as used to solve the hydrogen atom. The solution takes the form

$$\rho_c\varphi + p = \sum_{l,m} c_{lm}(t)\frac{r_0^l}{R_0^l}\bar{Y}_l^m(\theta_0,\phi_0)$$

where the c_{lm} are small unknown coefficients and the $\bar{Y}_l^m$ are real spherical harmonics. The precise form of the $\bar{Y}_l^m$ is not of importance in the analysis.

Plugging the solution for the pressure into Newton's second law shows that the velocity can be written as

$$\vec{v} = \sum_{l,m} v_{lm}(t)R_0\nabla\left(\frac{r_0^l}{R_0^l}\bar{Y}_l^m(\theta_0,\phi_0)\right)$$

where the coefficients v_{lm} are multiples of time integrals of the c_{lm}. What multiples is irrelevant as the potential and pressure will no longer be used.

(You might wonder about the integration constant in the time integration. It is assumed that the droplet was initially spherical and at rest before some surface perturbation put it into motion. If the drop was initially rotating, the analysis here would need to be modified. More generally, if the droplet was not at rest initially, it must be assumed that the initial velocity is "irrotational," meaning that $\nabla \times \vec{v} = 0$.)

Since the velocity is the time-derivative of position, the positions of the fluid elements are

$$\vec{r} = \vec{r}_0 + \sum_{l,m} r_{lm}(t)R_0\nabla\frac{r_0^l}{R_0^l}\bar{Y}_l^m(\theta_0,\phi_0)$$

where $\vec{r}_0$ is the unperturbed position of the fluid element and the coefficients of velocity are related to those of position by

$$v_{lm} = \dot{r}_{lm}$$

What will be important in the coming derivations is the radial displacement of the liquid surface away from the spherical shape. It follows from taking the

radial component of the displacement evaluated at the surface $r_0 = R_0$. That produces

$$R(\theta_0, \phi_0) = R_0 + \delta(\theta_0, \phi_0) \qquad \delta = \sum_{l,m} r_{lm}(t) l \bar{Y}_l^m(\theta_0, \phi_0) \tag{A.266}$$

To be sure, in the analysis δ will be defined to be the radial surface displacement as a function of the physical angles θ and ϕ. However, the difference between physical and unperturbed angles can be ignored because the perturbations are assumed to be infinitesimal.

The kinetic energy is defined by

$$T = \int \tfrac{1}{2} \rho_m \vec{v} \cdot \vec{v} \, \mathrm{d}^3 \vec{r}_0$$

Putting in the expression for the velocity field in terms of the r_{lm} position coefficients gives

$$T = \int \tfrac{1}{2} \rho_m \sum_{l,m} \dot{r}_{lm} R_0 \left(\nabla \frac{r_0^l}{R_0^l} \bar{Y}_l^m \right) \cdot \sum_{\underline{l},\underline{m}} \dot{r}_{\underline{l}\underline{m}} R_0 \left(\nabla \frac{r_0^{\underline{l}}}{R_0^{\underline{l}}} \bar{Y}_{\underline{l}}^{\underline{m}} \right) \, \mathrm{d}^3 \vec{r}_0$$

To simplify this, a theorem is useful. If any two functions F and G are solutions of the Laplace equation, then the integral of their gradients over the volume of a sphere can be simplified to an integral over the surface of that sphere:

$$\int_V (\nabla F) \cdot (\nabla G) \, \mathrm{d}^3 \vec{r}_0 = \int_V \nabla \left(F \cdot \nabla G \right) \, \mathrm{d}^3 \vec{r}_0 = \int_S F \frac{\partial G}{\partial r_0} \, \mathrm{d}S_0 \tag{A.267}$$

The first equality is true because the first term obtained in differentiating out the product $F \cdot \nabla G$ is the left hand side, while the second term is zero because G satisfies the Laplace equation. The second equality is the divergence theorem applied to the sphere. Further, the surface element of a sphere is in spherical coordinates:

$$\mathrm{d}S_0 = R_0^2 \sin\theta_0 \, \mathrm{d}\theta_0 \mathrm{d}\phi_0$$

Applying these results to the integral for the kinetic energy, noting that $r_0 = R_0$ on the surface of the droplet, gives

$$T = \tfrac{1}{2} \rho_m \sum_{l,m} \sum_{\underline{l},\underline{m}} \dot{r}_{lm} \dot{r}_{\underline{l}\underline{m}} \underline{l} R_0^3 \int \int \bar{Y}_l^m \bar{Y}_{\underline{l}}^{\underline{m}} \sin\theta_0 \, \mathrm{d}\theta_0 \mathrm{d}\phi_0$$

Now the spherical harmonics are orthonormal on the unit sphere; that means that the final integral is zero unless $l = \underline{l}$ and $m = \underline{m}$, and in that case the integral is one. Therefore, the final expression for the kinetic energy becomes

$$T = \tfrac{1}{2} \rho_m R_0^3 \sum_{l,m} l \dot{r}_{lm}^2 \tag{A.268}$$

A.43.3 Energy due to surface tension

From here on, the analysis will return to physical coordinates rather than unperturbed ones.

The potential energy due to surface tension is simply the surface tension times the surface of the deformed droplet. To evaluate that, first an expression for the surface area of the droplet is needed.

The surface can be described using the spherical angular coordinates θ and ϕ as $r = R(\theta, \phi)$. An infinitesimal coordinate element $\mathrm{d}\theta\mathrm{d}\phi$ corresponds to a physical surface element that is approximately a parallelogram. Specifically, the sides of that parallelogram are

$$\mathrm{d}\vec{r}_1 = \frac{\partial \vec{r}_{\rm surface}}{\partial \theta}\mathrm{d}\theta \qquad \mathrm{d}\vec{r}_2 = \frac{\partial \vec{r}_{\rm surface}}{\partial \phi}\mathrm{d}\phi$$

To get the surface area $\mathrm{d}S$, take a vectorial product of these two vectors and then the length of that. To work it out, note that in terms of the orthogonal unit vectors of a spherical coordinate system,

$$\vec{r}_{\rm surface} = \hat{\imath}_r R \qquad \frac{\partial \hat{\imath}_r}{\partial \theta} = \hat{\imath}_\theta \quad \frac{\partial \hat{\imath}_r}{\partial \phi} = \sin\theta \hat{\imath}_\phi$$

That way, the surface area works out to be

$$S = \int\int \sqrt{1 + \left(\frac{1}{R}\frac{\partial R}{\partial \theta}\right)^2 + \left(\frac{1}{R\sin\theta}\frac{\partial R}{\partial \phi}\right)^2}\, R^2 \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi$$

Multiply by the surface tension σ and you have the potential energy due to surface tension.

Of course, this expression is too complicated to work with. What needs to be done, first of all, is to write the surface in the form

$$R = R_0 + \delta$$

where δ is the small deviation away from the radius R_0 of a perfect spherical drop. This can be substituted into the integral, and the integrand can then be expanded into a Taylor series in terms of δ. That gives the potential energy $V_s = \sigma S$ as

$$\begin{aligned} V_s &= \sigma \int\int R_0^2 \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi + \sigma \int\int 2R_0\delta \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi + \sigma \int\int \delta^2 \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi \\ &\quad + \sigma \int\int \tfrac{1}{2}\left[\left(\frac{1}{R_0}\frac{\partial \delta}{\partial \theta}\right)^2 + \left(\frac{1}{R_0\sin\theta}\frac{\partial \delta}{\partial \phi}\right)^2\right] R_0^2 \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi \end{aligned}$$

where the final integral comes from expanding the square root and where terms of order of magnitude δ^3 or less have been ignored. The first integral in the

result can be ignored; it is the potential energy of the undeformed droplet, and only differences in potential energy are important. However, the second integral is one problem, and the final one another.

The second integral is first. Its problem is that if you plug in a valid approximate expression for δ, you are still not going to get a valid approximate result for the integral. The radial deformation δ is both negative and positive over the surface of the cylinder, and if you integrate, the positive parts integrate away against the negative parts, and what you have left is mainly the errors.

Why is δ both positive and negative? Because the volume of the liquid must stay the same, and if δ was all positive, the volume would increase. The condition that the volume must remain the same means that

$$\frac{4\pi}{3}R_0^2 = \int\int\int r^2 \sin\theta \,\mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi = \int\int \tfrac{1}{3}R^3 \sin\theta \mathrm{d}\theta\mathrm{d}\phi$$

the first because of the expression for volume in spherical coordinates and the second from integrating out r. Writing again $R = R_0 + \delta$ and expanding in a Taylor series gives after rearranging

$$-\int\int R_0^2\delta \sin\theta \mathrm{d}\theta\mathrm{d}\phi = \int\int R_0\delta^2 \sin\theta \mathrm{d}\theta\mathrm{d}\phi$$

where the integral of δ^3 has been ignored. Now the integral in the left hand side is essentially the one needed in the potential energy. According to this equation, it can be replaced by the integral in the right hand side. And that one can be accurately evaluated using an approximate expression for δ: since the integrand is all positive, there is no cancellation that leaves only the errors. Put more precisely, if the used expression for δ has an error of order δ^2, direct evaluation of the integral in the left hand side gives an unacceptable error of order δ^2, but evaluation of the integral in the right hand side gives an acceptable error of order δ^3.

If this is used in the expression for the potential energy, it produces

$$\begin{aligned} V_s &= V_{s,0} - \sigma\int\int \delta^2 \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi \\ &\quad +\sigma\int\int \tfrac{1}{2}\left[\left(\frac{1}{R_0}\frac{\partial\delta}{\partial\theta}\right)^2 + \left(\frac{1}{R_0\sin\theta}\frac{\partial\delta}{\partial\phi}\right)^2\right] R_0^2 \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi \end{aligned}$$

Now δ can be written in terms of the spherical harmonics defined in the previous subsection as

$$\delta = \sum_{l,m} \delta_{lm} \bar{Y}_l^m$$

where the δ_{lm} are time dependent coefficients still to be found. If this is substituted into the expression for the potential, the first integral is similar to the one

encountered in the previous subsection; it is given by the orthonormality of the spherical harmonics. However, the final term involves an integral of the form

$$I = \int\int \left[\frac{\partial \bar{Y}_l^m}{\partial \theta} \frac{\partial \bar{Y}_{\underline{l}}^{\underline{m}}}{\partial \theta} + \frac{1}{\sin^2\theta} \frac{\partial \bar{Y}_l^m}{\partial \phi} \frac{\partial \bar{Y}_{\underline{l}}^{\underline{m}}}{\partial \phi} \right] \sin\theta \, \mathrm{d}\theta \mathrm{d}\phi$$

This integral can be simplified by using the same theorem (A.267) used earlier for the kinetic energy. Just take $F = r^l \bar{Y}_l^m$ and $G = r^{\underline{l}} \bar{Y}_{\underline{l}}^{\underline{m}}$ and integrate over a sphere of unit radius. The theorem then produces an equality between a volume integral and a surface one. The surface integral can be evaluated using the orthonormality of the spherical harmonics. The volume integral can be integrated explicitly in the radial direction to produce a multiple of I above and a second term that can once more be evaluated using the orthonormality of the spherical harmonics. It is then seen that $I = 0$ unless $l = \underline{l}$ and $m = \underline{m}$ and then $I = l(l+1)$.

Putting it all together, the potential energy due to surface tension becomes

$$V_s = V_{s,0} + \sum_{l,m} \tfrac{1}{2}(l-1)(l+2)\sigma\delta_{lm}^2 \tag{A.269}$$

A.43.4 Energy due to Coulomb repulsion

The potential energy due to the Coulomb forces is tricky. You need to make sure that the derivation is accurate enough. What is needed is the change in potential energy when the radial position of the surface of the droplet changes from the spherical value $r = R_0$ to the slightly perturbed value $r = R_0 + \delta$. The change in potential energy must be accurate to the order of magnitude of δ^2.

Trying to write a six-dimensional integral for the Coulomb energy would be a mess. Instead, assume that the surface perturbation is applied in small increments, as a perturbation δ' that is gradually increased from zero to δ. Imagine that you start with the perfect sphere and cumulatively add thin layers of charged liquid $\mathrm{d}\delta'$ until the full surface perturbation δ is achieved. (With adding negative amounts $\mathrm{d}\delta'$ understood as removing charged liquid. At each stage, just as much liquid is removed as added, so that the total volume of liquid stays the same.) The change in potential energy due to addition of an infinitesimal amount of charge equals the amount of charge times the surface potential at the point where it is added.

The surface radius perturbation and its differential change can be written in terms of spherical harmonics:

$$\delta' = \sum_{l,m} \delta'_{lm} \bar{Y}_l^m \qquad \mathrm{d}\delta' = \sum_{l,m} \mathrm{d}\delta'_{lm} \bar{Y}_l^m$$

The amount of charge added per unit solid angle $\mathrm{d}\Omega = \sin\theta \mathrm{d}\theta \mathrm{d}\phi$ will be called γ'. It is given in terms of the charge density ρ_c and δ' as

$$\gamma' = \rho_c \left(\tfrac{1}{3}(R_0 + \delta')^3 - \tfrac{1}{3}R_0^3 \right)$$

To first approximation, the incremental amount of charge laid down per unit solid angle is

$$\mathrm{d}\gamma' \sim \rho_c R_0^2 \mathrm{d}\delta'$$

However, if $\mathrm{d}\gamma$ is written in terms of spherical harmonics,

$$\mathrm{d}\gamma' = \sum_{l,m} \mathrm{d}\gamma'_{lm} \bar{Y}_l^m \qquad \mathrm{d}\gamma'_{lm} \sim \rho_c R_0^2 \mathrm{d}\delta'_{lm}$$

then the coefficient $\mathrm{d}\gamma'_{00}$ is zero exactly, because the net volume, and hence the net charge, remains unchanged during the build-up process. (The spherical harmonic Y_0^0 is independent of angular position and gives the average; the average charge added must be zero if the net charge does not change.) The coefficient $\mathrm{d}\delta'_{00}$ is zero to good approximation, but not exactly.

Now the surface potential is needed for the deformed drop. There are two contributions to this potential: the potential of the original spherical drop and the potential of the layer δ' of liquid that has been laid down, (the removed liquid here counting as negative charge having been laid down.) For the spherical drop, to the needed accuracy

$$V_{0,\mathrm{surface}} \sim \frac{Ze}{4\pi\epsilon_0(R_0+\delta')} \sim \frac{Ze}{4\pi\epsilon_0 R_0} - \frac{Ze}{4\pi\epsilon_0 R_0^2}\delta'$$

For the surface potential of the laid-down layer, fortunately only a leading order approximation is needed. That means that the thickness of the layer can be ignored. That turns it into a spherical shell of negligible thickness at $r = R_0$. The potential inside the shell can always be written in the form

$$V_{1,\mathrm{inside}} = \sum_{l,m} V_{lm} \frac{r^l}{R_0^l} \bar{Y}_l^m$$

though the coefficients V_{lm} are still unknown. The potential outside the shell takes the form

$$V_{1,\mathrm{outside}} = \sum_{l,m} V_{lm} \frac{R_0^{l+1}}{r^{l+1}} \bar{Y}_l^m$$

where the coefficients V_{lm} are approximately the same as those inside the shell because the shell is too thin for the potential to vary significantly across it.

However, the electric field strength does vary significantly from one side of the shell to the other, and it is that variation that determines the coefficients V_{lm}. First, integrate Maxwell's first equation over a small surface element of the shell. Since the shell has thickness δ', you get

$$\rho_c \delta' \,\mathrm{d}S = (\mathcal{E}_{r,\text{immediately outside}} - \mathcal{E}_{r,\text{immediately inside}})\,\mathrm{d}S$$

where $\mathrm{d}S$ is the area of the shell element. Note that $\mathcal{E}_r = -\partial V_1/\partial r$, and substitute in the inside and outside expressions for V_1 above, differentiated with respect to r and evaluated at $r = R_0$. That gives the V_{lm} and then

$$V_{1,\text{surface}} = \sum_{l,m} \frac{\rho_c R_0}{(2l+1)\epsilon_0} \delta'_{lm} \bar{Y}_l^m \qquad \rho_c = \frac{Ze}{\frac{4}{3}\pi R_0^3}$$

Multiplying the two surface potentials by the amount of charge laid down gives the incremental change in Coulomb potential of the drop as

$$\mathrm{d}V_c = \left[\frac{Ze}{4\pi\epsilon_0 R_0} - \frac{Ze}{4\pi\epsilon_0 R_0^2}\delta' + \sum_{l,m} \frac{3Ze}{(2l+1)4\pi\epsilon_0 R_0^2} \delta'_{l,m} \bar{Y}_l^m \right] \mathrm{d}\gamma' \sin\theta \, \mathrm{d}\theta \mathrm{d}\phi$$

Substituting in $\delta' = \sum_{l,m} \delta'_{lm} \bar{Y}_l^m$ and $\mathrm{d}\gamma' = \sum_{l,m} \mathrm{d}\gamma'_{\underline{l}\underline{m}} \bar{Y}_{\underline{l}}^{\underline{m}}$, the integrals can be evaluated using the orthonormality of the spherical harmonics. In particular, the first term of the surface potential integrates away since it is independent of angular position, therefore proportional to $\bar{Y}_0^0$, and $\mathrm{d}\gamma'_{00}$ is zero. For the other terms, it is accurate enough to set $\mathrm{d}\gamma'_{lm} = \rho_c R_0^2 \, \mathrm{d}\delta'_{lm}$ and then δ' can be integrated from zero to δ to give the Coulomb potential of the fully deformed sphere:

$$V_c = V_{c,0} - \sum_{l,m} \frac{l-1}{2l+1} \frac{Ze}{4\pi\epsilon_0} \rho_c \delta_{lm}^2 \tag{A.270}$$

A.43.5 Frequency of vibration

Having found the kinetic energy, (A.268), and the potential energy, (A.269) plus (A.270), the motion of the drop can be determined.

A rigorous analysis would so using a Lagrangian analysis, {A.1}. It would use the coefficients r_{lm} as generalized coordinates, getting rid of the δ_{lm} in the potential energy terms using (A.266). But this is a bit of an overkill, since the only thing that the Lagrangian analysis really does is show that each coefficient r_{lm} evolves completely independent of the rest.

If you are willing to take that for granted, just assume $r_{lm} = \varepsilon \sin(\omega t - \varphi)$ with ε and φ unimportant constants, and then equate the maximum kinetic energy to the maximum potential energy to get ω. The result is

$$\omega^2 = \frac{(l-1)l(l+2)}{3} \frac{C_s}{R_A^2 m_\mathrm{p}} \frac{1}{A} - \frac{2(l-1)l}{2l+1} \frac{e^2}{4\pi\epsilon_0 R_A^3 m_\mathrm{p}} \frac{Z^2}{A^2}$$

Note that this is zero if $l = 0$. There cannot be any vibration of a type $\delta = \delta_{00} Y_0^0$ because that would be a uniform radial expansion or compression of the drop, and its volume must remain constant. The frequency is also zero for $l = 1$. In that case, the potential energy does not change according to the derived expressions. If kinetic energy cannot be converted into potential energy, the

droplet must keep moving. Indeed, solutions for $l = 1$ describe that the droplet is translating at a constant speed without deformation. Vibrations occur for $l \geqslant 2$, and the most important ones are the ones with the lowest frequency, which means $l = 2$.

A.44 Relativistic neutrinos

It is certainly dubious to describe beta decay nonrelativistically. Neutrinos are highly relativistic particles with almost zero rest mass. So [16, p. 258] rightly states that it is absurd to treat them nonrelativistically. Then he immediately proceeds to do it anyway. (But he confirms the results later using a proper analysis.)

One big problem is that relativistically, spin and orbital angular momentum become mixed-up. That problem was encountered earlier for the photon, which is an even more relativistic particle. Chapter 7.4.3 needed some contortions to talk around that problem.

But note that while the problem for the neutrino is qualitatively similar, the details are quite different. While the photon is a boson, the neutrino is a fermion. Fermions, unlike bosons, are described by the Dirac equation, chapter 12.12 versus {A.21}.

To understand the Dirac equation requires some linear algebra, in particular matrices. But to understand the discussion here, the information in the notations section should be plenty.

Consider first the Dirac Hamiltonian eigenvalue problem for the electron and its antiparticle, the positron, in empty space. It is simplest thought of as a system of two equations:

$$\boxed{mc^2\vec{\psi}^- + c\widehat{\vec{p}} \cdot \vec{\sigma}\vec{\psi}^+ = E\vec{\psi}^- \qquad - mc^2\vec{\psi}^+ + c\widehat{\vec{p}} \cdot \vec{\sigma}\vec{\psi}^- = E\vec{\psi}^+} \tag{A.271}$$

Here c is the speed of light and m the rest mass of the electron or positron. So mc^2 is the rest mass energy according to Einstein's famous relation. Further $\widehat{\vec{p}}$ is the linear momentum operator. And $\vec{\sigma}$ is the spin angular momentum operator, except for a scale factor. More precisely, $\vec{\sigma} = \widehat{\vec{S}}/\frac{1}{2}\hbar$. The three components of $\vec{\sigma}$ are the famous "Pauli spin matrices."

In the nonrelativistic limit, $\vec{\psi}^-$ becomes the wave function of an electron. Recall from chapter 5.5.1 that this wave function can be thought of as a vector with two components; the first component corresponds to the spin-up state and the second to the spin-down state. Similarly $\vec{\psi}^+$ becomes the wave function of a positron.

The nonrelativistic limit means mathematically that the rest mass energies mc^2 are much larger than the kinetic energies. Or more simply, it is the limit $c \to \infty$. Under those conditions, by approximation,

$$mc^2\vec{\psi}^- \approx E\vec{\psi}^- \qquad - mc^2\vec{\psi}^+ \approx E\vec{\psi}^+$$

So in the nonrelativistic limit the energy of the electron is approximately the rest mass energy, as it should be. Note however that the value of E for a positron is negative. That should not be taken to mean that the rest mass energy of a positron is negative. It is the same as the one of an electron, positive. The positron is an antiparticle, and the value of E of an antiparticle picks up an extra minus sign because these particles move backward in time. Quantum mechanics relates E to the time derivative as $E \rightarrow \mathrm{i}\hbar\partial/\partial t$, as seen in the Schrödinger equation, so a sign change in E is equivalent to a reversal in time.

If you improve the accuracy of the nonrelativistic approximations a bit using some mathematical manipulation, the energy will also include the classical kinetic energy. To see how that works, (without using more proper matrix eigenvalue methods), premultiply the second equation in (A.271) by $\widehat{\vec{p}} \cdot \vec{\sigma}/2mc$ and add it to the first. In doing so, note that $(\widehat{\vec{p}} \cdot \vec{\sigma})^2$ is simply $\widehat{p}^2$. That can be verified by multiplying out the Pauli matrices given in chapter 12.10. In particular,

$$(\widehat{\vec{p}} \cdot \vec{\sigma})(\widehat{\vec{p}} \cdot \vec{\sigma}) = \sum_{i=1}^{3} \sum_{j=1}^{3} \widehat{p}_i \sigma_i \widehat{p}_j \sigma_j = \sum_{i=1}^{3} \widehat{p}_i \widehat{p}_i \quad \text{since} \quad \left\{ \begin{array}{l} \sigma_i \sigma_i = 1 \text{ for all } i \\ \sigma_i \sigma_j + \sigma_j \sigma_i = 0 \text{ if } i \neq j \end{array} \right.$$

Taking that into account, you get

$$\left(mc^2 + \frac{\widehat{p}^2}{2m} \right) \vec{\psi}^- + \frac{c\widehat{\vec{p}} \cdot \vec{\sigma}}{2} \vec{\psi}^+ = E \left(\vec{\psi}^- + \frac{c\widehat{\vec{p}} \cdot \vec{\sigma}}{2mc^2} \vec{\psi}^+ \right)$$

This expression can be rewritten as

$$\left(mc^2 + \frac{\widehat{p}^2}{2m} \right) \left(\vec{\psi}^- + \frac{c\widehat{\vec{p}} \cdot \vec{\sigma}}{2mc^2} \vec{\psi}^+ \right) = E \left(\vec{\psi}^- + \frac{c\widehat{\vec{p}} \cdot \vec{\sigma}}{2mc^2} \vec{\psi}^+ \right)$$

(If you multiply this out, you get an additional term, but it is negligibly small compared to the rest.) Now note that this is an eigenvalue problem for an electron whose wave function has picked up a little bit of $\vec{\psi}^+$. You might say that the slightly relativistic electron picks up a bit of a nonrelativistic positron wave function. Also its energy has picked up an additional term $\widehat{p}^2/2m$, the classical kinetic energy.

Derivation {D.81} shows how take this further, and do it rigorously using proper linear algebra procedures.

But this addendum is not really interested in the nonrelativistic limit. Instead it is interested in the ultrarelativistic limit, where it is the rest mass that is negligible compared to the kinetic energy. And for zero rest mass, the Dirac eigenvalue problem (A.271) becomes

$$c\widehat{\vec{p}} \cdot \vec{\sigma} \vec{\psi}^+ = E \vec{\psi}^- \qquad c\widehat{\vec{p}} \cdot \vec{\sigma} \vec{\psi}^- = E \vec{\psi}^+ \tag{A.272}$$

To make some sense out of this, define two new partial wave functions as

$$\vec{\psi}_{\rm R} = \frac{\vec{\psi}^- + \vec{\psi}^+}{2} \qquad \vec{\psi}_{\rm L} = \frac{\vec{\psi}^- - \vec{\psi}^+}{2}$$

By taking sums and differences of the ultrarelativistic equations above, you then get

$$\boxed{c\widehat{\vec{p}} \cdot \vec{\sigma}\vec{\psi}_{\rm R} = E\vec{\psi}_{\rm R} \qquad -c\widehat{\vec{p}} \cdot \vec{\sigma}\vec{\psi}_{\rm L} = E\vec{\psi}_{\rm L}} \tag{A.273}$$

The mathematician Weyl first noted how remarkable these equations really are. The equations for the two partial wave functions are not coupled! Partial wave function $\vec{\psi}_{\rm L}$ is not in the equation for $\vec{\psi}_{\rm R}$ and $\vec{\psi}_{\rm R}$ is not in the one for $\vec{\psi}_{\rm L}$. So it seems like each equation might describe a particle whose wave function is a simple two-dimensional vector. In other words, each equation could describe a particle without a partner. Note incidentally that the second equation is the negative E version of the first. So presumably the second equation could merely describe the antiparticle of the particle of the first equation.

Reasonably speaking, these particles should be electrically neutral. After all, their wave functions are equal combinations of $\vec{\psi}^-$, the equivalent of the negatively charged nonrelativistic electron in this system, and $\vec{\psi}^+$, the equivalent of the positively charged nonrelativistic positron. So these particles should be neutral as well as massless. In short, they should be neutrinos.

There are a couple of problems however. The Dirac equation includes both the electron and the positron, the time-reversed electron. If you write an equation for a particle that does not inherently include a time-reversed partner, are you violating time-reversal symmetry?

The other problem is what happens if you look at nature in the mirror. Or rather, what you really want to do is "inversion:" swap the positive direction of all three axes in a Cartesian coordinate system. This has the effect of creating a mirror image of nature, since the coordinate system is now left-handed instead of right-handed. And its effects are mathematically easier to describe.

First consider what happens to the original Dirac equation (A.271). The linear momentum operator $\widehat{\vec{p}}$ introduces a minus sign under inversion. That is because this operator is proportional to the gradient operator ∇, and every Cartesian coordinate in ∇ changes sign. However, the spin operators $\vec{\sigma}$ are like orbital angular momentum, $\vec{r} \times \widehat{\vec{p}}$, so they introduce two minus signs, which means no sign change.

(These arguments as presented look at the dot product purely algebraically. If you consider the gradient as a vector, there is an additional sign change since the unit vectors change direction. But the same holds for the $\vec{\sigma}$ vector in the dot product, so there is no difference in the final answer.)

At face value then, the $\widehat{\vec{p}} \cdot \vec{\sigma}$ terms in the Dirac system change sign. That would mean that nature does not behave in the same way when seen in the mirror; the Dirac equation is not the same. And that is definitely wrong; it is

very accurately established that electromagnetic systems involving electrons *do* behave in exactly the same way when seen in the mirror.

Fortunately there is a loophole to save the mathematics of the Dirac equation. It can be assumed that either the electron or the positron has negative "intrinsic" parity. In other words it can be assumed that either nonrelativistic wave function changes sign under inversion. You can check that if either $\vec{\psi}^-$ or $\vec{\psi}^+$ changes sign under inversion, then the Dirac system stays the same under inversion. Which one of the two changes sign is not important, since the sign of the wave function does not make a difference for the physics. It is convention to assume that the antiparticles of fermions have the negative intrinsic parity. In that case, you do not need to worry about intrinsic parity if you have no antiparticles present.

But now reconsider the Hamiltonian eigenvalue problem (A.273) for the neutrinos. The trick no longer works! If you can have one of these particles by itself, not tied to a partner, the physics is no longer the same when seen in the mirror.

There is an important consequence to that. If nature is the same when seen in the mirror, there is a conserved mathematical quantity called "parity." That is just like there is a conserved quantity called angular momentum because nature behaves in the same way when you look at it in a rotated coordinate system, chapter 7.3. What it boils down to is that Weyl was saying that parity might not be conserved in nature.

At the time, Pauli lambasted Weyl for daring to suggest something so obviously stupid like that. However, Pauli lived long enough to learn in 1956 that experiments showed that indeed neutrinos do not conserve parity. (Weyl had already died.)

The fact that the Weyl Hamiltonians do not conserve parity can be described more abstractly. Let Π be the "parity operator" that expresses what happens to a wave function when the axes are inverted. The above discussion can then be summarized abstractly as

$$\Pi(c\widehat{\vec{p}}\cdot\vec{\sigma})\vec{\psi} = -(c\widehat{\vec{p}}\cdot\vec{\sigma})\Pi\vec{\psi}$$

In words, when you invert axes, the $\widehat{\vec{p}}\cdot\vec{\sigma}$ operator introduces a minus sign. Also the wave function $\vec{\psi}$ changes into its mirror image. Now the expression in parentheses is the first Weyl Hamiltonian, or minus the second. So the expression above can be read that if you interchange the order of the parity operator and either Hamiltonian, it introduces a minus sign. It is said that the parity operator and the Hamiltonian do not commute: the order in which they are applied makes a difference.

It is quite generally true that if an operator does not commute with the Hamiltonian, the corresponding quantity is not preserved. See for example chapter 4.5.1, 7.1.4 and {A.19}. So you might ask whether orbital angular momentum is preserved by the Weyl Hamiltonian. If you grind out the commutator

of an orbital angular momentum component with the Weyl Hamiltonian using the rules of chapters 4.5.4 and 5.5.3, you find that you get something nonzero. So orbital angular momentum is not conserved by the Weyl neutrinos. That was no big surprise to physicists, because the Dirac equation does not conserve orbital angular momentum either. Similarly, if you work out the commutator of a component of spin, you find that spin is not preserved either.

However, if you add the two operators, you get the net angular momentum, orbital plus spin. That operator *does* commute with the Weyl Hamiltonians. So the Weyl Hamiltonians do conserve net angular momentum, like the Dirac equation does. That is very fortunate, because without doubt Pauli would have had a fit if Weyl had suggested that nature does not conserve net angular momentum.

But relativity does mix up orbital angular momentum with spin. Often this book describes particles as being in states of specific orbital angular momentum with specific values of the spin. For example, that is how the nonrelativistic hydrogen atom was described, chapter 4.3. Such a description is not really justified for particles at relativistic speeds. Fortunately the electron in a hydrogen atom has a kinetic energy much less than its rest mass energy. Things are much worse in beta decay, where the neutrino is emitted at almost the speed of light.

You might ask what else the Weyl Hamiltonians commute with. It can be seen that they commute with the linear momentum operators: different linear momentum operators commute because basically they are just derivatives, and spin commutes with nonspin. Different spin components do not commute, but everything commutes with itself. Since the Hamiltonian only involves the spin component in the direction of the linear momentum, (because of the dot product), the Hamiltonian commutes with the spin in that direction.

It is instructive to form a mental picture of these neutrinos. According to the above, the neutrinos can be in states with definite linear momentum and with definite spin in the direction of that linear momentum. In those states the neutrinos move in a specific direction and also rotate around an axis in the same direction. You can think of it macroscopically as screws; screws too rotate around their axis of motion while they move in or out of their hole.

Now there are two kinds of screws. Normal, "right-handed," screws move into the hole if you rotate them clockwise with a screwdriver. Some special applications use "left-handed" screws, which are machined so that they move into the hole when you rotate them counter-clockwise. (One of the screws that keep the pedals on a bicycle is normally a left-handed screw, to prevent you from loosening it up when pedaling.)

Experiments show that all antineutrinos observed in a normal lab setting act like right-handed screws. Neutrinos act like left-handed screws. Since a right-handed screw turns into a left-handed screw when seen in a mirror, nature is not the same when seen in a mirror. If, say, a beta decay is observed in a mirror, the antineutrino that comes out is left-handed, rather than right-handed

as it should.

(Using arguments like in {D.70}, it can be seen that $\vec{\psi}_{\rm R}$ above is righthanded and $\vec{\psi}_{\rm L}$ left handed, [53, p. 95].)

Note that theoretical physicists do not know about screws. So they came up with a different way to express that antineutrinos are right-handed like normal screws, while neutrinos are left-handed. They call the component of spin in the direction of motion, scaled so that its values are ± 1, the "helicity." So according to theoretical physicists, antineutrinos as seen in a lab have helicity 1, while neutrinos have helicity -1.

It is ironic that Pauli in turn did not live long enough to learn that Weyl's neutrinos were wrong on at least one major count. Around 1998, experiments found that neutrinos and antineutrinos are not massless as was long thought. Their mass is extremely tiny, (far too tiny to be accurately measured by experimental methods available at the time of writing), but it is not zero.

Fundamentally, it makes a big difference, because then their speed must be less than the speed of light. Hypothetically, an observer can then move with a speed faster than some antineutrino produced in a lab. The antineutrino then seems to go backward for that observer. Since the spin is still in the same direction, the observer sees a left-handed antineutrino, not a right-handed one. So the handedness, or helicity, of antineutrinos, and similarly neutrinos, is not fundamental. A process like the beta decay of an nucleus moving at a speed extremely close to the speed of light could produce a left-handed antineutrino trailing the nucleus. But since this does not happen for a nucleus at rest, nature is still not the same when seen in the mirror.

At the time of writing, the precise nature of neutrinos is not yet fully understood. Usually, it is assumed that neutrinos are distinct from antineutrinos. Neutrinos for which that is true are called Dirac neutrinos. However, so-called Majorana neutrinos would be their own antiparticles; neutrinos and antineutrinos would simply differ in the spin state.

A.45 Fermi theory

This note needs more work, but as far as I know is basically OK. Unfortunately, a derivation of electron capture for zero spin transitions is not included.

This note derives the Fermi theory of beta decay. In particular, it gives the ballparks that were used to create figure 14.54. It also describes the Fermi integral plotted in figure 14.52, as well as Fermi's (second) golden rule. There is also a final subsection on electron capture, A.45.7.

When beta decay was first observed, it was believed that the nucleus simply ejected an electron. However, problems quickly arose with energy and momentum conservation. To solve them, Pauli proposed in 1931 that in addition to the electron, also a neutral particle was emitted. Fermi called it the "neutrino," for

"small neutral one." Following ideas of Pauli in 1933, Fermi in 1934 developed a comprehensive theory of beta decay. The theory justifies the various claims made about allowed and forbidden beta decays. It also allows predictions of the decay rate and the probability that the electron and antineutrino will come out with given kinetic energies. This note gives a summary. The ballparks as described in this note are the ones used to produce figure 14.54.

A large amount of work has gone into improving the accuracy of the Fermi theory, but it is outside the scope of this note. To get an idea of what has been done, you might start with [23] and work backwards. One point to keep in mind is that the derivations below are based on expanding the electron and neutrino wave functions into plane waves, waves of definite linear momentum. For a more thorough treatment, it may be a better idea to expand into spherical waves, because nuclear states have definite angular momentum. That idea is worked out in more detail in the note on gamma decay, {A.25}. *That is news to the author. But it was supposed to be there, I think.*

A.45.1 Form of the wave function

A classical quantum treatment will not do for beta decay. To see why, note that in a classical treatment the wave function state before the decay is taken to be of the form

$$\psi_1(\vec{r}_1, S_{z,1}, \vec{r}_2, S_{z,2}, \ldots, \vec{r}_A, S_{z,A})$$

where 1 through A number the nucleons. However, the decay creates an electron and a antineutrino out of nothing. Therefore, after the decay the classical wave function is of the form

$$\psi_2(\vec{r}_1, S_{z,1}, \vec{r}_2, S_{z,2}, \ldots, \vec{r}_A, S_{z,A}, \vec{r}_e, S_{z,e}, \vec{r}_{\bar{\nu}}, S_{z,\bar{\nu}})$$

There is no way to describe how ψ_1 could evolve into ψ_2. You cannot just scribble in two more arguments into a function somewhere half way during the evolution. That would be voodoo mathematics. And there is also a problem with one nucleon turning from a neutron into a proton. You should really cross out the argument corresponding to the old neutron, and write in an argument for the new proton.

You might think that maybe the electron and antineutrino were always there to begin with. But that has some major problems. A lone neutron falls apart into a proton, an electron and an antineutrino. So supposedly the neutron would consist of a proton, an electron, and an antineutrino. But to confine light particles like electrons and neutrinos to the size of a nucleon would produce huge kinetic energies. According to the Heisenberg uncertainty relation $p \sim \hbar/\Delta x$, where the energy for relativistic particles is about pc, so the kinetic energy of a light particle confined to a 1 fm range is about 200 MeV. What conceivable force could be strong enough to hold electrons and neutrinos that hot? And how

come the effects of this mysterious force never show up in the *atomic* electrons that *can* be very accurately observed? How come that electrons come out in beta decays with only a few MeV, rather than 200 MeV?

Further, a high-energy antineutrino can react with a proton to create a neutron and a positron. That neutron is supposed to consist of a proton, an electron, and an antineutrino. So, following the same reasoning as before, the original proton before the reaction would consist of a positron, an electron, and a *proton*. That proton in turn would supposedly also consist of a positron, an electron, and an proton. So the original proton consists of a positron, an electron, a positron, an electron, and a proton. And so on until a proton consists of a proton and infinitely many electron / positron pairs. Not just one electron with very high kinetic energy would need to be confined inside a nucleon, but an infinite number of them, and positrons to boot. And all these electrons and positrons would somehow have to be prevented from annihilating each other.

It just does not work. There is plenty of solid evidence that neutrons and protons each contain three quarks, *not* other nucleons along with electrons, positrons, and neutrinos. The electron and antineutrino are created out of pure energy during beta decay, as allowed by Einstein's famous relativistic expression $E = mc^2$. A relativistic quantum treatment is therefore necessary.

In particular, it is necessary to deal mathematically with the appearance of the electron and an antineutrino out of nothing. To do so, a more general, more abstract way must be used to describe the states that nature can be in. Consider a decay that produces an electron and an antineutrino of specific momenta $\vec{p}_e$, respectively $\vec{p}_{\bar{\nu}}$. The final state is written as

$$\psi_2 = \psi_{2,\text{nuc}} |1e, \vec{p}_e\rangle |1\bar{\nu}, \vec{p}_{\bar{\nu}}\rangle \tag{A.274}$$

where $\psi_{2,\text{nuc}}$ is the nuclear part of the final wave function. The electron ket $|1e, \vec{p}_e\rangle$ is a "Fock-space ket," and should be read as "one electron in the state with angular momentum $\vec{p}_e$." The antineutrino ket should be read as "one antineutrino in the state with angular momentum $\vec{p}_{\bar{\nu}}$."

Similarly, the state before the decay is written as

$$\psi_1 = \psi_{1,\text{nuc}} |0e, \vec{p}_e\rangle |0\bar{\nu}, \vec{p}_{\bar{\nu}}\rangle \tag{A.275}$$

where $|0e, \vec{p}_e\rangle$ means "zero electrons in the state with angular momentum $\vec{p}_e$," and similar for the antineutrino ket. Written in this way, the initial and final wave functions are no longer inconsistent. What is different is not the *form* of the wave function, but merely how many electrons and antineutrinos are in the states with momentum $\vec{p}_e$, respectively $\vec{p}_{\bar{\nu}}$. Before the decay, the "occupation numbers" of these states are zero electrons and zero antineutrinos. After the decay, the occupation numbers are one electron and one neutrino. It is not that the initial state does not *have* occupation numbers for these states, (which would make ψ_1 and ψ_2 inconsistent), but merely that these occupation numbers have the value zero, (which does not).

(You could also add kets for different momentum states that the final electron and antineutrino are *not* in after the decay. But states that have zero electrons and neutrinos both before and after the considered decay are physically irrelevant and can be left away.)

That leaves the nuclear part of the wave function. You could use Fock-space kets to deal with the disappearance of a neutron and appearance of a proton during the decay. However, there is a neater way. The total number of nucleons remains the same during the decay. The only thing that happens is that a nucleon changes type from a neutron into a proton. The mathematical trick is therefore to take the particles to be nucleons, instead of protons and neutrons. If you give each nucleon a "nucleon type" property, then the only thing that happens during the decay is that the nucleon type of one of the nucleons flips over from neutron to proton. No nucleons are created or destroyed. Nucleon type is typically indicated by the symbol T_3 and is *defined* to be $\frac{1}{2}$ if the nucleon is a proton and $-\frac{1}{2}$ if the nucleon is a neutron. (Some older references may define it the other way around.) The general form of the nuclear wave function therefore becomes

$$\Psi_N(\vec{r}_1, S_{z,1}, T_{3,1}, \vec{r}_2, S_{z,2}, T_{3,2}, \ldots, \vec{r}_A, S_{z,A}, T_{3,A}; t)$$

During the decay, the T_3 value of one nucleon will change from $-\frac{1}{2}$ to $\frac{1}{2}$.

Of course, the name "nucleon type" for T_3 is not really acceptable, because it is understandable. In the old days, the names "isobaric spin" or "isotopic spin" were used, because nucleon type has absolutely nothing to do with spin. However, it was felt that these nonsensical names could cause some smart outsiders to suspect that the quantity being talked about was not really spin. Therefore the modern term "isospin" was introduced. This term contains nothing to give the secret away that it is not spin at all.

A.45.2 Source of the decay

Next the source of the decay must be identified. Ultimately that must be the Hamiltonian, because the Hamiltonian describes the time evolution of systems according to the Schrödinger equation.

In a specific beta decay process, two states are involved. A state ψ_1 describes the nucleus before the decay, and a state ψ_2 describes the combination of nucleus, electron, and antineutrino after the decay. That makes the system into a so-called "two state system." The unsteady evolution of such systems was discussed in chapter 7.6 and {D.38}. The key to the solution were the "Hamiltonian coefficients." The first one is:

$$E_1 \equiv H_{11} \equiv \langle \psi_1 | H \psi_1 \rangle$$

where H is the (relativistic) Hamiltonian. The value of H_{11} is the expectation value of the energy E_1 when nature is in the state ψ_1. Assuming that the

nucleus is initially at rest, the relativistic energy is just the rest mass energy of the nucleus. It is given in terms of its mass by Einstein's famous relation $E_1 = m_{\mathrm{N1}}c^2$.

The Hamiltonian coefficient for the final state is similarly

$$E_2 \equiv H_{22} \equiv \langle \psi_2 | H \psi_2 \rangle$$

Using the form given in the previous section for the final wave function, that becomes

$$E_2 = \langle 1\bar{\nu}, \vec{p}_{\bar{\nu}} | \langle 1e, \vec{p}_e | \psi_{2,\mathrm{nuc}} | H \psi_{2,\mathrm{nuc}} | 1e, \vec{p}_e \rangle | 1\bar{\nu}, \vec{p}_{\bar{\nu}} \rangle$$

It is the expectation value of energy after the decay. It consists of the sum of the rest mass energies of final nucleus, electron, and antineutrino, as well as their kinetic energies.

The Hamiltonian coefficient that describes the interaction between the two states is crucial, because it is the one that causes the decay. It is

$$H_{21} \equiv \langle \psi_2 | H \psi_1 \rangle$$

Using the form for the wave functions given in the previous section:

$$H_{21} = \langle 1\bar{\nu}, \vec{p}_{\bar{\nu}} | \langle 1e, \vec{p}_e | \psi_{2,\mathrm{nuc}} | H \psi_{1,\mathrm{nuc}} | 0e, \vec{p}_e \rangle | 0\bar{\nu}, \vec{p}_{\bar{\nu}} \rangle$$

If H_{21} is zero, no decay will occur. And most of the Hamiltonian does not produce a contribution to H_{21}. But there is a small part of the Hamiltonian, call it H', that does produce a nonzero interaction. That part is due to the weak force.

Unfortunately, Fermi had no clue what H' was. He assumed that beta decay would not be that much different from the better understood decay of excited atomic states in atoms. Gamma decay is the direct equivalent of atomic decay for excited nuclei. Beta decay is definitely different, but maybe not that different. In atomic decay an electromagnetic photon is created, rather than an electron and antineutrino. Still the general idea seemed similar.

In atomic decay H' is essentially proportional to the product of the charge of the excited electron, times the spatial eigenstate of the photon, times a "photon creation" operator $\widehat{a}^\dagger$:

$$H' \propto e \psi_{\mathrm{photon}}(\vec{r}) \widehat{a}^\dagger$$

In words, it says that the interaction of the electron with the electromagnetic field can create photons. The magnitude of that effect is proportional to the amplitude of the photon at the location of the electron, and also to the electric charge of the electron. The electric charge acts as a "coupling constant" that links electrons and photons together. If the electron was uncharged, it would not be able to create photons. So it would not be able to create an electric field. Further, the fact that the coupling between the electron and the photon occurs

at the location of the electron eliminates some problems that relativity has with action at a distance.

There is another term in H' that involves an annihilation operator $\widehat{a}$ instead of a creation operator. An annihilation operator destroys photons. However, that does not produce a contribution to H_{21}; if you try to annihilate the nonexisting photon in the initial wave function, you get a zero wave function. On the other hand for the earlier term, the creation operator is essential. It turns the initial state with no photon into a state with one photon. States with different numbers of particles are orthogonal, so the Hamiltonian coefficient H_{12} would be zero without the creation operator. Looked at the other way around, the presence of the creation operator in the Hamiltonian ensures that the final state must have one more photon for the decay to occur. (See addendum {A.15} for more details on electromagnetic interactions, including a more precise description of H'. See also {A.25}.)

Fermi assumed that the general ideas of atomic decay would also hold for beta decay of nuclei. Electron and antineutrino creation operators in the Hamiltonian would turn the zero-electron and zero-antineutrino kets into one-electron and one-antineutrino ones. Then the inner products of the kets are equal to one pairwise. Therefore both the creation operators and the kets drop out of the final expression. In that way the Hamiltonian coefficient simplifies to

$$H_{21} = \langle \psi_{2,\text{nuc}} | \sum_{i=1}^{A} g h_i \psi_{e,\vec{p}_e}(\vec{r}_i) \psi_{\bar{\nu},\vec{p}_{\bar{\nu}}}(\vec{r}_i) | \psi_{1,\text{nuc}} \rangle$$

where the index i is the nucleon number and gh_i is the remaining still unknown part of the Hamiltonian. In indicating this unknown part by gh_i, the assumption is that it will be possible to write it as some generic dimensional constant g times some simple nondimensional operator h_i acting on nucleon number i.

To write expressions for the wave functions of the electron and antineutrino, you face the complication that unbound states in infinite space are not normalizable. That produced mathematical complications for momentum eigenstates in chapter 7.9.2, and similar difficulties resurface here. To simplify things, the mathematical trick is to assume that the decaying nucleus is not in infinite space, but in an extremely large "periodic box." The assumption is that nature repeats itself spatially; a particle that exits the box through one side reenters it through the opposite side. Space "wraps around" if you want, and opposite sides of the box are assumed to be physically the same location. It is like on the surface of the earth: if you move along the straightest-possible path on the surface of the earth, you travel around the earth along a big circle and return to the same point that you started out at. Still, on a local scale, the surface on the earth looks flat. The idea is that the empty space around the decaying nucleus has a similar property, in each of the three Cartesian dimensions. This trick is also commonly used in solid mechanics, chapter 10.

In a periodic box, the wave function of the antineutrino is

$$\psi_{\bar\nu,\vec p_{\bar\nu}} = \frac{1}{\sqrt{\mathcal{V}}} e^{\mathrm{i}\vec p_{\bar\nu}\cdot\vec r/\hbar} = \frac{1}{\sqrt{\mathcal{V}}} e^{\mathrm{i}(p_{x,\bar\nu}x + p_{y,\bar\nu}y + p_{z,\bar\nu}z)/\hbar}$$

where $\mathcal{V}$ is the volume of the periodic box. It is easy to check that this wave function is indeed normalized. Also, it is seen that it is indeed an eigenfunction of the x-momentum operator $\hbar\partial/\mathrm{i}\partial x$ with eigenvalue p_x, and similar for the y- and z-momentum operators.

The wave function of the electron will be written in a similar way:

$$\psi_{e,\vec p_e}(\vec r) = \frac{1}{\sqrt{\mathcal{V}}} e^{\mathrm{i}\vec p_e\cdot\vec r/\hbar} = \frac{1}{\sqrt{\mathcal{V}}} e^{\mathrm{i}(p_{x,e}x + p_{y,e}y + p_{z,e}z)/\hbar}$$

This however has an additional problem. It works fine far from the nucleus, where the momentum of the electron is by definition the constant vector $\vec p_e$. However, near the nucleus the Coulomb field of the nucleus, and to some extent that of the atomic electrons, affects the kinetic energy of the electron, and hence its momentum. Therefore, the energy eigenfunction that has momentum $\vec p$ far from the nucleus differs significantly from the above exponential closer to the nucleus. And this wave function must be evaluated at nucleon positions inside the nucleus! The problem is particularly large when the momentum $\vec p_e$ is low, because then the electron has little kinetic energy and the Coulomb potential is relatively speaking more important. The problem gets even worse for low-energy positron emission, because a positively-charged positron is repelled by the positive nucleus and must tunnel through to reach it.

The usual way to deal with the problem is to stick with the exponential electron wave function for now, and fix up the problem later in the final results. The fix-up will be achieved by throwing in an additional fudge factor. While "Fermi fudge factor" alliterates nicely, it does not sound very respectful, so physicists call the factor the "Fermi function."

The bottom line is that for now

$$H_{21} = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\mathrm{nuc}} | \sum_{i=1}^{A} h_i e^{\mathrm{i}(\vec p_e + \vec p_{\bar\nu})\cdot\vec r_i/\hbar} \psi_{1,\mathrm{nuc}} \rangle \tag{A.276}$$

That leaves the still unknown operator gh_i. The constant g is simply *defined* so that the operator h_i has a magnitude that is of order one. That means that $\langle\psi_{2,\mathrm{nuc}}|h_i\psi_{1,\mathrm{nuc}}\rangle$ should never be greater than about one, though it could be much less if $\psi_{2,\mathrm{nuc}}$ and $h_i\psi_{1,\mathrm{nuc}}$ turn out to be almost orthogonal. It is found that g has a rough value of about 100 eV fm^3, depending a bit on whether it is a Fermi or Gamow-Teller decay. Figure 14.54 simply used 100 MeV fm^3.

A.45.3 Allowed or forbidden

The question of allowed and forbidden decays is directly related to the Hamiltonian coefficient H_{21}, (A.276), derived in the previous subsection, that causes the decay.

First note that the emitted electron and antineutrino have quite small momentum values, on a nuclear scale. In particular, in their combined wave function

$$\frac{1}{\mathcal{V}} e^{\mathrm{i}(\vec{p}_e+\vec{p}_{\bar{\nu}})\cdot\vec{r}_i/\hbar}$$

the argument of the exponential is small. Typically, its magnitude is only a few percent, It is therefore possible to approximate the exponential by one, or more generally by a Taylor series:

$$e^{\mathrm{i}(\vec{p}_e+\vec{p}_{\bar{\nu}})\cdot\vec{r}_i/\hbar} \approx 1 + \frac{\mathrm{i}(\vec{p}_e+\vec{p}_{\bar{\nu}})\cdot\vec{r}_i}{\hbar} + \frac{1}{2!}\left(\frac{\mathrm{i}(\vec{p}_e+\vec{p}_{\bar{\nu}})\cdot\vec{r}_i}{\hbar}\right)^2 + \ldots$$

Since the first term in the Taylor series is by far the largest, you would expect that the value of H_{21}, (A.276), can be well approximated by replacing the exponential by 1, giving:

$$H^0_{21} = \frac{g}{\sqrt{\mathcal{V}}}\langle\psi_{2,\mathrm{nuc}}|\sum_{i=1}^{A} h_i\psi_{1,\mathrm{nuc}}\rangle$$

However, clearly this approximation does not work if the value of H^0_{21} is zero for some reason:

> *If the simplified coefficient H^0_{21} is nonzero, the decay is allowed. If it is zero, the decay is forbidden.*

If the decay is forbidden, higher order terms in the Taylor series will have to be used to come up with a nonzero value for H_{21}. Since these higher order terms are much smaller, and H_{21} drives the decay, a forbidden decay will proceed much slower than an allowed one.

Why would a decay not be allowed? In other words why would $\psi_{2,\mathrm{nuc}}$ and $h_i\psi_{1,\mathrm{nuc}}$ be *exactly* orthogonal? If you took two random wave functions for $\psi_{2,\mathrm{nuc}}$ and $\psi_{1,\mathrm{nuc}}$, they definitely would not be. But $\psi_{2,\mathrm{nuc}}$ and $\psi_{1,\mathrm{nuc}}$ are not random wave functions. They satisfy a significant amount of symmetry constraints.

One very important one is symmetry with respect to coordinate system orientation. An inner product of two wave functions is independent of the angular orientation of the coordinate system in which you evaluate it. Therefore, you can average the inner product over all directions of the coordinate system. However, the angular variation of a wave function is related to its angular momentum; see chapter 7.3 and its note. In particular, if you average a wave function of definite angular momentum over all coordinate system orientations, you get zero unless

the angular momentum is zero. So, if it was just $\psi_{1,\mathrm{nuc}}$ in the inner product in H^0_{21}, the inner product would be zero unless the initial nucleus had zero spin. However, the final state is also in the inner product, and being at the other side of it, its angular variation acts to counteract that of the initial nucleus. Therefore, H^0_{21} will be zero unless the initial angular momentum is exactly balanced by the net final angular momentum. And that is angular momentum conservation. The decay has to satisfy it.

Note that the linear momenta of the electron and antineutrino have become ignored in H^0_{21}. Therefore, their orbital angular momentum is approximated to be zero too. Under these condition H^0_{21} is zero unless the angular momentum of the final nucleus plus the spin angular momentum of electron and antineutrino equals the angular momentum of the original nucleus. Since the electron and antineutrino can have up to one unit of combined spin, the nuclear spin cannot change more than one unit. That is the first selection rule for allowed decays given in chapter 14.19.6.

Another important constraint is symmetry under the parity transformation $\vec{r} \rightarrow -\vec{r}$. This transformation too does not affect inner products, so you can average the values before and after the transform. However, a wave function that has odd parity changes sign under the transform and averages to zero. So the inner product in H^0_{21} is zero if the total integrand has odd parity. For a nonzero value, the integrand must have even parity, and that means that the parity of the initial nucleus must equal the combined parity of the final nucleus electron, and antineutrino.

Since the electron and antineutrino come out without orbital angular momentum, they have even parity. So the nuclear parity must remain unchanged under the transition. (To be sure, this is not absolutely justified. Nuclear wave functions actually have a tiny uncertainty in parity because the weak force does not conserve parity, chapter 14.19.8. This effect is usually too small to be observed and will be ignored here.)

So what if either one of these selection rules is violated? In that case, maybe the second term in the Taylor series for the electron and antineutrino wave functions produces something nonzero that can drive the decay. For that to be true,

$$H^1_{21} = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\mathrm{nuc}} | \sum_{i=1}^{A} h_i \frac{\mathrm{i}(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\hbar} \psi_{1,\mathrm{nuc}} \rangle$$

has to be nonzero. If it is, the decay is a first-forbidden one. Now the spherical harmonics Y_1^m of orbital angular momentum are of the generic form, {D.14}

$$r Y_1^m = \sum_j c_j r_j$$

with the c_j some constants. Therefore, the factor $\vec{r}_i$ in H^1_{21} brings in angular variation corresponding to one unit of angular momentum. That means that

the total spin can now change by up to one unit, and therefore the nuclear spin by up to two units. That is indeed the selection rule for first forbidden decays.

And note that because $\vec{r}_i$ changes sign when every $\vec{r}$ is replaced by $-\vec{r}$, the initial and final nuclear parities must now be opposite for H^1_{21} not to be zero. That is indeed the parity selection rule for first-forbidden decays.

The higher order forbidden decays go the same way. For an ℓth-forbidden decay,

$$H^\ell_{21} = \frac{g}{\ell!\sqrt{\mathcal{V}}} \langle \psi_{2,\mathrm{nuc}} | \sum_{i=1}^{A} h_i \left(\frac{\mathrm{i}(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\hbar} \right)^\ell \psi_{1,\mathrm{nuc}} \rangle \tag{A.277}$$

must be the first nonzero inner product. Note that an ℓth-forbidden decay has a coefficient H_{21} proportional to a factor of order $(pR/\hbar)^\ell$, with R the nuclear radius. Since the decay rate turns out to be proportional to $|H_{21}|^2$, an ℓth-forbidden decay is slowed down by a factor of order $(pR/\hbar)^{2\ell}$, making highly forbidden decays extremely slow.

A.45.4 The nuclear operator

This subsection will have a closer look at the nuclear operator h_i. While the discussion will be kept simple, having some idea about the nature of this operator can be useful. It can help to understand why some decays have relatively low decay rates that are not explained by just looking at the electron and antineutrino wave functions, and the nuclear spins and parities. The discussion will mainly focus on allowed decays.

Although Fermi did not know what h_i was, Pauli had already established the possible generic forms for it allowed by relativity. It could take the form of a scalar (S), a vector (V), an axial vector (A, a vector like angular momentum, one that inverts when the physics is seen in a mirror), a pseudoscalar (P, a scalar like the scalar triple product of vectors that changes sign when the physics is seen in the mirror), or a tensor (T, a multiple-index object like a matrix that transforms in specific ways.) Fermi simply assumed the interaction was of the vector, V, type in analogy with the decay of excited atoms.

Fermi ignored the spin of the electron and antineutrino. However, Gamow & Teller soon established that to allow for decays where the two come out with spin, (Gamow-Teller decays), gh_i also should have terms with axial, A, and/or tensor, T, character. Work of Fierz combined with experimental evidence showed that the Hamiltonian could not have both S and V, nor both A and T terms. Additional evidence narrowed h_i down to STP combinations or VA ones.

Finally, it was established in 1953 that the correct one was the STP combination, because experimental evidence on RaE, (some physicists cannot spell bismuth-210), showed that P was present. Unfortunately, it did not. For one, the conclusion depended to an insane degree on the accuracy of a correction term.

However, in 1955 it was established that it was STP anyway, because experimental evidence on helium-6 clearly showed that the Gamow-Teller part of the decay was tensor. The question was therefore solved satisfactorily. It was STP, or maybe just ST. Experimental evidence had redeemed itself.

However, in 1958, a quarter century after Fermi, it was found that beta decay violated parity conservation, chapter 14.19.8, and theoretically that was not really consistent with STP. So experimentalists had another look at their evidence and quickly came back with good news: "The helium-6 evidence does not show Gamow-Teller is tensor after all."

The final answer is that gh_i is VA. Since so much of our knowledge about nuclei depends on experimental data, it may be worthwhile to keep this cautionary tale, taken from the Stanford Encyclopedia of Philosophy, in mind.

It may next be noted that gh_i will need to include a isospin creation operator to be able to turn a neutron into a proton. In Fermi decays, h_i is assumed to be just that operator. The constant of proportionality g, usually called the coupling constant $g_{\rm F}$, describes the strength of the weak interaction. That is much like the unit electric charge e describes the strength of the electromagnetic interaction between charged particles and photons. In Fermi decays it is found that $g_{\rm F}$ is about 88 eV fm^3. Note that this is quite small compared to the MeV scale of nuclear forces. If you ballpark relative strengths of forces, [31, p. 285] the nuclear force is strongest, the electromagnetic force about hundred times smaller, the weak force another thousand times smaller than that, and finally gravity is another 10^{34} times smaller than that. The decay rates turn out to be proportional to the square of the interaction, magnifying the relative differences.

In Gamow-Teller decays, h_i is assumed to consist of products of isospin creation operators times spin creation or annihilation operators. The latter operators allow the spin of the neutron that converts to the proton to flip over. Suitable spin creation and annihilation operators are given by the so-called "Pauli spin matrices," chapter 12.10 When they act on a nucleon, they produce states with the spin in an orthogonal direction flipped over. That allows the net spin of the nucleus to change by one unit. The appropriate constant of proportionality g_{GT} is found to be a bit larger than the Fermi one.

The relevant operators then become, [5],

$$h_i = \tau_1 \pm {\rm i}\tau_2 \qquad h_i = (\tau_1 \pm {\rm i}\tau_2) \sum_{j=1}^{3} \sigma_j$$

for Fermi and Gamow-Teller decays respectively. Here the three σ_j are the Pauli spin matrices of chapter 12.10. The τ_i are the equivalents of the Pauli spin matrices for isospin; in the combinations shown above they turn neutrons into protons, or vice-versa. Please excuse: using the clarity now made possible by modern physical terminology, they create, respectively annihilate, isospin. The upper sign is relevant for beta-minus decay and the lower for beta-plus

decay. The Gamow-Teller operator absorbs the spin part of the electron and antineutrino wave functions, in particular the averaging over the directions of their spin.

So how do these nuclear operators affect the decay rate? That is best understood by going back to the more physical shell-model picture. In beta minus decay, a neutron is turned into a proton. That proton usually occupies a different spatial state in the proton shells than the original neutron in the neutron shells. And different spatial states are supposedly orthogonal, so the inner product $\langle \psi_{2,\mathrm{nuc}} | h_i \psi_{1,\mathrm{nuc}} \rangle$ will usually be pretty small, if the decay is allowed at all. There is one big exception, though: mirror nuclei. In a decay between mirror nuclei, a nucleus with a neutron number $N_1 = Z_1 \pm 1$ decays into one with neutron number $N_2 = Z_2 \mp 1$. In that case, the nucleon that changes type remains in the same spatial orbit. Therefore, the Fermi inner product equals one, and the Gamow Teller one is maximal too. Allowed decays of this type are called "superallowed." The simplest example is the beta decay of a free neutron.

If you allow for beta decay to excited states, more superallowed decays are possible. States that differ merely in nucleon type are called isobaric analog states, or isospin multiplets, chapter 14.18. There are about twenty such superallowed decays in which the initial and final nuclei both have spin zero and positive parity. These twenty are particularly interesting theoretically, because only Fermi decays are possible for them. And the Fermi inner product is $\sqrt{2}$. (The reason that it is $\sqrt{2}$ instead of 1 like for mirror nuclei can be seen from thinking of isospin as if it is just normal spin. Mirror nuclei have an odd number of nucleons, so the net nuclear isospin is half integer. In particular the net isospin will be $\frac{1}{2}$ in the ground state. However, nuclei with zero spin have an even number of nucleons, hence integer net isospin. The isospin of the twenty decays is one; it cannot be zero because at least one nucleus must have a nonzero net nucleon type $T_{3,\mathrm{net}}$. The net nucleon type is only zero if the number of protons is the same as the number of neutrons. It is then seen from (12.9) and (12.10) in chapter 12 that the isospin creation or annihilation operators will produce a factor $\sqrt{2}$.)

These decays therefore allow the value of the Fermi coupling constant g_{F} to be determined from the decay rates. It turns out to be about 88 eV fm^3, regardless of the particular decay used to compute it. That seems to suggest that the interaction with neighboring nucleons in a nucleus does not affect the Fermi decay process. Indeed, if the value of g_{F} is used to analyze the decay rates of the mirror nuclei, including the free neutron that has no neighbors, the data show no such effect. The hypothesis that neighboring nucleons do not affect the Fermi decay process is known as the "conserved vector current hypothesis." What name could be clearer than that? Unlike Fermi decays, Gamow-Teller decays are somewhat affected by the presence of neighboring nuclei.

Besides the spin and parity rules already mentioned, Fermi decays must satisfy the approximate selection rule that the magnitude of isospin must be

unchanged. They can be slowed down by several orders of magnitude if that rule is violated.

Gamow-Teller decays are much less confined than Fermi ones because of the presence of the electron spin operator. As the shell model shows, nucleon spins are uncertain in energy eigenstates. Therefore, the nuclear symmetry constraints are a lot less restrictive.

A.45.5 Fermi's golden rule

The previous four subsections have focussed on finding the Hamiltonian coefficients of the decay from a state ψ_1 to a state ψ_2. Most of the attention was on the coefficient H^ℓ_{21} that drives the decay. The next step is solution of the Schrödinger equation to find the evolution of the decay process.

The quantum amplitude of the pre-decay state ψ_1 will be indicated by $\bar{a}$ and the quantum amplitude of the final decayed state ψ_2 by $\bar{b}$. The Schrödinger equation implies that $\bar{b}$ increases from zero according to

$$\mathrm{i}\hbar\dot{\bar{b}} = H^\ell_{21} e^{\mathrm{i}(E_2-E_1)t/\hbar}\bar{a}$$

(To use this expression, the quantum amplitudes must include an additional phase factor, but it is of no consequence for the probability of the states. See chapter 7.6 and {D.38} for details.)

Now picture the following. At the initial time there are a large number of pre-decay nuclei, all with $\bar{a} = 1$. All these nuclei then evolve according to the Schrödinger equation, above, over a time interval t_c that is short enough that $\bar{a}$ stays close to one. (Because the perturbation of the nucleus by the weak force is small, the magnitudes of the coefficients only change slowly on the relevant time scale.) In that case, $\bar{a}$ can be dropped from the equation and its solution is then seen to be

$$\bar{b} = -H^\ell_{21}\frac{e^{\mathrm{i}(E_2-E_1)t_c/\hbar}-1}{(E_2-E_1)}$$

Half of the exponential can be factored out to produce a real ratio:

$$\bar{b} = -H^\ell_{21} e^{\mathrm{i}\frac{1}{2}(E_2-E_1)t_c/\hbar}\frac{\mathrm{i}}{\hbar}\frac{\sin\left(\frac{1}{2}(E_2-E_1)t_c/\hbar\right)}{\frac{1}{2}(E_2-E_1)t_c/\hbar}t_c$$

Then at the final time t_c, assume that the state of all the nuclei is "measured." The macroscopic surroundings of the nuclei establishes whether or not electron and antineutrino pairs have come out. The probability that a give nucleus has emitted such a pair is given by the square magnitude $|\bar{b}|^2$ of the amplitude of the decayed state. Therefore, a fraction $|\bar{b}|^2$ of the nuclei will be found to have decayed and $1-|\bar{b}|^2$ will be found to be still in the pre-decay state ψ_1. After this "measurement," the entire process then repeats for the remaining $1-|\bar{b}|^2$ nuclei that did not decay.

The bottom line is however that a fraction $|\bar{b}|^2$ did. Therefore, the ratio $|\bar{b}|^2/t_c$ gives the specific decay rate, the relative amount of nuclei that decay per unit time. Plugging in the above expression for $\bar{b}$ gives:

$$\lambda_{\text{single final state}} = \frac{|H_{21}^{\ell}|^2}{\hbar^2} \frac{\sin^2\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)}{\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)^2} t_c \tag{A.278}$$

To get the total decay rate, you must still sum over all possible final states. Most importantly, you need to sum the specific decay rates together for all possible electron and antineutrino momenta.

And there may be more. If the final nuclear state has spin you also need to sum over all values of the magnetic quantum number of the final state. (The amount of nuclear decay should not depend on the angular orientation of the initial nucleus in empty space. However, if you expand the electron and neutrino wave functions into spherical waves, you need to average over the possible initial magnetic quantum numbers. It may also be noted that the total coefficient $|H_{21}^{\ell}|$ for the decay $1 \to 2$ will not be the same as the one for $2 \to 1$: you average over the initial magnetic quantum number, but sum over the final one.) If there are different excitation levels of the final nucleus that can be decayed to, you also need to sum over these. And if there is more than one type of decay process going on at the same time, they too need to be added together.

However, all these details are of little importance in finding a ballpark for the dominant decay process. The real remaining problem is summing over the electron and antineutrino momentum states. The total ballparked decay rate must be found from

$$\lambda = \sum_{\text{all } \vec{p}_e, \vec{p}_{\bar{\nu}}} \frac{|H_{21}^{\ell}|^2}{\hbar^2} \frac{\sin^2\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)}{\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)^2} t_c$$

Based on energy conservation, you would expect that decays should only occur when the total energy E_2 of the nucleus, electron and antineutrino after the decay is exactly equal to the energy E_1 of the nucleus before the decay. However, the summation above shows that that is not quite true. For a final state that has E_2 exactly equal to E_1, the last fraction in the summation is seen to be unity, using l'Hospital. For a final state with an energy E_2 of, for example, $E_1 + \hbar/t_c$, the ratio is quite comparable. Therefore decay to such a state proceeds at a comparable rate as to a state that conserves energy exactly. There is "slop" in energy conservation.

How can energy not be conserved? The reason is that neither the initial state nor the final state is an energy eigenstate, strictly speaking. Energy eigenstates are stationary states. The very fact that decay occurs assures that these states are not really energy eigenstates. They have a small amount of uncertainty

in energy. The nonzero value of the Hamiltonian coefficient H^ℓ_{21} assures that, chapter 5.3, and there may be more decay processes adding to the uncertainty in energy. If there is some uncertainty in energy, then $E_2 = E_1$ is not an exact relationship.

To narrow this effect down more precisely, the fraction is plotted in figure 7.7. The spikes in the figure indicate the energies E_2 of the possible final states. Now the energy states are almost infinitely densely spaced, if the periodic box in which the decay is assumed to occur is big enough. And the box must be assumed very big anyway, to simulate decay in infinite space. Therefore, the summation can be replaced by integration, as follows:

$$\lambda = \int_{\text{all } E_2} \frac{|H^\ell_{21}|^2}{\hbar^2} \frac{\sin^2\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)}{\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)^2} t_c \frac{\mathrm{d}N}{\mathrm{d}E_2}\,\mathrm{d}E_2$$

where $\mathrm{d}N/\mathrm{d}E_2$ is the number of final states per unit energy range, often called the density of states $\rho(E_2)$.

Now assume that the complete problem is cut into bite-size pieces for each of which $|H^\ell_{21}|$ is about constant. It can then be taken out of the integral. Also, the range of energy in figure 7.7 over which the fraction is appreciable, the energy slop, is very small on a normal nuclear energy scale: beta decay is a slow process, so the initial and final states do remain energy eigenstates to a very good approximation. Energy conservation is almost exactly satisfied. Because of that, the density of states $\mathrm{d}N/\mathrm{d}E_2$ will be almost constant over the range where the integrand is nonzero. It can therefore be taken out of the integral too. What is left can be integrated analytically, [41, 18.36]. That gives:

$$\lambda = \frac{|H^\ell_{21}|^2}{\hbar^2} t_c \frac{\mathrm{d}N}{\mathrm{d}E} \frac{2\pi\hbar}{t_c}$$

That is "Fermi's (second) golden rule." It describes how energy slop increases the total decay rate. It is not specific to beta decay but also applies to other forms of decay to a continuum of states. Note that it no longer depends on the artificial length t_c of the time interval over which the system was supposed to evolve without "measurement." That is good news, since that time interval was obviously poorly defined.

Because of the assumptions involved, like dividing the problem into bite-size pieces, the above expression is not very intuitive to apply. It can be rephrased into a more intuitive form that does not depend on such an assumption. The obtained decay rate is exactly the same as if in an energy slop range

$$\Delta E_{\text{slop}} = \frac{2\pi\hbar}{t_c}$$

all states contribute just as much to the decay as one that satisfies energy conservation exactly, while no states contribute outside of that range.

(Note that if you ballpark t_c as the half-life, then even a half-life as short as 10^{-15} s gives an energy slop of a few eV, almost impossible to measure. And any nuclear decay with a half life of 10^{-15} s should surely produce an amount of energy equal to very many MeV. Outside the mathematics of the Fermi theory, the energy slop is imperceptible under normal conditions.)

The good news is that phrased this way, it indicates the relevant physics much more clearly than the earlier purely mathematical expression for Fermi's golden rule. The bad news is that it suffers esthetically from still involving the poorly defined time t, instead of already having shoved t under the mat. Therefore, it is more appealing to write things in terms of the energy slop altogether:

$$\boxed{\lambda_{\text{single final state}} = \frac{2\pi}{\hbar\varepsilon}|H^\ell_{21}|^2 \qquad \Delta E_{\text{slop}} \equiv \varepsilon \qquad \varepsilon t_c \sim 2\pi\hbar} \tag{A.279}$$

Here ε is the amount that energy conservation seems to be violated, and is related to a typical time t_c between collisions by the energy-time uncertainty relationship shown.

It may be noted that the golden rule does not apply if the evolution is not to a continuum of states. It also does not apply if the slop range ε is so large that $\mathrm{d}N/\mathrm{d}E$ is not constant over it. And it does not apply for systems that evolve without being perturbed over times long enough that the decay probability becomes significant before the system is "measured." (If $\bar{b}$ becomes appreciable, $\bar{a}$ can no longer be close to one since the probabilities $|\bar{a}|^2$ and $|\bar{b}|^2$ must add to one.) "Measurements," or rather interactions with the larger environment, are called "collisions." Fermi's golden rule applies to so-called "collision-dominated" conditions. Typically examples where the conditions are not collision dominated are in NMR and atomic decays under intense laser light.

Mathematically, the conditions for Fermi's golden rule can be written as

$$|H_{21}| \ll \varepsilon \ll E \qquad \varepsilon \equiv \frac{2\pi\hbar}{t_c} \tag{A.280}$$

The first inequality means that the perturbation causing the decay must be weak enough that there is only a small chance of decay before a collision occurs. The second inequality means that there must be enough time between collisions that an apparent energy conservation from initial to final state applies. Roughly speaking, collisions must be sufficiently frequent on the time scale of the decay process, but rare on the quantum time scale $\hbar/E$.

It should also be noted that the rule was derived by Dirac, not Fermi. The way Fermi got his name on it was that he was the one who named it a "golden rule." Fermi had a flair for finding memorable names. God knows how he ended up being a physicist.

A.45.6 Mopping up

The previous subsections derived the basics for the rate of beta decay. The purpose of this section is to pull it all together and get some actual ballpark estimates for beta decay.

First consider the possible values for the momenta $\vec{p}_e$ and $\vec{p}_{\bar{\nu}}$ of the electron and antineutrino. Their wave functions were approximately of the form

$$\psi_{\vec{p}} = \frac{1}{\sqrt{\mathcal{V}}} e^{\mathrm{i}(p_x x + p_y y + p_z z)/\hbar}$$

where $\mathcal{V} = \ell^3$ is the volume of the periodic box in which the decay is assumed to occur.

In a periodic box the wave function must be the same at opposite sides of the box. For example, the exponential factor $e^{\mathrm{i}p_x x/\hbar}$ is 1 at x=0, and it must be 1 again at $x = \ell$. That requires $p_x \ell/\hbar$ to be a whole multiple of 2π. Therefore p_x must be a whole multiple of $2\pi\hbar/\ell$. Successive possible p_x values are therefore spaced the finite amount $2\pi\hbar/\ell$ apart. And so are successive p_y and p_z values.

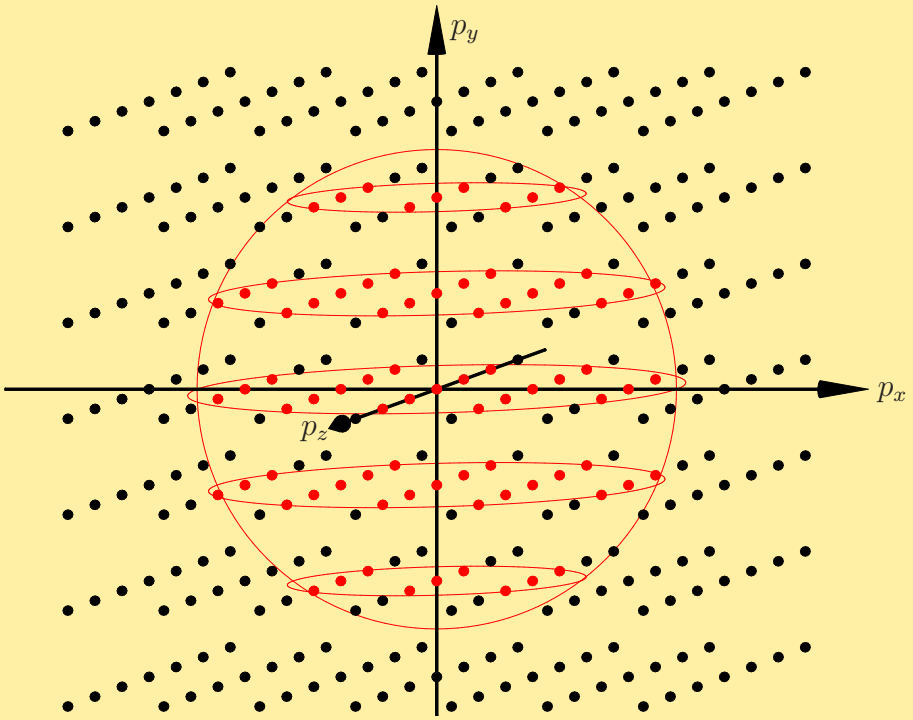


Figure A.27: Possible momentum states for a particle confined to a periodic box. The states are shown as points in momentum space. States that have momentum less than some example maximum value are in red.

Graphically this can by visualized by plotting the possible momentum values as points in a three-dimensional p_x, p_y, p_z axis system. That is done in figure

A.27. Each point correspond to one possible momentum state. Each point is the center of its own little cube with sides $2\pi\hbar/\ell$. The "volume" (in this three-dimensional momentum plot, not physical volume) of that little cube is $(2\pi\hbar/\ell)^3$. Since ℓ^3 is the physical volume $\mathcal{V}$ of the periodic box, the "volume" in momentum space taken up by each momentum state is $(2\pi\hbar)^3/\mathcal{V}$

That allows the number of different momentum states to be computed. In particular, consider how many states have magnitude of momentum $|\vec{p}\,'|$ less than some maximum value p. For some example value of p, these are the red states in figure A.27. Note that they form a sphere of radius p. That sphere has a "volume" equal to $\frac{4}{3}\pi p^3$. Since each state takes up a "volume" $(2\pi\hbar)^2/\mathcal{V}$, the number of states N is given by the number of such "volumes" in the sphere:

$$N_{|\vec{p}\,'|\leqslant p} = \frac{\frac{4}{3}\pi p^3}{(2\pi\hbar)^3/\mathcal{V}}$$

The number of electron states that have momentum in a range from p_e to $p_e + \mathrm{d}p_e$ can be found by taking a differential of the expression above:

$$\mathrm{d}N_e = \frac{\mathcal{V} p_e^2}{2\pi^2\hbar^3}\mathrm{d}p_e$$

(Here the range $\mathrm{d}p_e$ is assumed small, but not so small that the fact that the number of states is discrete would show up.) Each momentum state still needs to be multiplied by the number of corresponding antineutrino states to find the number of states of the complete system.

Now the kinetic energy of the antineutrino $T_{\bar{\nu}}$ is fixed in terms of that of the electron and the energy release of the decay Q by:

$$T_{\bar{\nu}} = Q - T_e$$

Here the kinetic energy of the final nucleus is ignored. The heavy final nucleus is unselfish enough to assure that momentum conservation is satisfied for whatever the momenta of the electron and antineutrino are, without demanding a noticeable share of the energy for itself. That is much like Mother Earth does not take any of the kinetic energy away if you shoot rockets out to space from different locations. You might write equations down for it, but the only thing they are going to tell you is that it is true as long as the speed of the nucleus does not get close to the speed of light. Beta decays do not release that much energy by far.

The electron and antineutrino kinetic energies are related to their momenta by Einstein's relativistic expression, chapter 1.1.2:

$$T_e = \sqrt{(m_e c^2)^2 + p_e^2 c^2} - m_e c^2 \qquad T_{\bar{\nu}} = p_{\bar{\nu}} c \tag{A.281}$$

where c is the speed of light and the extremely small rest mass of the neutrino was ignored. With the neutrino energy fixed, so is the magnitude of the neutrino

momentum:

$$p_{\bar\nu} = \frac{1}{c}(Q - T_e)$$

These result shows that the neutrino momentum is fixed for given electron momentum p_e. Therefore there should not be a neutrino momentum range $\mathrm{d}p_{\bar\nu}$ and so no neutrino states. However, Fermi's golden rule says that the theoretical energy after the decay does not need to be exactly the same as the one before it, because both energies have a bit of uncertainty. This slop in the energy conservation equation allows a range of energies

$$\Delta E_{\text{slop}} = \Delta T_{\bar\nu} = \Delta p_{\bar\nu} c \equiv \varepsilon$$

Therefore the total amount of neutrino states for a given electron momentum is not zero, but

$$\Delta N_{\bar\nu} = \frac{\mathcal{V} p_{\bar\nu}^2}{2\pi^2\hbar^3}\frac{1}{c}\varepsilon \qquad p_{\bar\nu} = \frac{1}{c}(Q - T_e)$$

The number of complete system states in an electron momentum range $\mathrm{d}p_e$ is the product of the number of electron states times the number of antineutrino states:

$$\mathrm{d}N = \mathrm{d}N_e \Delta N_{\bar\nu} = \frac{\mathcal{V}^2}{4\pi^4\hbar^6 c}\, p_e^2 p_{\bar\nu}^2\, \varepsilon\, \mathrm{d}p_e$$

Each of these states adds a contribution to the specific decay rate given by

$$\lambda_{\text{single final state}} = \frac{2\pi}{\hbar\varepsilon}|H_{21}^\ell|^2$$

Therefore the total specific decay rate is

$$\lambda = \int_{p_e=0}^{p_{e,\max}} \frac{\mathcal{V}^2 |H_{21}^\ell|^2}{2\pi^3\hbar^7 c} p_e^2 p_{\bar\nu}^2\, \mathrm{d}p_e$$

where the maximum electron momentum $p_{e,max}$ can be computed from the Q-value of the decay using (A.281). (For simplicity it will be assumed that $|H_{21}^\ell|^2$ has already been averaged over all directions of the electron and antineutrino momentum.)

The derived expression (A.277) for the Hamiltonian coefficient H_{21}^ℓ can be written in the form

$$|H_{21}^\ell|^2 = \frac{g^2}{\mathcal{V}^2}\frac{1}{(\ell!)^2}\left(\frac{\sqrt{p_e^2 + p_{\bar\nu}^2}R}{\hbar}\right)^{2\ell} C_N^\ell$$

$$C_N^\ell \equiv \overline{\left|\left\langle \psi_{2,\text{nuc}} \middle| \sum_{i=1}^{A} h_i \left(\frac{(\vec{p}_e + \vec{p}_{\bar\nu})\cdot\vec{r}_i}{\sqrt{p_e^2 + p_{\bar\nu}^2}R}\right)^\ell \middle| \psi_{1,\text{nuc}} \right\rangle\right|^2}$$

where the overline indicates some suitable average over the directions of the electron and antineutrino momenta.

It is not easy to say much about C_N^ℓ in general, beyond the fact that its magnitude should not be much more than one. This book will essentially ignore C_N^ℓ to ballpark the decay rate, assuming that its variation will surely be much less than that of the beta decay lifetimes, which vary from milliseconds to 10^{17} year.

The decay rate becomes after clean up

$$\lambda = \frac{1}{2\pi^3}\frac{g^2 m_{\rm e}^4 c^2}{\hbar^6}\frac{m_{\rm e}c^2}{\hbar}\frac{\tilde{R}^{2\ell}}{(\ell!)^2}C_N^\ell \int_{\tilde{p}_e=0}^{\tilde{p}_{e,\max}} (\tilde{p}_e^2+\tilde{p}_{\bar{\nu}}^2)^\ell \tilde{p}_e^2 \tilde{p}_{\bar{\nu}}^2 F^\ell \mathrm{d}\tilde{p}_e \tag{A.282}$$

where the $\tilde{p}$ indicate the electron and antineutrino momenta nondimensionalized with $m_{\rm e}c$. Also,

$$\tilde{Q} \equiv \frac{Q}{m_{\rm e}c^2} \quad \tilde{p}_{e,\max} = \sqrt{\tilde{Q}^2+2\tilde{Q}} \quad \tilde{p}_{\bar{\nu}} = \tilde{Q} - \sqrt{1+\tilde{p}_e^2}+1 \quad \tilde{R} \equiv \frac{m_{\rm e}cR}{\hbar} \tag{A.283}$$

Here $\tilde{Q}$ is the Q-value or kinetic energy release of the decay in units of the electron rest mass, and the next two relations follow from the expression for the relativistic kinetic energy. The variable $\tilde{R}$ a suitably nondimensionalized nuclear radius, and is small.

The factor F^ℓ that popped up out of nothing in the decay rate is thrown in to correct for the fact that the wave function of the electron is not really just an exponential. The nucleus pulls on the electron with its charge, and so changes its wave function locally significantly. The correction factor F^0 for allowed decays is called the "Fermi function" and is given by

$$F(\tilde{p}_e, Z_2, A) = \frac{2(1+\xi)}{\Gamma^2(1+2\xi)}\frac{1}{(2\tilde{p}_e\tilde{R})^{2-2\xi}}e^{\pi\eta}|\Gamma(\xi+\mathrm{i}\eta)|^2 \tag{A.284}$$

$$\xi \equiv \sqrt{1-(\alpha Z_2)^2} \quad \eta \equiv \alpha Z_2 \frac{\sqrt{1+\tilde{p}_e^2}}{\tilde{p}_e} \qquad \alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137}$$

where α is the fine structure constant and Γ the gamma function. The nonrelativistic version follows for letting the speed of light go to infinity, while keeping $p_e = m_{\rm e}c\tilde{p}_e$ finite. That gives $\xi = 1$ and

$$F(p_e, Z_2) = \frac{2\pi\eta}{1-e^{-2\pi\eta}} \qquad \eta = \frac{e^2}{4\pi\epsilon_0\hbar}\frac{m_{\rm e}}{p_e}$$

For beta-plus decay, just replace Z_2 by $-Z_2$, because an electron is just as much repelled by a negatively charged nucleus as a positron is by a positively charged one.

To ballpark the effect of the nuclear charge on the electron wave function, this book will use the relativistic Fermi function above whether it is an allowed decay or not.

For allowed decays, the factor in the decay rate that is governed by the Q-value and nuclear charge is

$$f = \int_{\tilde{p}_e=0}^{\tilde{p}_{e,\max}} \tilde{p}_e^2 \tilde{p}_\nu^2 F \mathrm{d}\tilde{p}_e \tag{A.285}$$

This quantity is known as the "Fermi integral." Typical values are shown in figure 14.52.

Note that f also depends a bit on the mass number through the nuclear radius in F. The figure used

$$A = 1.82 + 1.9\, Z_2 + 0.012\,71\, Z_2^2 - 0.000\,06\, Z_2^3 \tag{A.286}$$

for beta-minus decay and

$$A = -1.9 + 1.96\, Z_2 + 0.007\,9\, Z_2^2 - 0.000\,02\, Z_2^3 \tag{A.287}$$

for beta-plus decay, [23].

A.45.7 Electron capture

Electron capture is much more simply to analyze than beta decay, because the captured electron is in a known initial state.

It will be assumed that a 1s, or K-shell, electron is captured, though L-shell capture may also contribute to the decay rate for heavy nuclei. The Hamiltonian coefficient that drives the decay is

$$H_{21} = \langle 1\bar{\nu}, \vec{p}_{\bar{\nu}}| \langle 0e, 1\mathrm{s}|\psi_{2,\mathrm{nuc}}|H'\psi_{1,\mathrm{nuc}}|1e, 1\mathrm{s}\rangle |0\bar{\nu}, \vec{p}_{\bar{\nu}}\rangle$$

In this case, it is an electron annihilation term in the Hamiltonian that will produce a nonzero term. However, the result will be the pretty much same; the Hamiltonian coefficient simplifies to

$$H_{21} = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\mathrm{nuc}}| \sum_{i=1}^{A} g h_i \psi_{100}(\vec{r}_i) e^{\mathrm{i}\vec{p}_{\bar{\nu}}\cdot\vec{r}_i/\hbar} \psi_{1,\mathrm{nuc}}\rangle$$

Here ψ_{100} is the hydrogen ground state wave function, but rescaled for a nucleus of charge Ze instead of e. It does not contribute to making forbidden decays possible, because ψ_{100} is spherically symmetric. In other words, the 1s electron has no orbital angular momentum and so cannot contribute to conservation of angular momentum and parity. Therefore, ψ_{100} can safely be approximated by its value at the origin, from chapter 4.3,

$$\psi_{100}(\vec{r}_i) \approx \frac{1}{\sqrt{\pi a_0^3}} \qquad a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_\mathrm{e} e^2 Z_1} = \frac{\hbar}{m_\mathrm{e} c \alpha Z_1}$$

where α is the fine-structure constant.

The square Hamiltonian coefficient for ℓth-forbidden decays then becomes

$$|H_{21}^{\ell}|^2 = \frac{g^2}{\mathcal{V}} \frac{m_{\mathrm{e}}^3 c^3 \alpha^3 Z^3}{\pi \hbar^3} \frac{1}{(\ell!)^2} \left(\frac{p_{\bar{\nu}} R}{\hbar} \right)^{2\ell} C_N^{\ell}$$

$$C_N^{\ell} =\equiv \overline{\left| \left\langle \psi_{2,\mathrm{nuc}} \right| \sum_{i=1}^{A} h_i \left(\frac{\vec{p}_{\bar{\nu}} \cdot \vec{r}_i}{p_{\bar{\nu}} R} \right)^{\ell} \left| \psi_{1,\mathrm{nuc}} \right\rangle \right|^2}$$

The decay rate for electron capture is

$$\lambda = \frac{2\pi}{\hbar \varepsilon} |H_{21}^{\ell}|^2 \frac{\mathcal{V} p_{\bar{\nu}}^2}{2\pi^2 \hbar^3} \Delta p_{\bar{\nu}} \qquad \Delta p_{\bar{\nu}} = \frac{\varepsilon}{c}$$

where the first ratio is the decay rate of a single state, with ε the energy slop implied by Fermi's golden rule.

Put it all together, including the fact that there are two K electrons, and the electron-capture decay rate becomes

$$\lambda = \frac{2}{\pi^2} \frac{g^2 m_{\mathrm{e}}^4 c^2}{\hbar^6} \frac{m_{\mathrm{e}} c^2}{\hbar} (\alpha Z)^3 \frac{\tilde{R}^{2\ell}}{(\ell!)^2} C_N^{\ell} \tilde{p}_{\bar{\nu}}^{2\ell} \tilde{p}_{\bar{\nu}}^2 \tag{A.288}$$

where the $\tilde{p}_{\bar{\nu}}$ indicate the neutrino momentum nondimensionalized with $m_{\mathrm{e}} c$. Also,

$$\tilde{Q} \equiv \frac{Q}{m_{\mathrm{e}} c^2} \qquad \tilde{p}_{\bar{\nu}} = \tilde{Q} \qquad \tilde{R} \equiv \frac{m_{\mathrm{e}} c R}{\hbar} \tag{A.289}$$

for the coefficients.

Appendix D

Derivations

This appendix gives various derivations. Sometimes you need to see the derivation to judge whether a result is applicable in given circumstances. And some people like to see the derivation period.

D.1 Generic vector identities

The rules of engagement are as follows:

- The Cartesian axes are numbered using an index i, with $i = 1$, 2, and 3 for x, y, and z respectively.
- Also, r_i indicates the coordinate in the i direction, x, y, or z.
- Derivatives with respect to a coordinate r_i are indicated by a simple subscript i.
- If the quantity being differentiated is a vector, a comma is used to separate the vector index from differentiation ones.
- Index $\bar{\imath}$ is the number immediately following i in the cyclic sequence $\ldots 123123\ldots$ and $\bar{\bar{\imath}}$ is the number immediately preceding i.

The first identity to be derived involves the "vectorial triple product:"

$$\nabla \times \nabla \times \vec{v} = \nabla(\nabla \cdot \vec{v}) - \nabla^2 \vec{v} \tag{D.1}$$

To do so, first note that the i-th component of $\nabla \times \vec{v}$ is given by

$$v_{\bar{\imath},\bar{\imath}} - v_{\bar{\imath},\bar{\bar{\imath}}}$$

Repeating the rule, the i-th component of $\nabla \times \nabla \times \vec{v}$ is

$$(v_{\bar{\imath},i} - v_{i,\bar{\imath}})_{\bar{\imath}} - (v_{i,\bar{\bar{\imath}}} - v_{\bar{\bar{\imath}},i})_{\bar{\bar{\imath}}}$$

That writes out to

$$v_{i,ii} + v_{\bar{\imath},\bar{\imath}i} + v_{\bar{\bar{\imath}},\bar{\bar{\imath}}i} - v_{i,ii} - v_{i,\bar{\imath}\bar{\imath}} - v_{i,\bar{\bar{\imath}}\bar{\bar{\imath}}}$$

since the first and fourth terms cancel each other. The first three terms can be recognized as the i-th component of $\nabla(\nabla \cdot \vec{v})$ and the last three as the i-th component of $-\nabla^2 \vec{v}_i$.

A second identity to be derived involves the "scalar triple product:"

$$(\vec{a} \times \vec{b}) \cdot \vec{c} = \vec{a} \cdot (\vec{b} \times \vec{c}) \tag{D.2}$$

This is easiest derived from simply writing it out. The left hand side is

$$a_y b_z c_x - a_z b_y c_x + a_z b_x c_y - a_x b_z c_y + a_x b_y c_z - a_y b_x c_z$$

while the right hand side is

$$a_x b_y c_z - a_x b_z c_y + a_y b_z c_x - a_y b_x c_z + a_z b_x c_y - a_z b_y c_x$$

Inspection shows it to be the same terms in a different order. Note that since no order changes occur, the three vectors may be noncommuting operators.

D.2 Some Green's functions

D.2.1 The Poisson equation

The so-called "Poisson equation" is

$$-\nabla^2 u = f \qquad \nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

Here f is supposed to be a given function and u an unknown function that is to be found.

The solution u to the Poisson equation in infinite space may be found in terms of its so-called "Green's function" $G(\vec{r})$. In particular:

$$u(\vec{r}) = \int_{\text{all } \underline{\vec{r}}} G(\vec{r} - \underline{\vec{r}}) f(\underline{\vec{r}})\, \mathrm{d}^3\underline{\vec{r}} \qquad G(\vec{r} - \underline{\vec{r}}) = \frac{1}{4\pi|\vec{r} - \underline{\vec{r}}|}$$

Loosely speaking, the above integral solution chops function f up into spikes $f(\underline{\vec{r}})\mathrm{d}^3\underline{\vec{r}}$. A spike at a position $\underline{\vec{r}}$ then makes a contribution $G(\vec{r} - \underline{\vec{r}})f(\underline{\vec{r}})\mathrm{d}^3\underline{\vec{r}}$ to u at $\vec{r}$. Integration over all such spikes gives the complete u.

Note that often, the Poisson equation is written without a minus sign. Then there will be a minus sign in G.

The objective is now to derive the above Green's function. To do so, first an intuitive derivation will be given and then a more rigorous one. (See also chapter 13.3.4 for a more physical derivation in terms of electrostatics.)

The intuitive derivation defines $G(\vec{r})$ as the solution due to a unit spike, i.e. a "delta function," located at the origin. That means that $G = G(\vec{r})$ is the solution to

$$\nabla^2 G = \delta^3 \qquad \text{with} \quad G(\vec{r}) \to 0 \quad \text{when} \quad \vec{r} \to \infty$$

Here $\delta^3 = \delta^3(\vec{r})$ is the three-dimensional delta function, defined as an infinite spike at the origin that integrates to 1.

By itself the above definition is of course meaningless: infinity is not a valid number. To give it meaning, it is necessary to define an approximate delta function, one that is merely a large spike rather than an infinite one. This approximate delta function $\delta^3_\varepsilon = \delta^3_\varepsilon(\vec{r})$ must still integrate to 1 and will be required to be zero beyond some small distance ε from the origin:

$$\int \delta^3_\varepsilon(\vec{r})\,\mathrm{d}^3\vec{r} = 1 \qquad \text{and} \qquad \delta^3_\varepsilon(\vec{r}) = 0 \quad \text{if} \quad r = |\vec{r}| \geqslant \varepsilon$$

In the above integral the region of integration should at least include the small region of radius ε around the origin. The approximate delta function will further be assumed to be nonnegative. It must have large values in the small vicinity around the origin where it is nonzero; otherwise the integral over the small vicinity would be small instead of 1. But the key is that the values are not infinite, just large. So normal mathematics can be used.

The corresponding approximate Green's function $G_\varepsilon = G_\varepsilon(\vec{r})$ of the Poisson equation satisfies

$$-\nabla^2 G_\varepsilon = \delta^3_\varepsilon \qquad \text{with} \quad G_\varepsilon \to 0 \quad \text{when} \quad \vec{r} \to \infty$$

In the limit $\varepsilon \to 0$, $\delta^3_\varepsilon(\vec{r})$ becomes the Dirac delta function $\delta^3(\vec{r})$ and $G_\varepsilon(\vec{r})$ becomes the exact Green's function $G(\vec{r})$.

To find the approximate Green's function, it will be assumed that $\delta^3_\varepsilon(\vec{r})$ only depends on the distance $r = |\vec{r}|$ from the origin. In other words, it is assumed to be spherically symmetric. Then so is G_ε. (Note that this assumption is not strictly necessary. That can be seen from the general solution for the Poisson equation given earlier. But it should at least be assumed that $\delta^3_\varepsilon(\vec{r})$ is nonnegative. If it could have arbitrarily large negative values, then G_ε could be anything.)

Now integrate both sides of the Poisson equation over a sphere of a chosen radius r:

$$-\int_{|\vec{\underline{r}}| \leqslant r} \nabla^2 G_\varepsilon \,\mathrm{d}^3\vec{\underline{r}} = \int_{|\vec{\underline{r}}| \leqslant r} \delta^3_\varepsilon \,\mathrm{d}^3\vec{\underline{r}}$$

As noted, the delta function integrates to 1 as long as the vicinity of the origin is included. That means that the right hand side is 1 as long as $r \geqslant \varepsilon$. This will now be assumed. The left hand side can be written out. That gives

$$-\int_{|\vec{\underline{r}}| \leqslant r} \nabla \cdot (\nabla G_\varepsilon)\,\mathrm{d}^3\vec{\underline{r}} = 1 \quad \text{if} \quad r \geqslant \varepsilon$$

According to the [divergence] [Gauss] [Ostrogradsky] theorem, the left hand side can be written as a surface integral to give

$$-\int_{|\vec{\underline{r}}|=r} \vec{n} \cdot (\nabla G_\varepsilon)\,\mathrm{d}S = 1 \quad \text{if} \quad r \geqslant \varepsilon$$

Here S stands for the surface of the sphere of radius r. The total surface is $4\pi r^2$. Also $\vec{n}$ is the unit vector orthogonal to the surface, in the outward direction. That is the radial direction. The total differential of calculus then implies that $\vec{n} \cdot \nabla G_\varepsilon$ is the radial derivative $\partial G_\varepsilon / \partial r$. So,

$$-\frac{\partial G_\varepsilon}{\partial r} 4\pi r^2 = 1 \quad \text{if} \quad r \geqslant \varepsilon$$

Because G_ε is required to vanish at large distances, this integrates to

$$G_\varepsilon = \frac{1}{4\pi r} \quad \text{if} \quad r \geqslant \varepsilon$$

The exact Green's function G has ε equal to zero, so

$$G = \frac{1}{4\pi r} \quad \text{if} \quad r \neq 0$$

Finally the rigorous derivation without using poorly defined things like delta functions. In the supposed general solution of the Poisson equation given earlier, change integration variable to $\vec{\rho} = \underline{\vec{r}} - \vec{r}$

$$u(\vec{r}) = \int_{\text{all } \vec{r}} G(\vec{r} - \underline{\vec{r}}) f(\underline{\vec{r}}) \, \mathrm{d}^3 \underline{\vec{r}} = \int_{\text{all } \vec{\rho}} G(\vec{\rho}) f(\vec{r} + \vec{\rho}) \, \mathrm{d}^3 \vec{\rho} \qquad G(\vec{\rho}) = \frac{1}{4\pi |\vec{\rho}|}$$

It is to be shown that the function u defined this way satisfies the Poisson equation $\nabla^2 u(\vec{r}) = f(\vec{r})$. To do so, apply ∇ twice:

$$\nabla^2 u(\vec{r}) = \int_{\text{all } \vec{\rho}} G(\vec{\rho}) \nabla^2 f(\vec{r} + \vec{\rho}) \, \mathrm{d}^3 \vec{\rho} = \int_{\text{all } \vec{\rho}} G(\vec{\rho}) \nabla_\rho^2 f(\vec{r} + \vec{\rho}) \, \mathrm{d}^3 \vec{\rho}$$

Here ∇_ρ means differentiation with respect to the components of $\vec{\rho}$ instead of the components of $\vec{r}$. Because f depends only on $\vec{r} + \vec{\rho}$, you get the same answer whichever way you differentiate.

It will be assumed that the function f is well behaved, at least continuous, and becomes zero reasonably quickly at infinity. In that case, you can get a valid approximation to the integral above if you exclude very small and very large values of $\vec{r}$:

$$\nabla^2 u(\vec{r}) \approx \int_{\varepsilon < |\vec{\rho}| < R} G(\vec{\rho}) \nabla_\rho^2 f(\vec{r} + \vec{\rho}) \, \mathrm{d}^3 \vec{\rho}$$

In particular, this approximation becomes exact in the limits where the constants $\varepsilon \to 0$ and $R \to \infty$. The integral can now be rewritten as

$$\nabla^2 u(\vec{r}) \approx \int_{\varepsilon < |\vec{\rho}| < R} \nabla_\rho [G(\vec{\rho}) \nabla_\rho f(\vec{r} + \vec{\rho})] - \nabla_\rho [f(\vec{r} + \vec{\rho}) \nabla_\rho G(\vec{\rho})] + f(\vec{r} + \vec{\rho}) \nabla_\rho^2 G(\vec{\rho}) \, \mathrm{d}^3 \vec{\rho}$$

as can be verified by explicitly differentiating out the three terms of the integrand. Next note that the third term is zero, because as seen above G satisfies the homogeneous Poisson equation away from the origin. And the other two terms can be written out using the [divergence] [Gauss] [Ostrogradsky] theorem much like before. This produces integrals over both the bounding sphere of radius R, as well as over the bounding sphere of radius ε. But the integrals over the sphere of radius R will be vanishingly small if f becomes zero sufficiently quickly at infinity. Similarly, the integral of the first term over the small sphere is vanishingly small, because G is $1/4\pi\varepsilon$ on the small sphere but the surface of the small sphere is $4\pi\varepsilon^2$. However, in the second term, the derivative of G in the negative radial direction is $1/4\pi\varepsilon^2$, which multiplies to 1 against the surface of the small sphere. Therefore the second term produces the average of $f(\vec{r}+\vec{\rho})$ over the small sphere, and that becomes $f(\vec{r})$ in the limit $|\vec{\rho}| = \varepsilon \to 0$. So the Poisson equation applies.

D.2.2 The screened Poisson equation

The so-called "screened Poisson equation" is

$$-\nabla^2 u + c^2 u = f \qquad \nabla \equiv \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

Here f is supposed to be a given function and u an unknown function that is to be found. Further c is a given constant. If c is zero, this is the Poisson equation. However, nonzero c corresponds to the inhomogeneous steady Klein-Gordon equation for a particle with nonzero mass.

The analysis of the screened Poisson equation is almost the same as for the Poisson equation given in the previous subsection. Therefore only the differences will be noted here. The approximate Green's function must satisfy, away from the origin,

$$-\nabla^2 G_\varepsilon + c^2 G_\varepsilon = 0 \quad \text{if} \quad r \geqslant \varepsilon$$

The solution to this that vanishes at infinity is of the form

$$G_\varepsilon = C\frac{e^{-cr}}{r} \quad \text{if} \quad r \geqslant \varepsilon$$

where C is some constant. To check this, plug it in, using the expression (N.5) for ∇^2 in spherical coordinates. To identify the constant C, integrate the full equation

$$-\nabla^2 G_\varepsilon + c^2 G_\varepsilon = \delta^3_\varepsilon$$

over a sphere of radius ε around the origin and apply the divergence theorem as in the previous subsection. Taking the limit $\varepsilon \to 0$ then gives $C = 1/4\pi$, which gives the exact Green's function as

$$G(\vec{r}) = \frac{e^{-cr}}{4\pi r} \quad \text{if} \quad r = |\vec{r}| \neq 0$$

The rigorous derivation is the same as before save for an additional c^2Gf term in the integrand, which drops out against the $-f\nabla_\rho^2 G$ one.

D.3 Lagrangian mechanics

This note gives the derivations for the addendum on the Lagrangian equations of motion.

D.3.1 Lagrangian equations of motion

To derive the nonrelativistic Lagrangian, consider the system to be build up from elementary particles numbered by an index j. You may think of these particles as the atoms you would use if you would do a molecular dynamics computation of the system. Because the system is assumed to be fully determined by the generalized coordinates, the position of each individual particle is fully fixed by the generalized coordinates and maybe time. (For example, it is implicit in a solid body approximation that the atoms are held rigidly in their relative position. Of course, that is approximate; you pay *some* price for avoiding a full molecular dynamics simulation.)

Newton's second law says that the motion of each individual particle j is governed by

$$m_j \frac{\mathrm{d}^2 \vec{r}_j}{\mathrm{d}t^2} = -\frac{\partial V}{\partial \vec{r}_j} + \vec{F}'_j$$

where the derivative of the potential V can be taken to be its gradient, if you (justly) object to differentiating with respect to vectors, and $\vec{F}'_j$ indicates any part of the force not described by the potential.

Now consider an infinitesimal virtual displacement of the system from its normal evolution in time. It produces an infinitesimal change in position $\delta\vec{r}_j(t)$ for each particle. After such a displacement, $\vec{r}_j + \delta\vec{r}_j$ of course no longer satisfies the correct equations of motion, but the kinetic and potential energies still exist.

In the equation of motion for the correct position $\vec{r}_j$ above, take the mass times acceleration to the other side, multiply by the virtual displacement, sum over all particles j, and integrate over an arbitrary time interval:

$$0 = \int_{t_1}^{t_2} \sum_j \left[-m_j \frac{\mathrm{d}^2 \vec{r}_j}{\mathrm{d}t^2} - \frac{\partial V}{\partial \vec{r}_j} + \vec{F}'_j \right] \cdot \delta\vec{r}_j \, \mathrm{d}t$$

Multiply out and integrate the first term by parts:

$$0 = \int_{t_1}^{t_2} \sum_j \left[m_j \frac{\mathrm{d}\vec{r}_j}{\mathrm{d}t} \cdot \delta \frac{\mathrm{d}\vec{r}_j}{\mathrm{d}t} - \frac{\partial V}{\partial \vec{r}_j} \cdot \delta\vec{r}_j + \vec{F}'_j \delta\vec{r}_j \right] \mathrm{d}t$$

The virtual displacements of interest here are only nonzero over a limited range of times, so the integration by parts did not produce any end point values.

Recognize the first two terms within the brackets as the virtual change in the Lagrangian due to the virtual displacement at that time. Note that this requires that the potential energy depends only on the position coordinates and time, and not also on the time derivatives of the position coordinates. You get

$$0 = \delta \int_{t_1}^{t_2} \mathcal{L}\,\mathrm{d}t + \int_{t_1}^{t_2} \sum_j \left[\vec{F}'_j \cdot \delta\vec{r}_j\right]\,\mathrm{d}t \tag{D.3}$$

In case that the additional forces $\vec{F}'_j$ are zero, this produces the action principle: the time integral of the Lagrangian is unchanged under infinitesimal virtual displacements of the system, assuming that they vanish at the end points of integration. More generally, for the virtual work by the additional forces to be zero will require that the virtual displacements respect the rigid constraints, if any. The infinite work done in violating a rigid constraint is not modeled by the potential V in any normal implementation.

Unchanging action is an integral equation involving the Lagrangian. To get ordinary differential equations, take the virtual change in position to be that due to an infinitesimal change $\delta q_k(t)$ in a single generic generalized coordinate. Represent the change in the Lagrangian in the expression above by its partial derivatives, and the same for $\delta\vec{r}_j$:

$$0 = \int_{t_1}^{t_2} \left[\frac{\partial \mathcal{L}}{\partial q_k}\delta q_k + \frac{\partial \mathcal{L}}{\partial \dot{q}_k}\delta\dot{q}_k\right]\,\mathrm{d}t + \int_{t_1}^{t_2} \sum_j \left[\vec{F}'_j \cdot \frac{\partial \vec{r}_j}{\partial q_k}\delta q_k\right]\,\mathrm{d}t$$

The integrand in the final term is by definition the generalized force Q_k multiplied by δq_k. In the first integral, the second term can be integrated by parts, and then the integrals can be combined to give

$$0 = \int_{t_1}^{t_2} \left[\frac{\partial \mathcal{L}}{\partial q_k} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_k}\right) + Q_k\right]\delta q_k\,\mathrm{d}t$$

Now suppose that there is any time at which the expression within the square brackets is nonzero. Then a virtual change δq_k that is only nonzero in a very small time interval around that time, and everywhere positive in that small interval, would produce a nonzero right hand side in the above equation, but it must be zero. Therefore, the expression within brackets must be zero at all times. That gives the Lagrangian equations of motion, because the expression between parentheses is defined as the canonical momentum.

D.3.2 Hamiltonian dynamics

To derive the Hamiltonian equations, consider the general differential of the Hamiltonian function (regardless of any motion that may go on). According to

the given definition of the Hamiltonian function, and using a total differential for $\mathrm{d}\mathcal{L}$,

$$\mathrm{d}H = \left(\sum_k p_k^{\mathrm{c}} \mathrm{d}\dot{q}_k\right) + \sum_k \dot{q}_k \mathrm{d}p_k^{\mathrm{c}} - \sum_k \frac{\partial \mathcal{L}}{\partial q_k}\mathrm{d}q_k - \left(\sum_k \frac{\partial \mathcal{L}}{\partial \dot{q}_k}\mathrm{d}\dot{q}_k\right) - \frac{\partial \mathcal{L}}{\partial t}\mathrm{d}t$$

The sums within parentheses cancel each other because of the definition of the canonical momentum. The remaining differences are of the arguments of the Hamiltonian function, and so by the very definition of partial derivatives,

$$\frac{\partial H}{\partial q_k} = -\frac{\partial \mathcal{L}}{\partial q_k} \qquad \frac{\partial H}{\partial p_k^{\mathrm{c}}} = \dot{q}_k \qquad \frac{\partial H}{\partial t} = -\frac{\partial \mathcal{L}}{\partial t}$$

Now consider an actual motion. For an actual motion, $\dot{q}_k$ is the time derivative of q_k, so the second partial derivative gives the first Hamiltonian equation of motion. The first partial derivative gives the second equation when combined with the Lagrangian equation of motion (A.2).

It is still to be shown that the Hamiltonian of a classical system is the sum of kinetic and potential energy if the position of the system does not depend explicitly on time. The Lagrangian can be written out in terms of the system particles as

$$\sum_j \sum_{\underline{k}=1}^{K} \sum_{\underline{\underline{k}}=1}^{K} \tfrac{1}{2} m_j \frac{\partial \vec{r}_j}{\partial q_{\underline{k}}} \cdot \frac{\partial \vec{r}_j}{\partial q_{\underline{\underline{k}}}} \dot{q}_{\underline{k}} \dot{q}_{\underline{\underline{k}}} - V(q_1, q_2, \ldots, q_K, t)$$

where the sum represents the kinetic energy. The Hamiltonian is defined as

$$\sum_k \dot{q}_k \frac{\partial \mathcal{L}}{\partial \dot{q}_k} - \mathcal{L}$$

and straight substitution shows the first term to be twice the kinetic energy.

D.3.3 Fields

As discussed in {A.1.5}, the Lagrangian for fields takes the form

$$\mathcal{L} = \mathcal{L}_0 + \int \pounds \, \mathrm{d}^3\vec{r}$$

Here the spatial integration is over all space. The first term depends only on the discrete variables

$$\mathcal{L}_0 = \mathcal{L}_0(\ldots; q_k, \dot{q}_k; \ldots)$$

where $q_k = q_k(t)$ denotes discrete variable number k. The dot indicates the time derivative of that variable. The Lagrangian density also depends on the fields

$$\pounds = \pounds(\ldots; \varphi_\alpha, \varphi_{\alpha_t}, \varphi_{\alpha_1}, \varphi_{\alpha_2}, \varphi_{\alpha_3}; \ldots; q_k; \dot{q}_k; \ldots)$$

where φ_α is field number α. A subscript t indicates the partial time derivative, and 1, 2, or 3 the partial x, y or z derivative.

The action is

$$\mathcal{S} = \int_{t_1}^{t_2} \left(\mathcal{L}_0 + \int \pounds \, \mathrm{d}^3\vec{r} \right) \mathrm{d}t$$

where the time range from t_1 to t_2 must include the times of interest. The action must be unchanged under small deviations from the correct evolution, as long as these deviations vanish at the limits of integration. That requirement defines the Lagrangian. (For simple systems the Lagrangian then turns out to be the difference between kinetic and potential energies. But it is not obvious what to make of that if there are fields.)

Consider now first an infinitesimal deviation $\delta q_k = \delta q_k(t)$ in a discrete variable q_k. The change in action that must be zero is then

$$0 = \delta\mathcal{S} = \int_{t_1}^{t_2} \left(\frac{\partial \mathcal{L}_0}{\partial q_k} \delta q_k + \frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} \delta \dot{q}_k + \int \frac{\partial \pounds}{\partial q_k} \mathrm{d}^3\vec{r}\, \delta q_k + \int \frac{\partial \pounds}{\partial \dot{q}_k} \mathrm{d}^3\vec{r}\, \delta \dot{q}_k \right) \mathrm{d}t$$

After an integration by parts of the second and fourth terms that becomes, noting that the deviation must vanish at the initial and final times,

$$0 = \delta\mathcal{S} = \int_{t_1}^{t_2} \left[\frac{\partial \mathcal{L}_0}{\partial q_k} - \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} + \int \frac{\partial \pounds}{\partial q_k} \mathrm{d}^3\vec{r} - \frac{\mathrm{d}}{\mathrm{d}t} \int \frac{\partial \pounds}{\partial \dot{q}_k} \mathrm{d}^3\vec{r} \right] \delta q_k \, \mathrm{d}t$$

This can only be zero for whatever you take $\delta q_k = \delta q_k(t)$ if the expression within square brackets is zero. That gives the final Lagrangian equation for the discrete variable q_k as

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} + \int \frac{\partial \pounds}{\partial \dot{q}_k} \mathrm{d}^3\vec{r} \right) = \frac{\partial \mathcal{L}_0}{\partial q_k} + \int \frac{\partial \pounds}{\partial q_k} \mathrm{d}^3\vec{r} \tag{1}$$

Next consider an infinitesimal deviation $\delta\varphi_\alpha = \delta\varphi_\alpha(\vec{r};t)$ in field φ_α. The change in action that must be zero is then

$$0 = \delta S = \int_{t_1}^{t_2} \int \left(\frac{\partial \pounds}{\partial \varphi_\alpha} \delta\varphi_\alpha + \frac{\partial \pounds}{\partial \varphi_{\alpha_t}} \delta\varphi_{\alpha_t} + \sum_{i=1}^{3} \frac{\partial \pounds}{\partial \varphi_{\alpha_i}} \delta\varphi_{\alpha_i} \right) \mathrm{d}^3\vec{r}\, \mathrm{d}t$$

Now integrate the derivative terms by parts in the appropriate direction to get, noting that the deviation must vanish at the limits of integration,

$$0 = \delta S = \int_{t_1}^{t_2} \int \left[\frac{\partial \pounds}{\partial \varphi_\alpha} - \frac{\partial}{\partial t} \left(\frac{\partial \pounds}{\partial \varphi_{\alpha_t}} \right) - \sum_{i=1}^{3} \frac{\partial}{\partial r_i} \left(\frac{\partial \pounds}{\partial \varphi_{\alpha_i}} \right) \right] \delta\varphi_\alpha \, \mathrm{d}^3\vec{r}\, \mathrm{d}t$$

Here r_i for $i = 1$, 2, or 3 stands for x, y, or z. If the above expression is to be zero for whatever you take the small change $\delta\varphi_\alpha = \delta\varphi_\alpha(\vec{r};t)$ to be, then the expression within square brackets will have to be zero at every position and time. That gives the equation for the field φ_α:

$$\frac{\partial}{\partial t}\left(\frac{\partial \pounds}{\partial \varphi_{\alpha_t}}\right) + \sum_{i=1}^{3} \frac{\partial}{\partial r_i}\left(\frac{\partial \pounds}{\partial \varphi_{\alpha_i}}\right) = \frac{\partial \pounds}{\partial \varphi_\alpha} \tag{2}$$

The canonical momenta are defined as

$$p_k^{\mathrm{c}} \equiv \frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} + \int \frac{\partial \pounds}{\partial \dot{q}_k}\,\mathrm{d}^3\vec{r} \qquad \pi_\alpha^{\mathrm{c}} \equiv \frac{\partial \pounds}{\partial \varphi_{\alpha_t}} \tag{3}$$

These are the quantities inside the time derivatives of the Lagrangian equations.

For Hamilton's equations, assume at first that there are no discrete variables. In that case, the Hamiltonian can be written in terms of a Hamiltonian density h:

$$H = \int h\,\mathrm{d}^3\vec{r} \qquad h = \sum_\alpha \pi_\alpha^{\mathrm{c}} \varphi_{\alpha_t} - \pounds$$

Take a differential of the Hamiltonian density

$$\mathrm{d}h = \sum_\alpha \left[\pi_\alpha^{\mathrm{c}} \mathrm{d}\varphi_{\alpha_t} + \varphi_{\alpha_t} \mathrm{d}\pi_\alpha^{\mathrm{c}} - \frac{\partial \pounds}{\partial \varphi_{\alpha_t}} \mathrm{d}\varphi_{\alpha_t} - \frac{\partial \pounds}{\partial \varphi_{\alpha_i}} \mathrm{d}\varphi_{\alpha_i} - \frac{\partial \pounds}{\partial \varphi_\alpha} \mathrm{d}\varphi_\alpha\right]$$

The first and third terms in the square brackets cancel because of the definition of the canonical momentum. Then according to calculus

$$\frac{\partial h}{\partial \pi_\alpha^{\mathrm{c}}} = \varphi_{\alpha_t} \qquad \frac{\partial h}{\partial \varphi_{\alpha_i}} = -\frac{\partial \pounds}{\partial \varphi_{\alpha_i}} \qquad \frac{\partial h}{\partial \varphi_\alpha} = -\frac{\partial \pounds}{\partial \varphi_\alpha}$$

The first of these expressions gives the time derivative of φ_α. The other expressions may be used to replace the derivatives of the Lagrangian density in the Lagrangian equations of motion (2). That gives Hamilton's equations as

$$\frac{\partial \varphi_\alpha}{\partial t} = \frac{\partial h}{\partial \pi_\alpha^{\mathrm{c}}} \qquad \frac{\partial \pi_\alpha^{\mathrm{c}}}{\partial t} = -\frac{\partial h}{\partial \varphi_\alpha} + \sum_{i=1}^{3} \frac{\partial}{\partial r_i}\left(\frac{\partial h}{\partial \varphi_{\alpha_i}}\right) \tag{4}$$

If there are discrete variables, this no longer works. The full Hamiltonian is then

$$H = \sum_k p_k^{\mathrm{c}} \dot{q}_k + \int \sum_\alpha \pi_\alpha^{\mathrm{c}} \varphi_{\alpha_t}\,\mathrm{d}^3\vec{r} - \mathcal{L}_0 - \int \pounds\,\mathrm{d}^3\vec{r}$$

To find Hamilton's equations, the integrals in this Hamiltonian must be approximated. The region of integration is mentally chopped into little pieces of the same volume $\mathrm{d}\mathcal{V}$. Then by approximation

$$\int \pounds\,\mathrm{d}^3\vec{r} \approx \sum_n \pounds_n \mathrm{d}\mathcal{V}$$

Here n numbers the small pieces and $\pounds_n$ stands for the value of $\pounds$ at the center point of piece n. Note that this is essentially the Riemann sum of calculus. A

similar approximation is made for the other integral in the Hamiltonian, and the one in the canonical momenta (3). Then the approximate Hamiltonian becomes

$$H_{\rm app} = \sum_k p_k^{\rm c} \dot{q}_k + \sum_{\alpha,n} \pi_{\alpha_n}^{\rm c} \varphi_{\alpha t_n} {\rm d}\mathcal{V} - \mathcal{L}_0 - \sum_n \pounds_n {\rm d}\mathcal{V}$$

The differential of this approximate Hamiltonian is

$$\begin{aligned}
{\rm d}H_{\rm app} \;=\; & \sum_k \dot{q}_k {\rm d}p_k^{\rm c} + \sum_k p_k^{\rm c} {\rm d}\dot{q}_k + \sum_{\alpha,n} \varphi_{\alpha t_n} {\rm d}\mathcal{V} {\rm d}\pi_{\alpha_n}^{\rm c} + \sum_{\alpha,n} \pi_{\alpha_n}^{\rm c} {\rm d}\mathcal{V} {\rm d}\varphi_{\alpha t_n} \\
& - \sum_k \frac{\partial \mathcal{L}_0}{\partial q_k} {\rm d}q_k - \sum_k \frac{\partial \mathcal{L}_0}{\partial \dot{q}_k} {\rm d}\dot{q}_k - \sum_{k,n} \frac{\partial \pounds_n}{\partial q_k} {\rm d}\mathcal{V} {\rm d}q_k - \sum_{k,n} \frac{\partial \pounds_n}{\partial \dot{q}_k} {\rm d}\mathcal{V} {\rm d}\dot{q}_k \\
& - \sum_{\alpha,n} \frac{\partial \pounds_n}{\partial \varphi_{\alpha_n}} {\rm d}\mathcal{V} {\rm d}\varphi_{\alpha_n} - \sum_{\alpha,n} \frac{\partial \pounds_n}{\partial \varphi_{\alpha t_n}} {\rm d}\mathcal{V} {\rm d}\varphi_{\alpha t_n} - \sum_{\alpha,n,i} \frac{\partial \pounds_n}{\partial \varphi_{\alpha i_n}} {\rm d}\mathcal{V} {\rm d}\varphi_{\alpha i_n}
\end{aligned}$$

The $\mathrm{d}\dot{q}_k$ and $\mathrm{d}\varphi_{\alpha t_n}$ terms drop out because of the definitions of the canonical momenta. The remainder allows expressions for the partial derivatives of the approximate Hamiltonian to be identified.

The $\mathrm{d}p_k^{\rm c}$ term allows the time derivative of q_k to be identified with the partial derivative of $H_{\rm app}$ with respect to $p_k^{\rm c}$. And the Lagrangian expression for the time derivative of $p_k^{\rm c}$, as given in (1), may be rewritten in terms of corresponding derivatives of the approximate Hamiltonian. Together that gives, in the limit $\mathrm{d}\mathcal{V} \to 0$,

$$\frac{{\rm d}q_k}{{\rm d}t} = \frac{\partial H}{\partial p_k^{\rm c}} \qquad \frac{{\rm d}p_k^{\rm c}}{{\rm d}t} = -\frac{\partial H}{\partial q_k} \tag{5}$$

For the field, consider an position $\vec{r}$ corresponding to the center of an arbitrary little volume $n = \underline{n}$. Then the $\mathrm{d}\pi_{\alpha_{\underline{n}}}^{\rm c}$ term allows the time derivative of φ_α at this arbitrary position to be identified in terms of the partial derivative of the approximate Hamiltonian with respect to $\pi_\alpha^{\rm c}$ at the same location. And the Lagrangian expression for the time derivative of $\pi_\alpha^{\rm c}$, as given by (2), may be rewritten in terms of corresponding derivatives of the approximate Hamiltonian. Together that gives, in the limit $\mathrm{d}\mathcal{V} \to 0$, and leaving $\underline{n}$ away since it can be any position,

$$\frac{\partial \varphi_\alpha}{\partial t} = \lim_{{\rm d}\mathcal{V}\to 0} \frac{1}{{\rm d}\mathcal{V}} \frac{\partial H_{\rm app}}{\partial \pi_\alpha^{\rm c}} \qquad \frac{\partial \pi_\alpha^{\rm c}}{\partial t} = -\lim_{{\rm d}\mathcal{V}\to 0} \frac{1}{{\rm d}\mathcal{V}} \frac{\partial H_{\rm app}}{\partial \varphi_\alpha} + \sum_{i=1}^{3} \frac{\partial}{\partial r_i} \lim_{{\rm d}\mathcal{V}\to 0} \frac{1}{{\rm d}\mathcal{V}} \frac{\partial H_{\rm app}}{\partial \varphi_{\alpha_i}} \tag{6}$$

Of course, in real life you would not actually write out these limits. Instead you simply differentiate the normal Hamiltonian H until you have to start differentiating inside an integral, like maybe,

$$\frac{\partial}{\partial \varphi_\alpha} \int \pounds \, {\rm d}^3\vec{r}$$

Then you think to yourself that you are not really evaluating this, but actually

$$\frac{\partial}{\partial \varphi_{\alpha_{\underline{n}}}} \sum_n \pounds_n \mathrm{d}\mathcal{V} = \frac{\partial \pounds_{\underline{n}}}{\partial \varphi_{\alpha_{\underline{n}}}} \mathrm{d}\mathcal{V}$$

where $\underline{n}$ indicates the position that you are considering the field at. And you are going to divide out the volume $\mathrm{d}\mathcal{V}$. That then boils down to

$$\frac{\partial}{\partial \varphi_\alpha} \int \pounds \, \mathrm{d}^3\vec{r} \quad \Longrightarrow \quad \frac{\partial \pounds}{\partial \varphi_\alpha}$$

even though the left hand side would mathematically be nonsense without discretization and division by $\mathrm{d}\mathcal{V}$.

D.4 Lorentz transformation derivation

This note derives the Lorentz transformation as discussed in chapter 1.2. The question is what is the relationship between the time and spatial coordinates $t_\mathrm{A}, x_\mathrm{A}, y_\mathrm{A}, z_\mathrm{A}$ that an observer A attaches to an arbitrary event versus the coordinates $t_\mathrm{B}, x_\mathrm{B}, y_\mathrm{B}, z_\mathrm{B}$ that an observer B attaches to them.

Note that since the choices what to define as time zero and as the origin are quite arbitrary, it can be arranged that $x_\mathrm{B}, y_\mathrm{B}, z_\mathrm{B}, t_\mathrm{B}$ are all zero when $x_\mathrm{A}, y_\mathrm{A}, z_\mathrm{A}, t_\mathrm{A}$ are all zero. That simplifies the mathematics, so it will be assumed. It will also be assumed that the axis systems of the two observers are taken to be parallel and that the x axes are along the direction of relative motion between the observers, figure 1.2.

It will further be assumed that the relationship between the coordinates is linear;

$$\begin{aligned} t_\mathrm{B} &= a_{tx}x_\mathrm{A} + a_{ty}y_\mathrm{A} + a_{tz}z_\mathrm{A} + a_{tt}t_\mathrm{A} & \qquad y_\mathrm{B} &= a_{yx}x_\mathrm{A} + a_{yy}y_\mathrm{A} + a_{yz}z_\mathrm{A} + a_{yt}t_\mathrm{A} \\ x_\mathrm{B} &= a_{xx}x_\mathrm{A} + a_{xy}y_\mathrm{A} + a_{xz}z_\mathrm{A} + a_{xt}t_\mathrm{A} & \qquad z_\mathrm{B} &= a_{zx}x_\mathrm{A} + a_{zy}y_\mathrm{A} + a_{zz}z_\mathrm{A} + a_{zt}t_\mathrm{A} \end{aligned}$$

where the $a_{..}$ are constants still to be found.

The biggest reason to assume that the transformation should be linear is that if space is populated with observers A and B, rather than just have a single one sitting at the origin of that coordinate system, then a linear transformation assures that all pairs of observers A and B see the exact same transformation. In addition, the transformation from $x_\mathrm{B}, y_\mathrm{B}, z_\mathrm{B}, t_\mathrm{B}$ back to $x_\mathrm{A}, y_\mathrm{A}, z_\mathrm{A}, t_\mathrm{A}$ should be of the same form as the one the other way, since the principle of relativity asserts that the two coordinate systems are equivalent. A linear transformation has a back transformation that is also linear.

Another way to look at it is to say that the spatial and temporal scales seen by normal observers are miniscule compared to the scales of the universe. Based on that idea you would expect that the relation between their coordinates would be a linearized Taylor series.

A lot of additional constraints can be put in because of physical symmetries that surely still apply even allowing for relativity. For example, the transformation to $x_{\rm B}, t_{\rm B}$ should not depend on the arbitrarily chosen positive directions of the y and z axes, so throw out the y and z terms in those equations. Seen in a mirror along the xy-plane, the y transformation should look the same, even if z changes sign, so throw out $z_{\rm A}$ from the equation for $y_{\rm B}$. Similarly, there goes $y_{\rm A}$ in the equation for $z_{\rm B}$. Since the choice of y and z axes is arbitrary, the remaining $a_{z\cdot}$ coefficients must equal the corresponding $a_{y\cdot}$ ones. Since the basic premise of relativity is that the coordinate systems A and B are equivalent, the y difference between tracks parallel to the direction of motion cannot get longer for B and shorter for A, nor vice-versa, so $a_{yy} = 1$. Finally, by the very definition of the relative velocity v of coordinate system B with respect to system A, $x_{\rm B} = y_{\rm B} = z_{\rm B} = 0$ should correspond to $x_{\rm A} = vt_{\rm A}$. And by the principle of relativity, $x_{\rm A} = y_{\rm A} = z_{\rm A} = 0$ should correspond to $x_{\rm B} = -vt_{\rm B}$.

You might be able to think up some more constraints, but this will do. Put it all together to get

$$\begin{aligned} t_{\rm B} &= a_{tx}x_{\rm A} + a_{xx}t_{\rm A} & \qquad y_{\rm B} &= a_{yx}x_{\rm A} + y_{\rm A} + a_{yt}t_{\rm A} \\ x_{\rm B} &= a_{xx}(x_{\rm A} - vt_{\rm A}) & \qquad z_{\rm B} &= a_{yx}x_{\rm A} + z_{\rm A} + a_{yt}t_{\rm A} \end{aligned}$$

Next the trick is to consider the wave front emitted by some light source that flashes at time zero at the then coinciding origins. Since according to the principle of relativity the two coordinate systems are fully equivalent, in both coordinate systems the wave front forms an expanding spherical shell with radius ct:

$$x_{\rm A}^2 + y_{\rm A}^2 + z_{\rm A}^2 = c^2t_{\rm A}^2 \qquad x_{\rm B}^2 + y_{\rm B}^2 + z_{\rm B}^2 = c^2t_{\rm B}^2$$

Plug the linearized expressions for $x_{\rm B}, y_{\rm B}, z_{\rm B}, t_{\rm B}$ in terms of $x_{\rm A}, y_{\rm A}, z_{\rm A}, t_{\rm A}$ into the second equation and demand that it is consistent with the first equation, and you obtain the Lorentz transformation. To get the back transformation giving $x_{\rm A}, y_{\rm A}, z_{\rm A}, t_{\rm A}$ in terms of $x_{\rm B}, y_{\rm B}, z_{\rm B}, t_{\rm B}$, solve the Lorentz equations for $x_{\rm A}$, $y_{\rm A}$, $z_{\rm A}$, and $t_{\rm A}$.

To derive the given transformations between the velocities seen in the two systems, take differentials of the Lorentz transformation formulae. Then take ratios of the corresponding infinitesimal position increments over the corresponding time increments.

D.5 Lorentz group property derivation

This note verifies the group property of the Lorentz transformation. It is not recommended unless you have had a solid course in linear algebra.

Note first that a much more simple argument can be given by defining the Lorentz transformation more abstractly, {A.4} (A.13). But that is cheating.

Then you have to prove that these Lorentz transform are always the same as the physical ones.

For simplicity it will be assumed that the observers still use a common origin of space and time coordinates.

The group property is easy to verify if the observers B and C are going in the same direction compared to A. Just multiply two matrices of the form (1.13) together and apply the condition that $\gamma^2 - \beta^2\gamma^2 = 1$ for each.

It gets much messier if the observers move in different directions. In that case the only immediate simplification that can be made is to align the coordinate systems so that both relative velocities are in the x, y planes. Then the transformations only involve z in a trivial way and the combined transformation takes the generic form

$$\Lambda_{C\leftarrow A} = \begin{pmatrix} \lambda^0{}_0 & \lambda^0{}_1 & \lambda^0{}_2 & 0 \\ \lambda^1{}_0 & \lambda^1{}_1 & \lambda^1{}_2 & 0 \\ \lambda^2{}_0 & \lambda^2{}_1 & \lambda^2{}_2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

It needs to be shown that this is a Lorentz transformation from A directly to C.

Now the spatial, x, y, coordinate system of observer C can be rotated to eliminate $\lambda^2{}_0$ and the spatial coordinate system of observer A can be rotated to eliminate $\lambda^0{}_2$. Next both Lorentz transformations preserve the inner products. Therefore the dot product between the four-vectors $(1, 0, 0, 0)$ and $(0, 0, 1, 0)$ in the A system must be the same as the dot product between columns 1 and 3 in the matrix above. And that means that $\lambda^1{}_2$ must be zero, because $\lambda^1{}_0$ will not be zero except in the trivial case that systems A and C are at rest compared to each other. Next since the proper length of the vector $(0, 0, 1, 0)$ equals one in the A system, it does so in the C system, so $\lambda^2{}_2$ must be one. (Or minus one, but a 180° rotation of the spatial coordinate system around the z-axis can take care of that.) Next, since the dot product of the vectors $(0, 1, 0, 0)$ and $(0, 0, 1, 0)$ is zero, so is $\lambda^2{}_1$.

That leaves the four values relating the time and x components. From the fact that the dot product of the vectors $(1, 0, 0, 0)$ and $(0, 1, 0, 0)$ is zero,

$$-\lambda^0{}_0\lambda^0{}_1 + \lambda^1{}_0\lambda^1{}_1 = 0 \qquad \Longrightarrow \qquad \frac{\lambda^0{}_1}{\lambda^1{}_1} = \frac{\lambda^1{}_0}{\lambda^0{}_0} \equiv \beta$$

where β is some constant. Also, since the proper lengths of these vectors are minus one, respectively one,

$$-\lambda^{02}_0 + \lambda^{12}_0 = -1 \qquad -\lambda^{02}_1 + \lambda^1{}_1 = 1$$

or substituting in for $\lambda^0{}_1$ and $\lambda^1{}_0$ from the above

$$-\lambda^{02}_0 + \beta^2\lambda^{02}_0 = -1 \qquad -\beta^2\lambda^{12}_1 + \lambda^1{}_1 = 1$$

It follows that ${\lambda^0}_0$ and ${\lambda^1}_1$ must be equal, (or opposite, but since both Lorentz transformations have unit determinant, so must their combination), so call them γ. The transformation is then a Lorentz transformation of the usual form (1.13). (Since the spatial coordinate system cannot just flip over from left handed to right handed at some point, γ will have to be positive.) Examining the transformation of the origin $x_{\mathrm{A}} = y_{\mathrm{A}} = z_{\mathrm{A}} = 0$ identifies β as V/c, with V the relative velocity of system A compared to B, and then the above two equations identify γ as the Lorentz factor.

Obviously, if any two Lorentz transformations are equivalent to a single one, then by repeated application any arbitrary number of them are equivalent to a single one.

D.6 Lorentz force derivation

To derive the given Lorentz force from the given Lagrangian, plug the canonical momentum and the Lagrangian into the Lagrangian equation of motion. That gives

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} + q\left(\frac{\partial A_i}{\partial t} + \frac{\partial A_i}{\partial x_j}v_j\right) = -q\frac{\partial \varphi}{\partial x_i} + q\frac{\partial A_j}{\partial x_i}v_j$$

This uses the Einstein convention that summation over j is to be understood. Reorder to get

$$\frac{\mathrm{d}p_i}{\mathrm{d}t} = q\left(-\frac{\partial \varphi}{\partial x_i} - \frac{\partial A_i}{\partial t}\right) + q\left(\frac{\partial A_j}{\partial x_i}v_j - \frac{\partial A_i}{\partial x_j}v_j\right)$$

The first parenthetical expression is the electric field as claimed. The quantity in the second parenthetical expression may be rewritten by expanding out the sums over j to give

$$\frac{\partial A_i}{\partial x_i}v_i - \frac{\partial A_i}{\partial x_i}v_i + \frac{\partial A_{\bar{\imath}}}{\partial x_i}v_{\bar{\imath}} - \frac{\partial A_i}{\partial x_{\bar{\imath}}}v_{\bar{\imath}} + \frac{\partial A_{\bar{\bar{\imath}}}}{\partial x_i}v_{\bar{\bar{\imath}}} - \frac{\partial A_i}{\partial x_{\bar{\bar{\imath}}}}v_{\bar{\bar{\imath}}}$$

where $\bar{\imath}$ follows i in the cyclic sequence $\ldots, 1, 2, 3, 1, 2, 3, \ldots$ and $\bar{\bar{\imath}}$ precedes it. The first two terms drop out and the others can be recognized as component number i of $\vec{v} \times (\nabla \times \vec{A})$. (For example, just write out the first component of $\vec{v} \times (\nabla \times \vec{A})$ and compare it the expression above for $\bar{\imath} = 2$ and $\bar{\bar{\imath}} = 3$.) Defining $\vec{\mathcal{B}}$ as $\nabla \times \vec{A}$, the Lorentz force law results.

D.7 Derivation of the Euler formula

To verify the Euler formula, write all three functions involved in terms of their Taylor series, [41, p. 136]

D.8 Completeness of Fourier modes

The purpose of this note is to show completeness of the "Fourier modes"

$$\ldots, \frac{e^{-3\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{-2\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{-\mathrm{i}x}}{\sqrt{2\pi}}, \frac{1}{\sqrt{2\pi}}, \frac{e^{\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{2\mathrm{i}x}}{\sqrt{2\pi}}, \frac{e^{3\mathrm{i}x}}{\sqrt{2\pi}}, \ldots$$

for describing functions that are periodic of period 2π. It is to be shown that "all" these functions can be written as combinations of the Fourier modes above. Assume that $f(x)$ is any reasonable smooth function that repeats itself after a distance 2π, so that $f(x+2\pi) = f(x)$. Then you can always write it in the form

$$f(x) = \ldots + c_{-2}\frac{e^{-2\mathrm{i}x}}{\sqrt{2\pi}} + c_{-1}\frac{e^{-\mathrm{i}x}}{\sqrt{2\pi}} + c_0\frac{1}{\sqrt{2\pi}} + c_1\frac{e^{\mathrm{i}x}}{\sqrt{2\pi}} + c_2\frac{e^{2\mathrm{i}x}}{\sqrt{2\pi}} + c_3\frac{e^{3\mathrm{i}x}}{\sqrt{2\pi}} + \ldots$$

or

$$f(x) = \sum_{k=-\infty}^{\infty} c_k \frac{e^{k\mathrm{i}x}}{\sqrt{2\pi}}$$

for short. Such a representation of a periodic function is called a "Fourier series." The coefficients c_k are called "Fourier coefficients." The factors $1/\sqrt{2\pi}$ can be absorbed in the definition of the Fourier coefficients, if you want.

Because of the Euler formula, the set of exponential Fourier modes above is completely equivalent to the set of real Fourier modes

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos(x)}{\sqrt{\pi}}, \frac{\sin(x)}{\sqrt{\pi}}, \frac{\cos(2x)}{\sqrt{\pi}}, \frac{\sin(2x)}{\sqrt{\pi}}, \frac{\cos(3x)}{\sqrt{\pi}}, \frac{\sin(3x)}{\sqrt{\pi}}, \ldots$$

so that 2π-periodic functions may just as well be written as

$$f(x) = a_0\frac{1}{\sqrt{2\pi}} + \sum_{k=1}^{\infty} a_k \frac{\cos(kx)}{\sqrt{\pi}} + \sum_{k=1}^{\infty} b_k \frac{\sin(kx)}{\sqrt{\pi}}.$$

The extension to functions that are periodic of some other period than 2π is a trivial matter of rescaling x. For a period 2ℓ, with ℓ any half period, the exponential Fourier modes take the more general form

$$\ldots, \frac{e^{-k_2\mathrm{i}x}}{\sqrt{2\ell}}, \frac{e^{-k_1\mathrm{i}x}}{\sqrt{2\ell}}, \frac{1}{\sqrt{2\ell}}, \frac{e^{k_1\mathrm{i}x}}{\sqrt{2\ell}}, \frac{e^{k_2\mathrm{i}x}}{\sqrt{2\ell}}, \ldots \qquad k_1 = \frac{1\pi}{\ell},\ k_2 = \frac{2\pi}{\ell},\ k_3 = \frac{3\pi}{\ell}, \ldots$$

and similarly the real version of them becomes

$$\frac{1}{\sqrt{2\ell}}, \frac{\cos(k_1x)}{\sqrt{\ell}}, \frac{\sin(k_1x)}{\sqrt{\ell}}, \frac{\cos(k_2x)}{\sqrt{\ell}}, \frac{\sin(k_2x)}{\sqrt{\ell}}, \frac{\cos(k_3x)}{\sqrt{\ell}}, \frac{\sin(k_3x)}{\sqrt{\ell}}, \ldots$$

See [41, p. 141] for detailed formulae.

Often, the functions of interest are not periodic, but are required to be zero at the ends of the interval on which they are defined. Those functions can be handled too, by extending them to a periodic function. For example, if the functions $f(x)$ relevant to a problem are defined only for $0 \leqslant x \leqslant \ell$ and must satisfy $f(0) = f(\ell) = 0$, then extend them to the range $-\ell \leqslant x \leqslant 0$ by setting $f(x) = -f(-x)$ and take the range $-\ell \leqslant x \leqslant \ell$ to be the period of a 2ℓ-periodic function. It may be noted that for such a function, the cosines disappear in the real Fourier series representation, leaving only the sines. Similar extensions can be used for functions that satisfy symmetry or zero-derivative boundary conditions at the ends of the interval on which they are defined. See again [41, p. 141] for more detailed formulae.

If the half period ℓ becomes infinite, the spacing between the discrete k values becomes zero and the sum over discrete k values turns into an integral over continuous k values. This is exactly what happens in quantum mechanics for the eigenfunctions of linear momentum. The representation is now no longer called a Fourier series, but a "Fourier integral." And the Fourier coefficients c_k are now called the "Fourier transform" $F(k)$. The completeness of the eigenfunctions is now called Fourier's integral theorem or inversion theorem. See [41, pp. 190-191] for more.

The basic completeness proof is a rather messy mathematical derivation, so read the rest of this note at your own risk. The fact that the Fourier modes are orthogonal and normalized was the subject of various exercises in chapter 2.6 and will be taken for granted here. See the solution manual for the details. What this note wants to show is that *any* arbitrary periodic function f of period 2π that has continuous first and second order derivatives can be written as

$$f(x) = \sum_{k=-\infty}^{k=\infty} c_k \frac{e^{k\mathrm{i}x}}{\sqrt{2\pi}},$$

in other words, as a combination of the set of Fourier modes.

First an expression for the values of the Fourier coefficients c_k is needed. It can be obtained from taking the inner product $\langle e^{lix}/\sqrt{2\pi}|f(x)\rangle$ between a generic eigenfunction $e^{lix}/\sqrt{2\pi}$ and the representation for function $f(x)$ above. Noting that all the inner products with the exponentials representing $f(x)$ will be zero except the one for which $k = l$, if the Fourier representation is indeed correct, the coefficients need to have the values

$$c_l = \int_{x=0}^{2\pi} \frac{e^{-lix}}{\sqrt{2\pi}} f(x)\,\mathrm{d}x,$$

a requirement that was already noted by Fourier. Note that l and x are just names for the eigenfunction number and the integration variable that you can change at will. Therefore, to avoid name conflicts later, the expression will be

renotated as

$$c_k = \int_{\bar{x}=0}^{2\pi} \frac{e^{-k\mathrm{i}\bar{x}}}{\sqrt{2\pi}} f(\bar{x})\,\mathrm{d}\bar{x},$$

Now the question is: suppose you compute the Fourier coefficients c_k from this expression, and use them to sum many terms of the infinite sum for $f(x)$, say from some very large negative value $-K$ for k to the corresponding large positive value K; in that case, is the result you get, call it $f_K(x)$,

$$f_K(x) \equiv \sum_{k=-K}^{k=K} c_k \frac{e^{k\mathrm{i}x}}{\sqrt{2\pi}},$$

a valid approximation to the true function $f(x)$? More specifically, if you sum more and more terms (make K bigger and bigger), does $f_K(x)$ reproduce the true value of $f(x)$ to any arbitrary accuracy that you may want? If it does, then the eigenfunctions are capable of reproducing $f(x)$. If the eigenfunctions are not complete, a definite difference between $f_K(x)$ and $f(x)$ will persist however large you make K. In mathematical terms, the question is whether $\lim_{K\to\infty} f_K(x) = f(x)$.

To find out, the trick is to substitute the integral for the coefficients c_k into the sum and then reverse the order of integration and summation to get:

$$f_K(x) = \frac{1}{2\pi} \int_{\bar{x}=0}^{2\pi} f(\bar{x}) \left[\sum_{k=-K}^{k=K} e^{k\mathrm{i}(x-\bar{x})} \right] \mathrm{d}\bar{x}.$$

The sum in the square brackets can be evaluated, because it is a geometric series with starting value $e^{-K\mathrm{i}(x-\bar{x})}$ and ratio of terms $e^{\mathrm{i}(x-\bar{x})}$. Using the formula from [41, item 21.4], multiplying top and bottom with $e^{-\mathrm{i}(x-\bar{x})/2}$, and cleaning up with, what else, the Euler formula, the sum is found to equal

$$\frac{\sin\Big((K+\frac{1}{2})(x-\bar{x})\Big)}{\sin\Big(\frac{1}{2}(x-\bar{x})\Big)}.$$

This expression is called the "Dirichlet kernel". You now have

$$f_K(x) = \int_{\bar{x}=0}^{2\pi} f(\bar{x}) \frac{\sin\Big((K+\frac{1}{2})(x-\bar{x})\Big)}{2\pi \sin\Big(\frac{1}{2}(x-\bar{x})\Big)}\,\mathrm{d}\bar{x}.$$

The second trick is to split the function $f(\bar{x})$ being integrated into the two parts $f(x)$ and $f(\bar{x}) - f(x)$. The sum of the parts is obviously still $f(\bar{x})$, but the first part has the advantage that it is constant during the integration over

$\bar{x}$ and can be taken out, and the second part has the advantage that it becomes zero at $\bar{x} = x$. You get

$$f_K(x) = f(x) \int_{\bar{x}=0}^{2\pi} \frac{\sin\Big((K+\frac{1}{2})(x-\bar{x})\Big)}{2\pi \sin\Big(\frac{1}{2}(x-\bar{x})\Big)}\,\mathrm{d}\bar{x} + \int_{\bar{x}=0}^{2\pi} \Big(f(\bar{x}) - f(x)\Big) \frac{\sin\Big((K+\frac{1}{2})(x-\bar{x})\Big)}{2\pi \sin\Big(\frac{1}{2}(x-\bar{x})\Big)}\,\mathrm{d}\bar{x}.$$

Now if you backtrack what happens in the trivial case that $f(x)$ is just a constant, you find that $f_K(x)$ is exactly equal to $f(x)$ in that case, while the second integral above is zero. That makes the first integral above equal to one. Returning to the case of general $f(x)$, since the first integral above is still one, it makes the first term in the right hand side equal to the desired $f(x)$, and the second integral is then the error in $f_K(x)$.

To manipulate this error and show that it is indeed small for large K, it is convenient to rename the K-independent part of the integrand to

$$g(\bar{x}) = \frac{f(\bar{x}) - f(x)}{2\pi \sin\Big(\frac{1}{2}(x-\bar{x})\Big)}$$

Using l'Hôpital's rule twice, it is seen that since by assumption f has a continuous second derivative, g has a continuous first derivative. So you can use one integration by parts to get

$$f_K(x) = f(x) + \frac{1}{K+\frac{1}{2}} \int_{\bar{x}=0}^{2\pi} g'(\bar{x}) \cos\Big((K+\tfrac{1}{2})(x-\bar{x})\Big)\,\mathrm{d}\bar{x}.$$

And since the integrand of the final integral is continuous, it is bounded. That makes the error inversely proportional to $K + \frac{1}{2}$, implying that it does indeed become arbitrarily small for large K. Completeness has been proved.

It may be noted that under the stated conditions, the convergence is uniform; there is a guaranteed minimum rate of convergence regardless of the value of x. This can be verified from Taylor series with remainder. Also, the more continuous derivatives the 2π-periodic function $f(x)$ has, the faster the rate of convergence, and the smaller the number $2K + 1$ of terms that you need to sum to get good accuracy is likely to be. For example, if $f(x)$ has three continuous derivatives, you can do another integration by parts to show that the convergence is proportional to $1/(K+\frac{1}{2})^2$ rather than just $1/(K+\frac{1}{2})$. But watch the end points: if a derivative has different values at the start and end of the period, then that derivative is not continuous, it has a jump at the ends. (Such jumps can be incorporated in the analysis, however, and have less effect

than it may seem. You get a better practical estimate of the convergence rate by directly looking at the integral for the Fourier coefficients.)

The condition for $f(x)$ to have a continuous second derivative can be relaxed with more work. If you are familiar with the Lebesgue form of integration, it is fairly easy to extend the result above to show that it suffices that the absolute integral of f^2 exists, something that will be true in quantum mechanics applications.

D.9 Momentum operators are Hermitian

To check that the linear momentum operators are Hermitian, assume that Ψ_1 and Ψ_2 are any two proper, reasonably behaved, wave functions. By definition:

$$\langle \Psi_1 | \widehat{p}_x \Psi_2 \rangle = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \Psi_1^* \frac{\hbar}{\mathrm{i}} \Psi_{2,x} \, \mathrm{d}x \mathrm{d}y \mathrm{d}z$$

Here the subscript x indicates differentiation with respect to x. This can be rewritten as

$$\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \left[\left(\Psi_1^* \frac{\hbar}{\mathrm{i}} \Psi_2 \right)_x - \Psi_{1,x}^* \frac{\hbar}{\mathrm{i}} \Psi_2 \right] \mathrm{d}x \mathrm{d}y \mathrm{d}z \tag{1}$$

as can be checked by simply differentiating out the product in the first term.

Now the first term in the integral can be integrated with respect to x and is then seen to produce zero. The reason is that Ψ_1 and Ψ_2 must become zero at large distances, otherwise their square integral cannot be zero. That leaves only the second term. And that equals

$$\langle \widehat{p}_x \Psi_1 | \Psi_2 \rangle = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \left(\frac{\hbar}{\mathrm{i}} \frac{\partial \Psi_1}{\partial x} \right)^* \Psi_2 \, \mathrm{d}x \mathrm{d}y \mathrm{d}z$$

(Recall that the complex conjugate of i is $-\mathrm{i}$, hence the minus sign.)

For the mathematically picky, it is maybe a good idea to examine the claim that the first term in the integral in (1) integrates to zero a bit more closely. The integral is definitely zero if the system is not in infinite space, but in a periodic box. The reason is that in that case the lower and upper limits of integration are equal and drop out against each other. To be rigorous in infinite space, you will at first need to limit the region of integration to a distance no more than some large number R away from the origin. (It will be assumed that the initial inner product is well defined, in the sense that the integral has a finite limit in the limit $R \to \infty$. In that case, the integral can be approximated to arbitrary accuracy by just taking R large enough.) Using the divergence theorem, the first integral is

$$\int_S \left(\Psi_1^* \frac{\hbar}{\mathrm{i}} \Psi_2 \right) n_x \, \mathrm{d}S$$

where S is the surface of the sphere $r = R$ and n_x the x-component of the unit vector $\hat{\imath}_r$ normal to the surface. Now since the absolute square integrals of the wave functions are finite, for large enough R their square integrals outside R become arbitrarily small. The Cauchy Schwartz inequality then says that the above integral integrated with respect to r must become vanishingly small. And that is not possible unless the integral itself becomes vanishingly small at almost all locations. So you can define a sequence for R where the inner product with the linear momentum swapped over approaches the original inner product. In particular, the two inner products are equal to the degree that they are well defined in the first place.

D.10 The curl is Hermitian

For later reference, it will be shown that the curl operator, $\nabla\times$ is Hermitian. In other words,

$$\int_{\text{all}} \vec{A}^* \cdot \nabla \times \vec{B} \mathrm{d}^3\vec{r} = \int_{\text{all}} \nabla \times \vec{A}^* \cdot \vec{B} \mathrm{d}^3\vec{r}$$

The rules of engagement are as follows:

- The Cartesian axes are numbered using an index i, with $i = 1$, 2, and 3 for x, y, and z respectively.
- Also, r_i indicates the coordinate in the i direction, x, y, or z.
- Derivatives with respect to a coordinate r_i are indicated by a simple subscript i.
- If the quantity being differentiated is a vector, a comma is used to separate the vector index from differentiation ones.
- Index $\bar{\imath}$ is the number immediately following i in the cyclic sequence $\ldots 123123\ldots$ and $\bar{\bar{\imath}}$ is the number immediately preceding i.
- A bare $\int$ integral sign is assumed to be an integration over all space, or over the entire box for particles in a box. The $\mathrm{d}^3\vec{r}$ is normally omitted for brevity and to be understood.
- A superscript * indicates a complex conjugate.

In index notation, the integral in the left hand side above reads:

$$\sum_i \int A_i^* (B_{\bar{\bar{\imath}},\bar{\imath}} - B_{\bar{\imath},\bar{\bar{\imath}}})$$

which is the same as

$$\sum_i \int [(A_i^* B_{\bar{\bar{\imath}}})_{\bar{\imath}} - (A_i^* B_{\bar{\imath}})_{\bar{\bar{\imath}}} - A^*_{i,\bar{\imath}} B_{\bar{\bar{\imath}}} + A^*_{i,\bar{\bar{\imath}}} B_{\bar{\imath}}]$$

as can be checked by differentiating out the first two terms. Now the third and fourth terms in the integral are $\nabla \times \vec{A}^* \cdot \vec{B}$, as you can see from moving all

indices in the third term one unit forward in the cyclic sequence, and those in the fourth term one unit back. (Such a shift does not change the sum; the same terms are simply added in a different order.)

So, if the integral of the first two terms is zero, the fact that curl is Hermitian has been verified. Note that the terms can be integrated. Then, if the system is in a periodic box, the integral is indeed zero because the upper and lower limits of integration are equal. An infinite domain will need to be truncated at some large distance R from the origin. Then shift indices and apply the divergence theorem to get

$$-\int_S (\vec{A}^* \times \vec{B}) \cdot \hat{\imath}_r \, \mathrm{d}S$$

where S is the surface of the sphere $r = R$ and $\hat{\imath}_r$ the unit vector normal to the sphere surface. It follows that the integral is zero if $\vec{A}$ and $\vec{B}$ go to zero at infinity quickly enough. Or at least their cross product has to go to zero quickly enough.

D.11 Extension to three-dimensional solutions

Maybe you have some doubt whether you really can just multiply one-dimensional eigenfunctions together, and add one-dimensional energy values to get the three-dimensional ones. Would a book that you find for free on the Internet lie? OK, let's look at the details then. First, the three-dimensional Hamiltonian, (really just the kinetic energy operator), is the sum of the one-dimensional ones:

$$H = H_x + H_y + H_z$$

where the one-dimensional Hamiltonians are:

$$H_x = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} \qquad H_y = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial y^2} \qquad H_z = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial z^2}$$

To check that any product $\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z)$ of one-dimensional eigenfunctions is an eigenfunction of the combined Hamiltonian H, note that the partial Hamiltonians only act on their own eigenfunction, multiplying it by the corresponding eigenvalue:

$$\begin{aligned}(H_x + H_y + H_z)&\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z)\\ &= E_x\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z) + E_y\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z) + E_z\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z)\end{aligned}$$

or

$$H\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z) = (E_x + E_y + E_z)\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z).$$

Therefore, by definition $\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z)$ is an eigenfunction of the three-dimensional Hamiltonian, with an eigenvalue that is the sum of the three one-dimensional ones. But there is still the question of completeness. Maybe the

above eigenfunctions are not complete, which would mean a need for additional eigenfunctions that are not products of one-dimensional ones.

The one-dimensional eigenfunctions $\psi_{n_x}(x)$ *are* complete, see [41, p. 141] and earlier exercises in this book. So, you can write any wave function $\Psi(x,y,z)$ *at given values of* y *and* z as a combination of x-eigenfunctions:

$$\Psi(x,y,z) = \sum_{n_x} c_{n_x} \psi_{n_x}(x),$$

but the coefficients c_{n_x} will be different for different values of y and z; in other words they will be functions of y and z: $c_{n_x} = c_{n_x}(y,z)$. So, more precisely, you have

$$\Psi(x,y,z) = \sum_{n_x} c_{n_x}(y,z) \psi_{n_x}(x),$$

But since the y-eigenfunctions are also complete, *at any given value of* z, you can write each $c_{n_x}(y,z)$ as a sum of y-eigenfunctions:

$$\Psi(x,y,z) = \sum_{n_x} \left(\sum_{n_y} c_{n_x n_y} \psi_{n_y}(y) \right) \psi_{n_x}(x),$$

where the coefficients $c_{n_x n_y}$ will be different for different values of z, $c_{n_x n_y} = c_{n_x n_y}(z)$. So, more precisely,

$$\Psi(x,y,z) = \sum_{n_x} \left(\sum_{n_y} c_{n_x n_y}(z) \psi_{n_y}(y) \right) \psi_{n_x}(x),$$

But since the z-eigenfunctions are also complete, you can write $c_{n_x n_y}(z)$ as a sum of z-eigenfunctions:

$$\Psi(x,y,z) = \sum_{n_x} \left(\sum_{n_y} \left(\sum_{n_z} c_{n_x n_y n_z} \psi_{n_z}(z) \right) \psi_{n_y}(y) \right) \psi_{n_x}(x).$$

Since the order of doing the summation does not make a difference,

$$\Psi(x,y,z) = \sum_{n_x} \sum_{n_y} \sum_{n_z} c_{n_x n_y n_z} \psi_{n_x}(x) \psi_{n_y}(y) \psi_{n_z}(z).$$

So, any wave function $\Psi(x,y,z)$ can be written as a sum of products of one-dimensional eigenfunctions; these products are complete.

D.12 The harmonic oscillator solution

If you really want to know how the harmonic oscillator wave function can be found, here it is. Read at your own risk.

The ODE (ordinary differential equation) to solve is

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi_x}{\partial x^2} + \tfrac{1}{2}m\omega^2 x^2\psi_x = E_x\psi_x$$

where the spring constant c was rewritten as the equivalent expression $m\omega^2$.

Now the first thing you always want to do with this sort of problems is to simplify it as much as possible. In particular, get rid of as much dimensional constants as you can by rescaling the variables: define a new scaled x-coordinate ξ and a scaled energy ϵ by

$$x \equiv \ell\xi \qquad E_x \equiv E_0\epsilon.$$

If you make these replacements into the ODE above, you can make the coefficients of the two terms in the left hand side equal by choosing $\ell = \sqrt{\hbar/m\omega}$. In that case both terms will have the same net coefficient $\frac{1}{2}\hbar\omega$. Then if you cleverly choose $E_0 = \frac{1}{2}\hbar\omega$, the right hand side will have that coefficient too, and you can divide it away and end up with no coefficients at all:

$$-\frac{\partial^2\psi_x}{\partial\xi^2} + \xi^2\psi_x = \epsilon\psi_x$$

Looks a lot cleaner, not?

Now examine this equation for large values of ξ (i.e. large x). You get approximately

$$\frac{\partial^2\psi_x}{\partial\xi^2} \approx \xi^2\psi_x + \ldots$$

If you write the solution as an exponential, you can ballpark that it must take the form

$$\psi_x = e^{\pm\frac{1}{2}\xi^2+\cdots}$$

where the dots indicate terms that are small compared to $\frac{1}{2}\xi^2$ for large ξ. The form of the solution is important, since $e^{+\frac{1}{2}\xi^2}$ becomes infinitely large at large ξ. That is unacceptable: the probability of finding the particle cannot become infinitely large at large x: the total probability of finding the particle must be one, not infinite. The *only* solutions that are acceptable are those that behave as $e^{-\frac{1}{2}\xi^2+\cdots}$ for large ξ.

Now split off the leading exponential part by defining a new unknown $h(\xi)$ by

$$\psi_x \equiv e^{-\frac{1}{2}\xi^2}h(\xi)$$

Substituting this in the ODE and dividing out the exponential, you get:

$$-\frac{\partial^2 h}{\partial \xi^2} + 2\xi \frac{\partial h}{\partial \xi} + h = \epsilon h$$

Now try to solve this by writing h as a power series, (say, a Taylor series):

$$h = \sum_p c_p \xi^p$$

where the values of p run over whatever the appropriate powers are and the c_p are constants. If you plug this into the ODE, you get

$$\sum_p p(p-1) c_p \xi^{p-2} = \sum_p (2p + 1 - \epsilon) c_p \xi^p$$

For the two sides to be equal, they must have the same coefficient for every power of ξ.

There must be a lowest value of p for which there is a nonzero coefficient c_p, for if p took on arbitrarily large negative values, h would blow up strongly at the origin, and the probability to find the particle near the origin would then be infinite. Denote the lowest value of p by q. This lowest power produces a power of ξ^{q-2} in the left hand side of the equation above, but there is no corresponding power in the right hand side. So, the coefficient $q(q-1)c_q$ of ξ^{q-2} will need to be zero, and that means either $q = 0$ or $q = 1$. So the power series for h will need to start as either $c_0 + \ldots$ or $c_1\xi + \ldots$. The constant c_0 or c_1 is allowed to have any nonzero value.

But note that the $c_q\xi^q$ term normally produces a term $(2q+1-\epsilon)c_q\xi^q$ in the right hand side of the equation above. For the left hand side to have a matching ξ^q term, there will need to be a further $c_{q+2}\xi^{q+2}$ term in the power series for h,

$$h = c_q\xi^q + c_{q+2}\xi^{q+2} + \ldots$$

where $(q+2)(q+1)c_{q+2}$ will need to equal $(2q+1-\epsilon)c_q$, so $c_{q+2} = (2q+1-\epsilon)c_q/(q+2)(q+1)$. This term in turn will normally produce a term $\Big(2(q+2)+1-\epsilon\Big)c_{q+2}\xi^{q+2}$ in the right hand side which will have to be canceled in the left hand side by a $c_{q+4}\xi^{q+4}$ term in the power series for h. And so on.

So, if the power series starts with $q = 0$, the solution will take the general form

$$h = c_0 + c_2\xi^2 + c_4\xi^4 + c_6\xi^6 + \ldots$$

while if it starts with $q = 1$ you will get

$$h = c_1\xi + c_3\xi^3 + c_5\xi^5 + c_7\xi^7 + \ldots$$

In the first case, you have a symmetric solution, one which remains the same when you flip over the sign of ξ, and in the second case you have an antisymmetric solution, one which changes sign when you flip over the sign of ξ.

You can find a general formula for the coefficients of the series by making the change in notations $p = 2 + \bar{p}$ in the left-hand-side sum:

$$\sum_{\bar{p}=q} (\bar{p}+2)(\bar{p}+1)c_{\bar{p}+2}\xi^{\bar{p}} = \sum_{p=q} (2p+1-\epsilon)c_p\xi^p$$

Note that you can start summing at $\bar{p} = q$ rather than $q - 2$, since the first term in the sum is zero anyway. Next note that you can again forget about the difference between $\bar{p}$ and p, because it is just a symbolic summation variable. The symbolic sum writes out to the exact same actual sum whether you call the symbolic summation variable p or $\bar{p}$.

So for the powers in the two sides to be equal, you must have

$$c_{p+2} = \frac{2p+1-\epsilon}{(p+2)(p+1)} c_p$$

In particular, for large p, by approximation

$$c_{p+2} \approx \frac{2}{p} c_p$$

Now if you check out the Taylor series of e^{ξ^2}, (i.e. the Taylor series of e^x with x replaced by ξ^2,) you find it satisfies the exact same equation. So, normally the solution h blows up something like e^{ξ^2} at large ξ. And since ψ_x was $e^{-\frac{1}{2}\xi^2}h$, normally ψ_x takes on the unacceptable form $e^{+\frac{1}{2}\xi^2+\cdots}$. (If you must have rigor here, estimate h in terms of $Ce^{\alpha\xi^2}$ where α is a number slightly less than one, plus a polynomial. That is enough to show unacceptability of such solutions.)

What are the options for acceptable solutions? The only possibility is that the power series terminates. There must be a highest power p, call it $p = n$, whose term in the right hand side is zero

$$0 = (2n+1-\epsilon)c_n\xi^n$$

In that case, there is no need for a further $c_{n+2}\xi^{n+2}$ term, the power series will remain a polynomial of degree n. But note that all this requires the scaled energy ϵ to equal $2n + 1$, and the actual energy E_x is therefore $(2n + 1)\hbar\omega/2$. Different choices for the power at which the series terminates produce different energies and corresponding eigenfunctions. But they are discrete, since n, as any power p, must be a nonnegative integer.

With ϵ identified as $2n + 1$, you can find the ODE for h listed in table books, like [41, 29.1], under the name "Hermite's differential equation." They then identify the polynomial solutions as the so-called "Hermite polynomials,"

except for a normalization factor. To find the normalization factor, i.e. c_0 or c_1, demand that the total probability of finding the particle anywhere is one, $\int_{-\infty}^{\infty} |\psi_x|^2 \, \mathrm{d}x = 1$. You should be able to find the value for the appropriate integral in your table book, like [41, 29.15].

Putting it all together, the generic expression for the eigenfunctions can be found to be:

$$h_n = \frac{1}{(\pi \ell^2)^{1/4}} \frac{H_n(\xi)}{\sqrt{2^n n!}} e^{-\xi^2/2} \qquad n = 0, 1, 2, 3, 4, 5, \ldots \tag{D.4}$$

where the details of the "Hermite polynomials" H_n can be found in table books like [41, pp. 167-168]. They are readily evaluated on a computer using the "recurrence relation" you can find there, for as far as computer round-off error allows (up to n about 70.)

Quantum field theory allows a much neater way to find the eigenfunctions. It is explained in addendum {A.15.5} or equivalently in {D.64}.

D.13 The harmonic oscillator and uncertainty

The given qualitative explanation of the ground state of the harmonic oscillator in terms of the uncertainty principle is questionable. In particular, position, linear momentum, potential energy, and kinetic energy are uncertain for the ground state. This note gives a solid argument, but it uses some advanced ideas discussed in chapter 4.4 and 4.5.3.

As explained more fully in chapter 4.4, the "expectation value" of the kinetic energy is defined as the average value expected for kinetic energy measurements. Similarly, the expectation value of the potential energy is defined as the average value expected for potential energy measurements.

From the precise form of expectation values in quantum mechanics, it follows that total energy must be the sum of the kinetic and potential energy expectation values. For the harmonic oscillator ground state, that gives

$$E_{x0} = \tfrac{1}{2}\hbar\omega = \frac{1}{2m} \left\langle p_x^2 \right\rangle + \frac{m}{2} \omega^2 \left\langle x^2 \right\rangle$$

Here $\langle . \rangle$ stands for the average of the enclosed quantity. Only the motion in the x-direction will be considered here. The y and z directions go exactly the same way.

Now any value of p_x can be written as equal to the average value $\langle p_x \rangle$ plus a deviation from that average Δp_x. Then

$$\left\langle p_x^2 \right\rangle = \Big\langle \left(\langle p_x \rangle + \Delta p_x \right)^2 \Big\rangle = \langle p_x \rangle^2 + 2 \langle p_x \rangle \langle \Delta p_x \rangle + \left\langle (\Delta p_x)^2 \right\rangle$$

Note that an average is a constant that is not affected by further averaging. Next note that the average of Δp_x is zero, otherwise the average of $p_x + \Delta p_x$

would not be $\langle p_x \rangle$. So:

$$\left\langle p_x^2 \right\rangle = \langle p_x \rangle^2 + \left\langle (\Delta p_x)^2 \right\rangle$$

Of course, a similar expression holds for $\langle x^2 \rangle$, so the ground state energy is

$$\tfrac{1}{2}\hbar\omega = \frac{1}{2m} \langle p_x \rangle^2 + \frac{m}{2}\omega^2 \langle x \rangle^2 + \frac{1}{2m} \left\langle (\Delta p_x)^2 \right\rangle + \frac{m}{2}\omega^2 \left\langle (\Delta x)^2 \right\rangle \tag{1}$$

Consider the last two terms. Call them a^2 and b^2 for now. Note that

$$(a-b)^2 \geqslant 0 \quad \Longrightarrow \quad a^2 + b^2 \geqslant 2ab = \omega \sqrt{\langle (\Delta p_x)^2 \rangle} \sqrt{\langle (\Delta x)^2 \rangle}$$

as follows from multiplying out the square. The $\geqslant$ becomes $=$ when a and b are equal.

Now the first square root above is a measure of the uncertainty in p_x. If Δp_x is always zero, then p_x is always its average value, without any uncertainty. Similarly, the second square root above is a measure of the uncertainty in x. The Heisenberg uncertainty principle can be made quantitative as, chapter 4.5.3,

$$\sqrt{\langle (\Delta p_x)^2 \rangle} \sqrt{\langle (\Delta x)^2 \rangle} \geqslant \tfrac{1}{2}\hbar$$

Therefore

$$a^2 + b^2 \geqslant \tfrac{1}{2}\hbar\omega$$

So the *minimum* value of the final two terms in the expression (1) for the ground state energy is the complete ground state energy. Therefore, in order that the right hand side in (1) does not exceed the left hand side, the first two terms must be zero. So the average particle momentum and position are both zero. In addition, for the estimates of the final two terms, equalities are needed, not inequalities. That means that a must be b. That then means that the expectation kinetic energy must be the expectation potential energy. And the two must be the very minimum allowed by the Heisenberg relation; otherwise there is still that inequality.

D.14 The spherical harmonics

This note derives and lists properties of the spherical harmonics.

D.14.1 Derivation from the eigenvalue problem

This analysis will derive the spherical harmonics from the eigenvalue problem of square angular momentum of chapter 4.2.3. It will use similar techniques as for the harmonic oscillator solution, {D.12}.

The imposed additional requirement that the spherical harmonics Y_l^m are eigenfunctions of L_z means that they are of the form $\Theta_l^m(\theta)e^{\mathrm{i}m\phi}$ where function

$\Theta_l^m(\theta)$ is still to be determined. (There is also an arbitrary dependence on the radius r, but it does not have anything to do with angular momentum, hence is ignored when people define the spherical harmonics.) Substitution into $\widehat{L}^2\psi = L^2\psi$ with $\widehat{L}^2$ as in (4.22) yields an ODE (ordinary differential equation) for $\Theta_l^m(\theta)$:

$$-\frac{\hbar^2}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Theta_l^m}{\partial\theta}\right)+\frac{\hbar^2m^2}{\sin^2\theta}\Theta_l^m = L^2\Theta_l^m$$

It is convenient define a scaled square angular momentum by $L^2 = \hbar^2\lambda^2$ so that you can divide away the $\hbar^2$ from the ODE.

More importantly, recognize that the solutions will likely be in terms of cosines and sines of θ, because they should be periodic if θ changes by 2π. If you want to use power-series solution procedures again, these transcendental functions are bad news, so switch to a new variable $x = \cos\theta$. At the very least, that will reduce things to algebraic functions, since $\sin\theta$ is in terms of $x = \cos\theta$ equal to $\sqrt{1-x^2}$. Converting the ODE to the new variable x, you get

$$-(1-x^2)\frac{\mathrm{d}^2\Theta_l^m}{\mathrm{d}x^2}+2x\frac{\mathrm{d}\Theta_l^m}{\mathrm{d}x}+\frac{m^2}{1-x^2}\Theta_l^m = \lambda^2\Theta_l^m$$

As you may guess from looking at this ODE, the solutions Θ_l^m are likely to be problematic near $x = \pm1$, (physically, near the z-axis where $\sin\theta$ is zero.) If you examine the solution near those points by defining a local coordinate ξ as in $x = \pm(1-\xi)$, and then deduce the leading term in the power series solutions with respect to ξ, you find that it is either $\xi^{m/2}$ or $\xi^{-m/2}$, (in the special case that $m = 0$, that second solution turns out to be $\ln\xi$.) Either way, the second possibility is not acceptable, since it physically would have infinite derivatives at the z-axis and a resulting expectation value of square momentum, as defined in chapter 4.4.3, that is infinite. You need to have that Θ_l^m behaves as $\xi^{m/2}$ at each end, so in terms of x it must have a factor $(1-x)^{m/2}$ near $x = 1$ and $(1+x)^{m/2}$ near $x = -1$. The two factors multiply to $(1-x^2)^{m/2}$ and so Θ_l^m can be written as $(1-x^2)^{m/2}f_l^m$ where f_l^m must have finite values at $x = 1$ and $x = -1$.

If you substitute $\Theta_l^m = (1-x^2)^{m/2}f_l^m$ into the ODE for Θ_l^m, you get an ODE for f_l^m:

$$-(1-x^2)\frac{\mathrm{d}^2f_l^m}{\mathrm{d}x^2}+2(1+m)x\frac{\mathrm{d}f_l^m}{\mathrm{d}x}+(m^2+m)f_l^m = \lambda^2f_l^m$$

Plug in a power series, $f_l^m = \sum c_px^p$, to get, after clean up,

$$\sum p(p-1)c_px^{p-2} = \sum\left[(p+m)(p+m+1)-\lambda^2\right]c_px^p$$

Using similar arguments as for the harmonic oscillator, you see that the starting power will be zero or one, leading to basic solutions that are again odd or even.

And just like for the harmonic oscillator, you must again have that the power series terminates; even in the least case that $m = 0$, the series for f_l^m at $|x| = 1$ is like that of $\ln(1 - x^2)$ and will not converge to the finite value stipulated. (For rigor, use Gauss's test.)

To get the series to terminate at some final power $p = n$, you must have according to the above equation that $\lambda^2 = (n+m)(n+m+1)$, and if you decide to call $n + m$ the azimuthal quantum number l, you have $\lambda^2 = l(l + 1)$ where $l \geqslant m$ since $l = n + m$ and n, like any power p, is greater or equal to zero.

The rest is just a matter of table books, because with $\lambda^2 = l(l + 1)$, the ODE for f_l^m is just the m-th derivative of the differential equation for the L_l Legendre polynomial, [41, 28.1], so the f_l^m must be just the m-th derivative of those polynomials. In fact, you can now recognize that the ODE for the Θ_l^m is just Legendre's associated differential equation [41, 28.49], and that the solutions that you need are the associated Legendre functions of the first kind [41, 28.50].

To normalize the eigenfunctions on the surface area of the unit sphere, find the corresponding integral in a table book, like [41, 28.63]. As mentioned at the start of this long and still very condensed story, to include negative values of m, just replace m by $|m|$. There is one additional issue, though, the sign pattern. In order to simplify some more advanced analysis, physicists like the sign pattern to vary with m according to the so-called "ladder operators." That requires, {D.64}, that starting from $m = 0$, the spherical harmonics for $m > 0$ have the alternating sign pattern of the "ladder-up operator," and those for $m < 0$ the unvarying sign of the "ladder-down operator." Physicists will still allow you to select your own sign for the $m = 0$ state, bless them.

The final solution is

$$Y_l^m(\theta, \phi) = (-1)^{\max(m,0)} \sqrt{\frac{2l+1}{4\pi} \frac{(l-|m|)!}{(l+|m|)!}} P_l^{|m|}(\cos\theta) e^{\mathrm{i}m\phi} \tag{D.5}$$

where the properties of the associated Legendre functions of the first kind $P_l^{|m|}$ can be found in table books like [41, pp. 162-166]. This uses the following definition of the associated Legendre polynomials:

$$P_l^m(x) \equiv (1 - x^2)^{m/2} \frac{\mathrm{d}^m P_l(x)}{\mathrm{d}x^m}$$

where P_l is the normal Legendre polynomial. Needless to say, some other authors use different definitions, potentially putting in a factor $(-1)^m$.

D.14.2 Parity

One special property of the spherical harmonics is often of interest: their "parity." The parity of a wave function is 1, or even, if the wave function stays the

same if you replace $\vec{r}$ by $-\vec{r}$. The parity is -1, or odd, if the wave function stays the same save for a sign change when you replace $\vec{r}$ by $-\vec{r}$. It turns out that the parity of the spherical harmonics is $(-1)^l$; so it is -1, odd, if the azimuthal quantum number l is odd, and 1, even, if l is even.

To see why, note that replacing $\vec{r}$ by $-\vec{r}$ means in spherical coordinates that θ changes into $\pi - \theta$ and ϕ into $\phi + \pi$. According to trig, the first changes $\cos\theta$ into $-\cos\theta$. That leaves $P_l(\cos\theta)$ unchanged for even l, since P_l is then a symmetric function, but it changes the sign of P_l for odd l. So the sign change is $(-1)^l$. The value of m has no effect, since while the factor $e^{\mathrm{i}m\phi}$ in the spherical harmonics produces a factor $(-1)^{|m|}$ under the change in ϕ, m also puts $|m|$ derivatives on P_l, and each derivative produces a compensating change of sign in $P_l^{|m|}(\cos\theta)$.

D.14.3 Solutions of the Laplace equation

The "Laplace equation" is

$$\nabla^2 u = 0$$

Solutions u to this equation are called "harmonic functions." In spherical coordinates, the Laplace equation has solutions of the form

$$r^l Y_l^m(\theta\phi)$$

This is a complete set of solutions for the Laplace equation inside a sphere. Any solution u of the Laplace equation inside a sphere is a linear combination of these solutions.

As you can see in table 4.3, each solution above is a power series in terms of Cartesian coordinates.

For the Laplace equation outside a sphere, replace r^l by $1/r^{l+1}$ in the solutions above. Note that these solutions are not acceptable inside the sphere because they blow up at the origin.

To check that these are indeed solutions of the Laplace equation, plug them in, using the Laplacian in spherical coordinates given in (N.5). Note here that the angular derivatives can be simplified using the eigenvalue problem of square angular momentum, chapter 4.2.3.

D.14.4 Orthogonal integrals

The spherical harmonics are orthonormal on the unit sphere:

$$\int_{\text{all}} Y_{\underline{l}}^{\underline{m}*} Y_l^m \,\mathrm{d}\Omega = \delta_{l\underline{l}}\delta_{m\underline{m}} \qquad \mathrm{d}\Omega \equiv sin\theta \,\mathrm{d}\theta\mathrm{d}\phi \tag{D.6}$$

Here $\delta_{l\underline{l}}$ is defined to be 0 if $\underline{l}$ and l are different, and 1 if they are equal, and similar for $\delta_{m\underline{m}}$. In other words, the integral above is 1 if $l = \underline{l}$ and $m = \underline{m}$, and

0 in every other case. This expresses physically that the spherical harmonics, as eigenfunctions of the Hermitian z and square angular momentum operators, are orthonormal. Mathematically, it allows you to integrate each spherical harmonic separately and quickly when you are finding $\int |\psi|^2 \, \mathrm{d}^3\vec{r}$ for a wave function ψ expressed in terms of spherical harmonics.

Further

$$\int_{\text{all}} \left(\frac{Y_{\underline{l}}^{\underline{m}*}}{\partial\theta} \frac{Y_l^m}{\partial\theta} + \frac{1}{\sin^2\theta} \frac{Y_{\underline{l}}^{\underline{m}*}}{\partial\phi} \frac{Y_l^m}{\partial\phi} \right) \mathrm{d}\Omega = l(l+1)\delta_{\underline{l}l}\delta_{\underline{m}m} \tag{D.7}$$

This expression simplifies your life when you are finding the $\int |\nabla\psi|^2 \, \mathrm{d}^3\vec{r}$ for a wave function ψ expressed in terms of spherical harmonics.

See the notations for more on spherical coordinates and ∇.

To verify the above expression, integrate the first term in the integral by parts with respect to θ and the second term with respect to ϕ to get

$$-\int \bar{Y} \left(\frac{1}{\sin\theta} (Y \sin\theta)_\theta + \frac{1}{\sin^2\theta} Y_{\phi\phi} \right) \mathrm{d}\Omega$$

and then apply the eigenvalue problem of chapter 4.2.3.

D.14.5 Another way to find the spherical harmonics

There is a more intuitive way to derive the spherical harmonics: they define the power series solutions to the Laplace equation. In particular, each $r^l Y_l^m$ is a different power series solution P of the Laplace equation $\nabla^2 P = 0$ in Cartesian coordinates. Each takes the form

$$\sum_{\alpha+\beta+\gamma=l} c_{\alpha\beta\gamma} x^\alpha y^\beta z^\gamma$$

where the coefficients $c_{\alpha\beta\gamma}$ are such as to make the Laplacian zero.

Even more specifically, the spherical harmonics are of the form

$$\sum_{2a+b=l-m} c_{ab} u^{a+m} v^a z^b \qquad a, b, m \geqslant 0$$

$$\sum_{2a+b=l-|m|} c_{ab} u^a v^{a+|m|} z^b \qquad a, b, -m \geqslant 0$$

where the coordinates $u = x + \mathrm{i}y$ and $v = x - \mathrm{i}y$ serve to simplify the Laplacian. That these are the basic power series solutions of the Laplace equation is readily checked.

To get from those power series solutions back to the equation for the spherical harmonics, one has to do an inverse separation of variables argument for the

solution of the Laplace equation in a sphere in spherical coordinates (compare also the derivation of the hydrogen atom.) Also, one would have to accept on faith that the solution of the Laplace equation is just a power series, as it is in 2D, with no additional nonpower terms, to settle completeness. In other words, you must assume that the solution is analytic.

D.14.6 Still another way to find them

The simplest way of getting the spherical harmonics is probably the one given later in derivation {D.64}.

D.15 The hydrogen radial wave functions

This will be child's play for harmonic oscillator, {D.12}, and spherical harmonics, {D.14}, veterans. If you replace the angular terms in (4.33) by $l(l+1)\hbar^2$, and then divide the entire equation by $\hbar^2$, you get

$$-\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right) + l(l+1) - 2\frac{m_\mathrm{e}e^2}{4\pi\epsilon_0\hbar^2}r = \frac{2m_\mathrm{e}}{\hbar^2}r^2E$$

Since $l(l+1)$ is nondimensional, all terms in this equation must be. In particular, the ratio in the third term must be the reciprocal of a constant with the dimensions of length; so, *define* the constant to be the Bohr radius a_0. It is convenient to also define a correspondingly nondimensionalized radial coordinate as $\rho = r/a_0$. The final term in the equation must be nondimensional too, and that means that the energy E must take the form $(\hbar^2/2m_\mathrm{e}a_0^2)\epsilon$, where ϵ is a nondimensional energy. In terms of these scaled coordinates you get

$$-\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\rho^2\frac{\mathrm{d}R}{\mathrm{d}\rho}\right) + l(l+1) - 2\rho = \rho^2\epsilon$$

or written out

$$-\rho^2R'' - 2\rho R' + [l(l+1) - 2\rho - \epsilon\rho^2]R = 0$$

where the primes denote derivatives with respect to ρ.

Similar to the case of the harmonic oscillator, you must have solutions that become zero at large distances ρ from the nucleus: $\int |\psi|^2\,\mathrm{d}^3\vec{r}$ gives the probability of finding the particle integrated over all possible positions, and if ψ does not become zero sufficiently rapidly at large ρ, this integral would become infinite, rather than one (certainty) as it should. Now the ODE above becomes for large ρ approximately $R'' + \epsilon R = 0$, which has solutions of the rough form $\cos(\sqrt{\epsilon}\rho + \alpha)$ for positive ϵ that do not have the required decay to zero. Zero scaled energy ϵ is still too much, as can be checked by solving in terms of Bessel

functions, so you must have that ϵ is negative. In classical terms, the earth can only hold onto the moon since the moon's total energy is less than the potential energy far from the earth; if it was not, the moon would escape.

Anyway, for bound states, you must have the scaled energy ϵ negative. In that case, the solution at large ρ takes the approximate form $R \approx e^{\pm\sqrt{-\epsilon}\rho}$. Only the negative sign is acceptable. You can make things a lot easier for yourself if you peek at the final solution and rewrite ϵ as being $-1/n^2$ (that is not really cheating, since you are not at this time claiming that n is an integer, just a positive number.) In that case, the acceptable exponential behavior at large distance takes the form $e^{-\frac{1}{2}\xi}$ where $\xi = 2\rho/n$. Split off this exponential part by writing $R = e^{-\frac{1}{2}\xi}\overline{R}$ where $\overline{R}(\xi)$ must remain bounded at large ξ. Substituting these new variables, the ODE becomes

$$-\xi^2\overline{R}'' + \xi(\xi-2)\overline{R}' + [l(l+1) - (n-1)\xi]\overline{R} = 0$$

where the primes indicate derivatives with respect to ξ.

If you do a power series solution of this ODE, you see that it must start with either power ξ^l or with power ξ^{-l-1}. The latter is not acceptable, since it would correspond to an infinite expectation value of energy. You could now expand the solution further in powers of ξ, but the problem is that tabulated polynomials usually do not start with a power l but with power zero or one. So you would not easily recognize the polynomial you get. Therefore it is best to split off the leading power by defining $\overline{R} = \xi^l\overline{\overline{R}}$, which turns the ODE into

$$\xi\overline{\overline{R}}'' + [2(l+1) - \xi]\overline{\overline{R}}' + [n-l-1]\overline{\overline{R}} = 0$$

Substituting in a power series $\overline{\overline{R}} = \sum c_p\xi^p$, you get

$$\sum p[p+2l+1]c_p\xi^{p-1} = \sum [p+l+1-n]c_p\xi^p$$

The acceptable lowest power p of ξ is now zero. Again the series must terminate, otherwise the solution would behave as e^ξ at large distance, which is unacceptable. Termination at a highest power $p = q$ requires that n equals $q + l + 1$. Since q and l are integers, so must be n, and since the final power q is at least zero, n is at least $l + 1$. The correct scaled energy $\epsilon = -1/n^2$ with $n > l$ has been obtained.

With n identified, you can identify the ODE as Laguerre's associated differential equation, e.g. [41, 30.26], the $(2l+1)$-th derivative of Laguerre's differential equation, e.g. [41, 30.1], and the polynomial solutions as the associated Laguerre polynomials L_{n+l}^{2l+1}, e.g. [41, 30.27], the $(2l+1)$-th derivatives of the Laguerre's polynomials L_{n+l}, e.g. [41, 30.2]. To normalize the wave function use an integral from a table book, e.g. [41, 30.46].

Putting it all together, the generic expression for hydrogen eigenfunctions are, drums please:

$$\psi_{nlm} = -\frac{2}{n^2}\sqrt{\frac{(n-l-1)!}{[(n+l)!a_0]^3}}\left(\frac{2\rho}{n}\right)^l L_{n+l}^{2l+1}\left(\frac{2\rho}{n}\right)e^{-\rho/n}Y_l^m(\theta,\phi) \tag{D.8}$$

The properties of the associated Laguerre polynomials $L_{n+l}^{2l+1}(2\rho/n)$ are in table books like [41, pp. 169-172], and the spherical harmonics were given earlier in chapter 4.2.3 and in derivation {D.14}, (D.5).

Do keep in mind that different references have contradictory definitions of the associated Laguerre polynomials. This book follows the notations of [41, pp. 169-172], who define

$$L_n(x) = e^x\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n e^{-x}\right), \quad L_n^m = \frac{\mathrm{d}^m}{\mathrm{d}x^m}L_n(x).$$

In other words, L_n^m is simply the m-th derivative of L_n, which certainly tends to simplify things. According to [25, p. 152], the "most nearly standard" notation defines

$$L_n^m = (-1)^m\frac{\mathrm{d}^m}{\mathrm{d}x^m}L_{n+m}(x).$$

Combine the messy definition of the spherical harmonics (D.5) with the uncertain definition of the Laguerre polynomials in the formulae (D.8) for the hydrogen energy eigenfunctions ψ_{nlm} above, and there is of course always a possibility of getting an eigenfunction wrong if you are not careful.

Sometimes the value of the wave functions at the origin is needed. Now from the above solution (D.8), it is seen that

$$\psi_{nlm} \propto r^l \quad \text{for} \quad r \to 0 \tag{D.9}$$

so only the eigenfunctions ψ_{n00} are nonzero at the origin. To find the value requires $L_n^1(0)$ where L_n^1 is the derivative of the Laguerre polynomial L_n. Skimming through table books, you can find that $L_n(0) = n!$, [41, 30.19], while the differential equation for these function implies that $L_n'(0) = -nL_n(0)$. Therefore:

$$\psi_{n00}(0) = \frac{1}{\sqrt{n^3\pi a_0^3}} \tag{D.10}$$

D.16 Constant spherical potentials derivations

This note gives the derivations for constant potentials in spherical coordinates.

D.16.1 The eigenfunctions

The derivation of the given spherical eigenfunction is almost comically trivial compared to similar problems in quantum mechanics.

Following the lines of the hydrogen atom derivation, chapter 4.3.2, the radial functions R_{El} are found to satisfy the equation

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R_{El}}{\mathrm{d}r}\right)+\left[\frac{p_{\mathrm{c}}^2}{\hbar^2}r^2-l(l+1)\right]R_{El}=0$$

To clean this up a bit more, define new dependent and independent variables. In particular, set $R_{El}=f_l$ and $r=x\hbar/p_{\mathrm{c}}$. That produces the spherical Bessel equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x^2\frac{\mathrm{d}f_l}{\mathrm{d}x}\right)+\left[x^2-l(l+1)\right]f_l=0$$

It is now to be shown that the solutions f_l to this equation are the Hankel and Bessel functions as given earlier.

To do so, make another change of dependent variable by setting $f_l=x^l g_l$. That gives for the g_l:

$$x\frac{\mathrm{d}^2g_l}{\mathrm{d}x^2}+2(l+1)\frac{\mathrm{d}g_l}{\mathrm{d}x}+xg_l=0$$

Check, by simply plugging it in, that $e^{\mathrm{i}x}/x$ is a solution for $l=0$.

Now make a further change in independent variable from x to $\xi=\frac{1}{2}x^2$ to give

$$2\xi\frac{\mathrm{d}^2g_l}{\mathrm{d}\xi^2}+2(l+1)\frac{\mathrm{d}g_l}{\mathrm{d}\xi}+g_l=0$$

Note that the equation for $l=1$ is obtained by differentiating the one for $l=0$, (taking g_l' as the new unknown.). That implies that the ξ-derivative of the solution for $l=0$ above is a solution for $l=1$. Keep differentiating to get solutions for all values of l. That produces the spherical Hankel functions of the first kind; the remaining constant is just an arbitrarily chosen normalization factor.

Since the original differential equation is real, the real and imaginary parts of these Hankel functions, as well as their complex conjugates, must be solutions too. That gives the spherical Bessel functions and Hankel functions of the second kind, respectively.

Note that all of them are just *finite* sums of elementary functions. And that physicists do not even disagree over their definition, just their names.

D.16.2 The Rayleigh formula

To derive the Rayleigh formula, convert the linear momentum eigenfunction to spherical coordinates by setting $z=r\cos\theta$. Also, for brevity set $x=p_\infty r/\hbar$.

That turns the linear momentum eigenfunction into

$$e^{\mathrm{i}x\cos\theta} = \sum_{\underline{l}=0}^{\infty} \frac{(\mathrm{i}x\cos\theta)^{\underline{l}}}{\underline{l}!}$$

the latter from Taylor series expansion of the exponential.

Now this is an energy eigenfunction. It can be written in terms of the spherical eigenfunctions

$$\psi_{Elm} = j_l(x) Y_l^m(\theta,\phi)$$

with the same energy because the ψ_{Elm} are complete. In addition, the only eigenfunctions needed are those with $m = 0$. The reason is that the spherical harmonics Y_l^m are simply Fourier modes in the ϕ direction, {D.14} (D.5), and the linear momentum eigenfunction above does not depend on ϕ. Therefore

$$\sum_{\underline{l}=0}^{\infty} \frac{(\mathrm{i}x\cos\theta)^{\underline{l}}}{\underline{l}!} = \sum_{l=0}^{\infty} c_{\mathrm{w},l} j_l(x) Y_l^0(\theta)$$

for suitable coefficients $c_{\mathrm{w},l}$.

To find these coefficients, find the lowest power of x in j_l by writing the sine in (A.19) as a Taylor series and then switching to x^2 as independent variable. Similarly, find the highest power of $\cos\theta$ in Y_l^0, {D.14} (D.5), by looking up the Rodrigue's formula for the Legendre polynomial appearing in it. That gives

$$\sum_{\underline{l}=0}^{\infty} \frac{(\mathrm{i}x\cos\theta)^{\underline{l}}}{\underline{l}!} = \sum_{l=0}^{\infty} c_{\mathrm{w},l} \left(\frac{2^l l!}{(2l+1)!} x^l + \ldots \right) \sqrt{\frac{2l+1}{4\pi}} \left(\frac{(2l)!}{2^l (l!)^2} \cos^l\theta + \ldots \right)$$

Each coefficient $c_{\mathrm{w},l}$ must be chosen to match the term with $\underline{l} = l$ in the first sum, because the terms for the other values for l do not have a low enough power of x or a high enough power of the cosine. That gives the Rayleigh values of the coefficients as listed earlier.

D.17 Inner product for the expectation value

To see that $\langle\Psi|A\rangle$ works for getting the expectation value, just write Ψ out in terms of the eigenfunctions α_n of A:

$$\langle c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \ldots | A | c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \ldots \rangle$$

Now by the definition of eigenfunctions $A\alpha_n = a_n\alpha_n$ for every n, so you get

$$\langle c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \ldots | c_1a_1\alpha_1 + c_2a_2\alpha_2 + c_3a_3\alpha_3 + \ldots \rangle$$

Since eigenfunctions are orthonormal:

$$\langle\alpha_1|\alpha_1\rangle = 1 \quad \langle\alpha_2|\alpha_2\rangle = 1 \quad \langle\alpha_3|\alpha_3\rangle = 1 \quad \ldots$$

$$\langle\alpha_1|\alpha_2\rangle = \langle\alpha_2|\alpha_1\rangle = \langle\alpha_1|\alpha_3\rangle = \langle\alpha_3|\alpha_1\rangle = \langle\alpha_2|\alpha_3\rangle = \langle\alpha_3|\alpha_2\rangle = \ldots = 0$$

So, multiplying out produces the desired result:

$$\langle\Psi|A\Psi\rangle = |c_1|^2 a_1 + |c_2|^2 a_2 + |c_3|^2 a_3 + \ldots \equiv \langle A\rangle$$

D.18 Eigenfunctions of commuting operators

Any two operators A and B that commute, $AB = BA$, have a common set of eigenfunctions, provided only that each has a complete set of eigenfunctions. (In other words, the operators do not necessarily have to be Hermitian. Unitary, anti-Hermitian, etcetera, operators all qualify.)

First note the following:

> *if α_i is an eigenfunction of A with eigenvalue a_i, then $B\alpha_i$ is either also an eigenfunction of A with eigenvalue a_i or is zero.*

To see that, note that since A and B commute $AB\alpha_i = BA\alpha_i$ which is $a_i B\alpha_i$. Comparing start and end, the combination $B\alpha_i$ must be an eigenfunction of A with eigenvalue a_i if it is not zero. (Eigenfunctions may not be zero.)

Now assume that there is just a single independent eigenfunction α_i for each distinct eigenvalue a_i of A. Then if $B\alpha_i$ is nonzero, it can only be a multiple of that single eigenfunction. By definition, that makes α_i an eigenfunction of B too, with as eigenvalue the multiple. On the other hand, if $B\alpha_i$ is zero, then α_i is still an eigenfunction of B, now with eigenvalue zero. So under the stated assumption, A and B have the exact same eigenfunctions, proving the assertion of this derivation.

However, frequently there is "degeneracy," i.e. there is more than one eigenfunction $\alpha_{i,1}, \alpha_{i,2}, \ldots$ for a single eigenvalue a_i. Then the fact that, say, $B\alpha_{i,1}$ is an eigenfunction of A with eigenvalue a_i no longer means that $B\alpha_{i,1}$ is a multiple of $\alpha_{i,1}$; it only means that $B\alpha_{i,1}$ is some combination of all of $\alpha_{i,1}, \alpha_{i,2}, \ldots$. Which means that $\alpha_{i,1}$ is not in general an eigenfunction of B.

To deal with that, it has to be assumed that the problem has been numerically approximated by some finite-dimensional one. Then A and B will be matrices, and the number of independent eigenfunctions (or rather, eigenvectors now) of A and B will be finite and equal. That allows the problem to be addressed one eigenfunction at a time.

Assume now that β is an eigenfunction of B, with eigenvalue b, that is not yet an eigenfunction of A too. By completeness, it can still be written as a combination of the eigenfunctions of A, and more particularly as $\beta = \beta_{a_i} + \beta_o$ where β_{a_i} is a combination of the eigenfunctions of A with eigenvalue a_i and β_o a combination of the eigenfunctions of A with other eigenvalues. There must be such eigenfunctions with β_{a_i} nonzero, because without using the α_i you cannot

create an equal number of independent eigenfunctions of B as of A. By definition

$$B\,(\beta_{a_i} + \beta_o) = b\,(\beta_{a_i} + \beta_o)$$

but that must mean that

$$B\beta_{a_i} = b\beta_{a_i}$$

since if it is not, $B\beta_0$ cannot make up the difference; as seen earlier, $B\beta_0$ only consists of eigenfunctions of A that do *not* have eigenvalue a_i. According to the above equation, β_{a_i}, which is already an eigenfunction of A with eigenvalue a_i, is also an eigenfunction of B with eigenvalue b. So replace one of the $\alpha_{i,1}$, $\alpha_{i,2}$,... with β_{a_i}. (If you write β_{a_i} in terms of the $\alpha_{i,1}$, $\alpha_{i,2}$,..., then the function you replace may not appear with a zero coefficient.) Similarly replace an eigenfunction of B with eigenvalue b with β_{a_i}. Then A and B have one more common eigenfunction. Keep going in this way and eventually all eigenfunctions of B are also eigenfunctions of A and vice versa.

Similar arguments can be used recursively to show that more generally, a set of operators $A, B, C, \ldots$ that all commute have a single common set of eigenfunctions. The trick is to define an artificial new operator, call it P, that has the common eigenfunctions of A and B, but whose eigenvalues are distinct for any two eigenfunctions unless these eigenfunctions have the same eigenvalues for both A and B. Then the eigenfunctions of P, even if you mess with them, remain eigenfunctions of A and B. So go find common eigenfunctions for P and C.

The above derivation assumed that the problem was finite-dimensional, or discretized some way into a finite-dimensional one like you do in numerical solutions. The latter is open to some suspicion, because even the most accurate numerical approximation is never truly exact. Unfortunately, in the infinite-dimensional case the derivation gets much trickier. However, as the hydrogen atom and harmonic oscillator eigenfunction examples indicate, typical infinite systems in nature do obey the theorem anyway.

D.19 The generalized uncertainty relationship

This note derives the generalized uncertainty relationship.

For brevity, define $A' = A - \langle A \rangle$ and $B' = B - \langle B \rangle$, then the general expression for standard deviation says

$$\sigma_A^2 \sigma_B^2 = \langle A'^2 \rangle \langle B'^2 \rangle = \langle \Psi | A'^2 \Psi \rangle \langle \Psi | B'^2 \Psi \rangle$$

Hermitian operators can be taken to the other side of inner products, so

$$\sigma_A^2 \sigma_B^2 = \langle A'\Psi | A'\Psi \rangle \langle B'\Psi | B'\Psi \rangle$$

Now the Cauchy-Schwartz inequality says that for any f and g,

$$|\langle f|g\rangle| \leqslant \sqrt{\langle f|f\rangle}\sqrt{\langle g|g\rangle}$$

(See the notations for more on this theorem.) Using the Cauchy-Schwartz inequality in reversed order, you get

$$\sigma_A^2 \sigma_B^2 \geqslant |\langle A'\Psi|B'\Psi\rangle|^2 = |\langle A'B'\rangle|^2$$

Now by the definition of the inner product, the complex conjugate of $\langle A'\Psi|B'\Psi\rangle$ is $\langle B'\Psi|A'\Psi\rangle$, so the complex conjugate of $\langle A'B'\rangle$ is $\langle B'A'\rangle$, and averaging a complex number with minus its complex conjugate reduces its size, since the real part averages away, so

$$\sigma_A^2 \sigma_B^2 \geqslant \left| \frac{\langle A'B'\rangle - \langle B'A'\rangle}{2} \right|^2$$

The quantity in the top is the expectation value of the commutator $[A', B']$. Writing it out shows that $[A', B'] = [A, B]$.

D.20 Derivation of the commutator rules

This note explains where the formulae of chapter 4.5.4 come from.

The general assertions are readily checked by simply writing out both sides of the equation and comparing. And some are just rewrites of earlier ones.

Position and potential energy operators commute since they are just ordinary numerical multiplications, and these commute.

The linear momentum operators commute because the order in which differentiation is done is irrelevant. Similarly, commutators between angular momentum in one direction and position in another direction commute since the other directions are not affected by the differentiation.

The commutator between the position X and linear momentum p_x in the x-direction was worked out in the previous subsection to figure out Heisenberg's uncertainty principle. Of course, three-dimensional space has no preferred direction, so the result applies the same in any direction, including the y and z directions.

The angular momentum commutators are simplest obtained by just grinding out

$$[\widehat{L}_x, \widehat{L}_y] = [\widehat{y}\widehat{p}_z - \widehat{z}\widehat{p}_y, \widehat{z}\widehat{p}_x - \widehat{x}\widehat{p}_z]$$

using the linear combination and product manipulation rules and the commutators for linear angular momentum. To generalize the result you get, you cannot just arbitrarily swap x, y, and z, since, as every mechanic knows, a right-handed screw is not the same as a left-handed one, and some axes swaps would turn one

into the other. But you can swap axes according to the "$xyzxyzx\ldots$" "cyclic permutation" scheme, as in:

$$x \to y, \quad y \to z, \quad z \to x$$

which produces the other two commutators if you do it twice:

$$[\widehat{L}_x, \widehat{L}_y] = \mathrm{i}\hbar\widehat{L}_z \quad \longrightarrow \quad [\widehat{L}_y, \widehat{L}_z] = \mathrm{i}\hbar\widehat{L}_x \quad \longrightarrow \quad [\widehat{L}_z, \widehat{L}_x] = \mathrm{i}\hbar\widehat{L}_y$$

For the commutators with square angular momentum, work out

$$[\widehat{L}_x, \widehat{L}_x^2 + \widehat{L}_y^2 + \widehat{L}_z^2]$$

using the manipulation rules and the commutators between angular momentum components.

A commutator like $[\widehat{x}, \widehat{L}_x] = [\widehat{x}, \widehat{y}\widehat{p}_z - \widehat{z}\widehat{p}_y]$ is zero because everything commutes in it. However, in a commutator like $[\widehat{x}, \widehat{L}_y] = [\widehat{x}, \widehat{z}\widehat{p}_x - \widehat{x}\widehat{p}_z]$, $\widehat{x}$ does not commute with $\widehat{p}_x$, so multiplying out and taking the $\widehat{z}$ out of $[\widehat{x}, \widehat{z}\widehat{p}_x]$ at its own side, you get $\widehat{z}[\widehat{x}, \widehat{p}_x]$, and the commutator left is the canonical one, which has value $\mathrm{i}\hbar$. Plug these results and similar into $[\widehat{x}^2 + \widehat{y}^2 + \widehat{z}^2, L_x]$ and you get zero.

For a commutator like $[\widehat{x}, \widehat{L}^2] = [\widehat{x}, \widehat{L}_x^2 + \widehat{L}_y^2 + \widehat{L}_z^2]$, the L_x^2 term produces zero because $\widehat{L}_x$ commutes with $\widehat{x}$, and in the remaining term, taking the various factors out at their own sides of the commutator produces

$$\begin{aligned}
[\widehat{x}, \widehat{L}^2] &= \widehat{L}_y[\widehat{x}, \widehat{L}_y] + [\widehat{x}, \widehat{L}_y]\widehat{L}_y + \widehat{L}_z[\widehat{x}, \widehat{L}_z] + [\widehat{x}, \widehat{L}_z]\widehat{L}_z \\
&= \mathrm{i}\hbar\widehat{L}_y\widehat{z} + \mathrm{i}\hbar\widehat{z}\widehat{L}_y - \mathrm{i}\hbar\widehat{L}_z\widehat{y} - \mathrm{i}\hbar\widehat{y}\widehat{L}_z
\end{aligned}$$

the final equality because of the commutators already worked out. Now by the nature of the commutator, you can swap the order of the terms in $\widehat{L}_y\widehat{z}$ as long as you add the commutator $[\widehat{L}_y, \widehat{z}]$ to make up for it, and that commutator was already found to be $\mathrm{i}\hbar\widehat{x}$, The same way the order of $\widehat{L}_z\widehat{y}$ can be swapped to give

$$[\widehat{x}, \widehat{L}^2] = -2\hbar^2\widehat{x} - 2\mathrm{i}\hbar(\widehat{y}\widehat{L}_z - \widehat{z}\widehat{L}_y)$$

and the parenthetical expression can be recognized as the x-component of $\widehat{\vec{r}} \times \widehat{\vec{L}}$, giving one of the expressions claimed.

Instead you can work out the parenthetical expression further by substituting in the definitions for $\widehat{L}_z$ and $\widehat{L}_y$:

$$[\widehat{x}, \widehat{L}^2] = -2\hbar^2\widehat{x} - 2\mathrm{i}\hbar\Big(\widehat{y}(\widehat{x}\widehat{p}_y - \widehat{y}\widehat{p}_x) - \widehat{z}(\widehat{z}\widehat{p}_x - \widehat{x}\widehat{p}_z) - \widehat{x}(\widehat{x}\widehat{p}_x - \widehat{x}\widehat{p}_x)\Big)$$

where the third term added within the big parentheses is self-evidently zero. This can be reordered to the x-component of the second claimed expression. And as always, the other components are of course no different.

The commutators between linear and angular momentum go almost identically, except for additional swaps in the order between position and momentum operators using the canonical commutator.

To derive the first commutator in (4.73), consider the z-component as the example:

$$[x\widehat{L}_y - y\widehat{L}_x, \widehat{L}^2] = [x, \widehat{L}^2]\widehat{L}_y - [y, \widehat{L}^2]L_x$$

because L^2 commutes with $\widehat{\vec{L}}$, and using (4.68)

$$[x\widehat{L}_y - y\widehat{L}_x, \widehat{L}^2] = -2\hbar^2 x\widehat{L}_y - 2\mathrm{i}\hbar(y\widehat{L}_z\widehat{L}_y - z\widehat{L}_y^2) + 2\hbar^2 y\widehat{L}_x + 2\mathrm{i}\hbar(z\widehat{L}_x^2 - x\widehat{L}_z\widehat{L}_x)$$

Now use the commutator $[\widehat{L}_y, \widehat{L}_z]$ to get rid of $\widehat{L}_z\widehat{L}_y$ and $[\widehat{L}_z, \widehat{L}_x]$ to get rid of $\widehat{L}_z\widehat{L}_x$ and clean up to get

$$[x\widehat{L}_y - y\widehat{L}_x, \widehat{L}^2] = 2\mathrm{i}\hbar\left(-y\widehat{L}_y\widehat{L}_z + z\widehat{L}_y^2 + z\widehat{L}_x^2 - x\widehat{L}_x\widehat{L}_z\right)$$

Now $\vec{r}\cdot\widehat{\vec{L}} = \vec{r}\cdot(\vec{r}\times\widehat{\vec{p}}) = 0$ so $x\widehat{L}_x + y\widehat{L}_y = -z\widehat{L}_z$, which gives the claimed expression. To verify the second equation of (4.73), use (4.68), the first of (4.73), and the definition of $[\vec{r}, \widehat{L}^2]$.

D.21 Solution of the hydrogen molecular ion

The key to the variational approximation to the hydrogen molecular ion is to be able to accurately evaluate the expectation energy

$$\langle E\rangle = \langle a\psi_\mathrm{l} + b\psi_\mathrm{r}|H|a\psi_\mathrm{l} + b\psi_\mathrm{r}\rangle$$

This can be multiplied out and simplified by noting that ψ_l and ψ_r are eigenfunctions of the partial Hamiltonians. For example,

$$H\psi_\mathrm{l} = E_1\psi_\mathrm{l} - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_\mathrm{r}}\psi_\mathrm{l}$$

where E_1 is the -13.6 eV hydrogen atom ground state energy. The expression can be further simplified by noting that by symmetry

$$\langle\psi_\mathrm{r}|r_\mathrm{l}^{-1}\psi_\mathrm{r}\rangle = \langle\psi_\mathrm{l}|r_\mathrm{r}^{-1}\psi_\mathrm{l}\rangle \qquad \langle\psi_\mathrm{l}|r_\mathrm{l}^{-1}\psi_\mathrm{r}\rangle = \langle\psi_\mathrm{r}|r_\mathrm{r}^{-1}\psi_\mathrm{l}\rangle$$

and that ψ_l and ψ_r are real, so that the left and right sides of the various inner products can be reversed. Also, a and b are related by the normalization requirement

$$a^2 + b^2 + 2ab\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle = 1$$

Cleaning up the expectation energy in this way, the result is

$$\langle E\rangle \;=\; E_1 - \frac{e^2}{4\pi\epsilon_0}\left[\left\langle\psi_{\rm l}|r_{\rm r}^{-1}\psi_{\rm l}\right\rangle - \frac{1}{d} + 2ab\,\langle\psi_{\rm l}|\psi_{\rm r}\rangle\left\{\frac{\left\langle\psi_{\rm l}|r_{\rm l}^{-1}\psi_{\rm r}\right\rangle}{\langle\psi_{\rm l}|\psi_{\rm r}\rangle} - \left\langle\psi_{\rm l}|r_{\rm r}^{-1}\psi_{\rm l}\right\rangle\right\}\right]$$

which includes the proton to proton repulsion energy (the $1/d$). The energy E_1 is the -13.6 eV amount of energy when the protons are far apart.

Numerical integration is not needed; the inner product integrals in this expression can be done analytically. To do so, take the origin of a spherical coordinate system (r,θ,ϕ) at the left proton, and the axis towards the right one, so that

$$r_{\rm l} = |\vec{r}-\vec{r}_{\rm lp}| = r \qquad r_{\rm r} = |\vec{r}-\vec{r}_{\rm rp}| = \sqrt{d^2+r^2-2dr\cos(\theta)}.$$

In those terms,

$$\psi_{\rm l} = \frac{1}{\sqrt{\pi a_0^3}}e^{-r/a_0} \qquad \psi_{\rm r} = \frac{1}{\sqrt{\pi a_0^3}}e^{-\sqrt{d^2+r^2-2dr\cos(\theta)}/a_0}.$$

Then integrate angles first using $\mathrm{d}^3\vec{r} = r^2\sin(\theta)\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi = -r^2\mathrm{d}r\,\mathrm{d}\cos(\theta)\,\mathrm{d}\phi$. Do not forget that $\sqrt{x^2} = |x|$, not x, e.g. $\sqrt{(-3)^2} = 3$, not -3. More details are in [25, pp. 305-307].

The "overlap integral" turns out to be

$$\langle\psi_{\rm l}|\psi_{\rm r}\rangle = e^{-d/a_0}\left[1+\frac{d}{a_0}+\frac{1}{3}\left(\frac{d}{a_0}\right)^2\right]$$

and provides a measure of how much the regions of the two wave functions overlap. The "direct integral" is

$$\left\langle\psi_{\rm l}|r_{\rm r}^{-1}\psi_{\rm l}\right\rangle = \frac{1}{d} - \left[\frac{1}{a_0}+\frac{1}{d}\right]e^{-2d/a_0}$$

and gives the classical potential of an electron density of strength $|\psi_{\rm l}|^2$ in the field of the right proton, except for the factor $-e^2/4\pi\epsilon_0$. The "exchange integral" is

$$\left\langle\psi_{\rm l}|r_{\rm l}^{-1}\psi_{\rm r}\right\rangle = \left[\frac{1}{a_0}+\frac{d}{a_0^2}\right]e^{-d/a_0}.$$

and is somewhat of a twilight term, since $\psi_{\rm l}$ suggests that the electron is around the left proton, but $\psi_{\rm r}$ suggests it is around the right one.

D.22 Unique ground state wave function

This derivation completes {A.8}. In particular, it proves that ground states are unique, given a real, noninfinite, potential that only depends on position. The derivation also proves that ground states cannot become zero. So they can be taken to be positive.

The basic idea is first to assume tentatively that there would be two independent ground state wave functions. These could then be taken to be orthonormal as usual. That means that the inner product of the two wave functions would be zero. However, it can be shown, see below, that ground state wave functions cannot cross zero. That means that both wave functions can be taken to be everywhere positive. (The one exception is at impenetrable boundaries, where the wave function is forced to be zero, rather than positive. But that exception does not change the argument here.) Now if you check the definition of the inner product, you see that the inner product of two positive wave functions is positive, not zero. But the orthonormality says that it is zero. So there is a contradiction. That means that the made assumption, that there are two independent ground states, must be wrong. So the ground state must be unique.

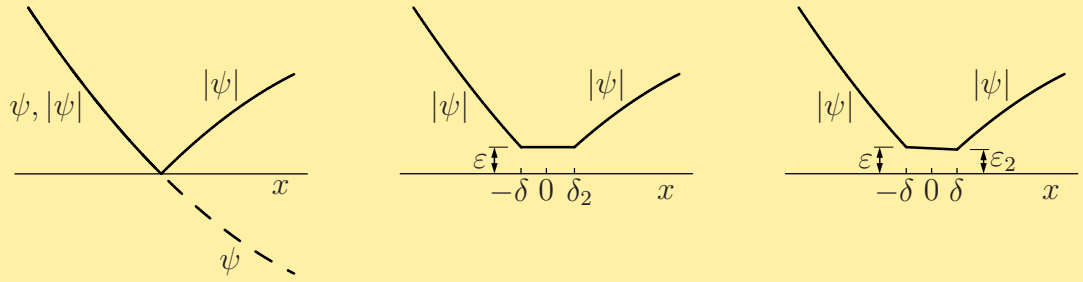


Figure D.1: Right: the absolute value of the wave function has a kink at a zero crossing. Middle: the kink has been slightly blunted. Right: an alternate way of blunting.

To finish the proof then, it must still be shown that ground states that cross zero are not possible. Tentatively suppose that you had a ground state whose wave function did cross zero. Near a zero crossing, the wave function ψ will then look something like the left part of figure D.1. (For a multi-dimensional wave function, you can take this figure to be some arbitrary one-dimensional cross section through the zero crossing.) Note from the figure that the absolute value $|\psi|$ has a kink at the zero crossing.

Next recall from {A.8} that $|\psi|$ has the ground state energy just like ψ. So it should not be possible to lower the energy below that of $|\psi|$. But the problem is now that the wave function shown in the middle in figure D.1 *does* have less energy. This wave function looks generally the same as $|\psi|$, except that the kink

has been blunted just a little bit. More precisely, this wave function has been prevented from becoming any smaller than some very small number ε.

To see why this wave function has less energy, compare what happens to the kinetic and potential energies. The energy is the sum of a kinetic energy integral I_T and a potential energy integral I_V. These are given by

$$I_T = \frac{\hbar^2}{2m} \int_{\text{all}} (\nabla\psi)^2 \, \mathrm{d}^3\vec{r} \qquad I_V = \int_{\text{all}} V\psi^2 \, \mathrm{d}^3\vec{r}$$

In the region that is blunted, the integrand of the kinetic energy integral is now zero, instead of whatever positive value it had in the left figure. The constant ε has zero derivatives. So the kinetic energy has been decreased noticeably.

You might at first think that the potential energy can compensate by increasing more than the kinetic energy decreases. But that does not work, because the integrand of the potential energy integral is proportional to $|\psi|^2$, and that is negligibly small in the blunted region. In fact, $|\psi|^2$ is no larger than ε^2, and ε was a very small number. So, if the kinetic energy decreases and the potential energy stays virtually the same, the conclusion is unavoidable. The energy decreases. The wave function in the middle in the figure has less energy than the ones on the left. So the ones on the left cannot be ground states. So ground state wave functions cannot cross zero.

That is basically it. Unfortunately, there are a few loose ends in the reasoning above. That is even if you ignore that "small" is not a valid mathematical concept. A number is either zero or not zero; that is all in mathematics. The correct mathematical statement is: "there is a number ε small enough that the kinetic energy decrease exceeds the potential energy increase." (Note that "small enough" does not imply small. All positive numbers less than 1 000 are small enough to be less than 1 000.) But that is so picky.

More importantly, you might object that after blunting, the wave function will no longer be normalized. But you can correct for that by dividing the given expression of the expectation energy by the square norm of the wave function. In particular, using a prime to indicate a quantity after blunting the wave function, the correct energy is

$$\langle E \rangle' = \frac{I'_T + I'_V}{\langle \psi' | \psi' \rangle} \qquad \langle \psi' | \psi' \rangle = \int_{\text{all}} (\psi')^2 \, \mathrm{d}^3\vec{r}$$

Now note that the square norm changes negligibly under the smoothing, because its integrand involves $|\psi|^2$ just like the potential energy. So dividing by the square norm should not make a difference.

In fact, there is a trick to avoid the normalization problem completely. Simply redefine the potential energy by a constant to make the expectation energy zero. You can always do that; changing the definition of the potential energy by constant does not make a difference physically. And if the expectation energy

$\langle E \rangle$ is zero, then so is $I_{\rm T} + I_{\rm V}$. Therefore the change in energy due to blunting becomes

$$\langle E \rangle' - \langle E \rangle = \langle E \rangle' = \frac{I'_T + I'_V}{\langle \psi' | \psi' \rangle} = \frac{I'_T + I'_V - (I_T + I_V)}{\langle \psi' | \psi' \rangle}$$

Comparing start and end, you see that the sign of the change in energy is the same as the sign of the change in the kinetic and potential energy integrals. Regardless of whether ψ' is normalized. And the sign of the change in energy is all that counts. If it is negative, you do not have a ground state. So if $I_T + I_V$ decreases due to blunting, you do not have a ground state. Because of this trick, the normalization problem can be ignored in the rest of the derivations.

You might further object that the given arguments do not account for the possibility that the wave function could cross zero with zero slope. In that case, the integrand of the original kinetic energy would be vanishingly small too. True.

But in one dimension, you can use the Cauchy-Schwartz inequality of the notations section on $|\psi|$ to show that the decrease in kinetic energy will still be more than the increase in potential energy. For simplicity, the coordinate x will be taken zero at the original zero crossing, as in the middle graph of figure D.1. Now consider the part of the blunted region at negative x. Here the original kinetic energy integral was:

$$\begin{aligned} I_T = \frac{\hbar^2}{2m} \int_{-\delta}^{0} |\psi_x|^2 \,{\rm d}x &= \frac{\hbar^2}{2m} \int_{-\delta}^{0} |\psi_x|^2 \,{\rm d}x \quad \frac{1}{\delta} \int_{-\delta}^{0} 1 \,{\rm d}x \\ &\geqslant \frac{\hbar^2}{2m\delta} \left(\int_{-\delta}^{0} |\psi_x| 1 \,{\rm d}x \right)^2 \\ &\geqslant \frac{\hbar^2}{2m\delta} \left(\int_{-\delta}^{0} -\psi_x \,{\rm d}x \right)^2 \\ &= \frac{\hbar^2 \varepsilon^2}{2m\delta} \end{aligned}$$

The first inequality above is the Cauchy-Schwartz inequality. The final equality applies because the change in ψ is the integral of its derivative. Comparing start and end above, the kinetic energy decrease is at least $\hbar^2\varepsilon^2/2m\delta$. On the other hand for the increase in potential energy

$$I'_V - I_V = \int_{-\delta}^{0} V(\varepsilon^2 - |\psi|^2) \,{\rm d}x \leqslant \int_{-\delta}^{0} V_{\rm max}\, \varepsilon^2 \,{\rm d}x = V_{\rm max}\, \varepsilon^2 \delta$$

where $V_{\rm max}$ is the maximum (redefined) potential in the region. (Note that if ψ is not monotonuous, $-\delta$ and δ_2 are defined as the points closest to the origin where ε is reached. So by definition $|\psi|$ does not exceed ε.) It is seen that the maximum potential energy decrease is proportional to δ. However, the

minimum kinetic energy decrease is proportional to $1/\delta$. So for small enough δ, potential energy increase cannot compete with kinetic energy decrease. More specifically, taking δ^2 small compared to $\hbar^2/2mV_{\max}$ makes the potential energy increase small compared to the kinetic energy decrease. So the net energy will decrease, showing that the original wave function is indeed not a ground state. (If $V_{\max}$ is negative, the potential energy will decrease, and net energy decrease is automatic.)

The same arguments normally apply for the blunted region at positive x. However, there is a possible exception. If after ψ reaches zero, it stays zero, there will be no position δ_2. At least not one vanishingly close to zero. To deal with this possibility, a slightly more sophisticated blunting can be used. That one is shown to the right in figure D.1. Here the blunting region is taken to be symmetric around the origin. The value of δ is taken as the smallest distance from the origin where ε is reached. Therefore once again $|\psi|$ does not exceed ε. Note that the modified wave function now has some kinetic energy left. In particular it has left

$$\frac{\hbar^2}{2m}\int_{-\delta}^{\delta}\left|\frac{\varepsilon-\varepsilon_2}{2\delta}\right|^2 \mathrm{d}x = \frac{\hbar^2(\varepsilon-\varepsilon_2)^2}{4m\delta} \leqslant \frac{\hbar^2\varepsilon^2}{4m\delta}$$

However, as seen above, the negative blunted part has kinetic energy of at least $\hbar^2\varepsilon^2/2m\delta$. So the kinetic energy decrease is still at least half of what it was. That is enough not to change the basic story.

Note that in neither approach, the zero crossing point can be at an impenetrable boundary. Neither blunted wave function is zero at $x = 0$ at it should be at an impenetrable boundary. That explains why ground state wave functions can indeed become zero at impenetrable boundaries. The ground state of the particle in a pipe provides an example, chapter 3.5.

Also note the need to assume that the potential does not become positive infinity. If the potential is positive infinity in a finite region, then the wave function *is* in fact zero inside that region. The particle cannot penetrate into such a region. Its surface acts just like an impenetrable boundary.

How about wave functions in more than one dimension? That is easy, if you will allow a very minor assumption. The minor assumption is that there is at least a single crossing point where the gradient of ψ is continuous and nonzero. It does not have to be true at all the zero crossing points, just at one of them. And in fact it does not even have to be true for either one of the two supposed ground states. It is enough if it is true for a single point in some linear combination of them. So it is very hard to imagine ground states for which the assumption would not apply.

Accepting that assumption, things are straightforward. Take the blunted wave function essentially like the middle graph in figure D.1. The x-direction is now the direction of the gradient at the point. However, rather than limiting the wave function to stay above ε, demand that it stays above $\varepsilon(\ell^2 - r^2)/\ell^2$.

Here r is the distance from the considered zero crossing point, and ℓ is a number small enough that the variation in the gradient is no more than say 50% within a distance ℓ from the zero crossing point. There is then again some kinetic energy left, but it is negligibly small. The estimates in each cross section of the blunted region are essentially the same as in the one-dimensional case.

However, all that does require that one minor assumption. You might wonder about pathological cases. For example, what if one wave function is only nonzero where the other is zero and vice-versa? With zero gradient at every single point of the zero crossings to boot? Of course, you and I know that ground states are not just stupidly going to be zero in sizable parts of the region. Why would the electron stay out of some region completely? Would not its uncertainty in position at least produce a very tiny probability for the elecron to be inside them? But proving it rigorously is another matter. Then there are somewhat more reasonable conjectures, like that a wave function would become zero at a single point not at the boundary. (That would still give a unique ground state. But would you not want to know whether it really could happen?) How about fractal wave functions? Or just discontinuous ones? In one dimension the wave function must be continuous, period. A discontinuity would produce a delta function in the derivative, which would produce infinite kinetic energy. But in multiple dimensions, things become much less obvious. (Note however that in real life, a noticeably singularity in ψ at a point would require quite a singular potential at that point.)

You might guess that you could use the approach of the right graph in figure D.1 in multiple dimensions, taking the x coordinate in the direction normal to the zero crossing surface. But first of all that requires that the zero crossing surface is rectifiable. That excludes lone zero crossing points, or fractal crossing surfaces. And in addition there is a major problem with trying to show that the derivatives in directions other than x remain small.

There is however a method somewhat similar to the one of the right graph that continues to work in more than one dimensions. In particular, in three dimensions this method uses a small sphere of radius δ around the supposed point of zero wave function. The method can show in, any number of dimensions, that $|\psi|$ cannot become zero. (Except at impenetrable boundaries as always.) The method does not make any unjustified a priory assumptions like a nonzero gradient. However, be warned: it is going to be messy. Only mathematically inclined students should read the rest of this derivation.

The discussion will use three dimensions as an example. That corresponds, for example, to the electron of the hydrogen molecular ion. But the same arguments can be made in any number of dimensions. For example, you might have a particle confined in a two-dimensional quantum well. In that case, the sphere around the point of zero wave function becomes a circle. Similarly, in a one-dimensional quantum wire, the sphere becomes the line segment $-\delta \leqslant x \leqslant \delta$. If you have two nonconfined electrons instead of just one, you are in six

dimensions. All these cases can be covered mathematically by generalizing the three-dimensional sphere to a "hypersphere." A two-dimensional hypersphere is physically a circle, and a one-dimensional hypersphere is a line segment. As discussed in the notations section, a general n-dimensional hypersphere has an n-dimensional "volume" and surface "area" given by:

$$\mathcal{V}_n = C_n \delta^n \qquad A_n = n C_n \delta^{n-1}$$

For example, $C_3 = 4\pi/3$, so the above expressions give the correct volume and surface of a sphere in three dimensions. In two dimensions, the "volume" is physically the area of the circle, and the "area" is its perimeter. The derivations will need that the n-dimensional infinitesimal integration element is

$$\mathrm{d}^n\vec{r} = \mathrm{d}A_n \mathrm{d}r$$

Here $\mathrm{d}A_n$ is an infinitesimal segment of the spherical surface of radius r. You can relate this to the way that you do integration in polar or spherical coordinates. However, the above expression does not depend on exactly how the angular coordinates on hypersphere areas are defined.

To show that points of zero wave function are not possible, once again first the opposite will be assumed. So it will be assumed that there is some point where $|\psi|$ becomes zero. Then a contradiction will be established. That means that the assumption must be incorrect; there are no points of zero wave function.

To find the contradiction, define a radial coordinate r as the distance away from the supposed point of zero $|\psi|$. Next at every distance r, define $\varphi(r)$ as the average value $|\psi|$ on the spherical surface of radius r:

$$\varphi(r) \equiv \int |\psi| \frac{\mathrm{d}A_n}{A_n}$$

Function $\varphi(r)$ will need to be continuous for $r \neq 0$, otherwise the implied jump in wave function values would produce infinite kinetic energy. For $|\psi|$ to become zero at $r = 0$, as assumed, requires that $\varphi(r)$ is also continuous at $r = 0$ and that $\varphi(0) = 0$. (Note that $|\psi|$ must be continuous and zero at $r = 0$. Otherwise it would have values that stay a finite amount above zero however close you get to $r = 0$. Then $|\psi|$ would not be zero in a meaningful sense. And here we want to exclude points of zero wave function. Excluding points of indeterminate wave function will be left as an exercise for the reader. But as already mentioned, that sort of singular behavior would require quite a singular potential.)

Take now some small sphere, of some small radius δ, around the supposed point of zero wave function. The value of φ on the outer surface of this sphere will be called ε. It will be assumed that there are no values of $\varphi(r)$ greater than ϵ inside the sphere. (If there are, you can always reduce the value of δ to get rid of them.)

The blunting inside this sphere will now be achieved by replacing the $\varphi(r)$ part of $|\psi|$ by ε. So the blunted wave function is:

$$\psi' \equiv |\psi| - \varphi(r) + \varepsilon$$

Consider now first what the corresponding increase in the potential energy integral inside the sphere is:

$$I'_V - I_V = \int \left[V(|\psi| - \varphi(r) + \varepsilon)^2 - V(|\psi|)^2\right] \mathrm{d}^3\vec{r}$$

Multiplying out the square, that becomes:

$$I'_V - I_V = \int 2V|\psi|(\varepsilon - \varphi(r))\,\mathrm{d}^3\vec{r} + \int V(\varepsilon - \varphi(r))^2\,\mathrm{d}^3\vec{r}$$

Since $\varphi(r)$ is nonnegative, it follows that the increase in potential energy is limited as

$$I'_V - I_V \leqslant 2V_{\max}\varepsilon \int |\psi|\,\mathrm{d}^3\vec{r} + V_{\max}\varepsilon^2 \int \mathrm{d}^3\vec{r} \qquad \int \mathrm{d}^3\vec{r} = C_n\delta^n$$

Note that the hypersphere formula for the volume of the sphere has been used. The purpose is to make the final result valid in any number n of dimensions, not just three dimensions. The remaining integral in the above expression can be rewritten as

$$\int |\psi|\,\mathrm{d}^3\vec{r} = \iint |\psi|\,\mathrm{d}A_n\mathrm{d}r = \int \left[\int |\psi|\,\mathrm{d}A_n/A_n\right] A_n\mathrm{d}r = \int \varphi A_n\mathrm{d}r \leqslant \varepsilon C_n\delta^n$$

So finally

$$I'_V - I_V \leqslant 3V_{\max}\varepsilon^2 C_n\delta^n$$

The next question is what happens to the kinetic energy. In three-dimensional spherical coordinates, the kinetic energy after blunting is

$$I'_T = \frac{\hbar^2}{2m}\int \left[\left(\frac{\partial\psi'}{\partial r}\right)^2 + \left(\frac{1}{r}\frac{\partial\psi'}{\partial\theta}\right)^2 + \left(\frac{1}{r\sin\theta}\frac{\partial\psi'}{\partial\phi}\right)^2\right] \mathrm{d}^3\vec{r}$$

The initial kinetic energy is given by a similar expression, with $|\psi|$ replacing ψ'. Now the expression for the angular derivatives in the integrand will be different in a different number of dimensions. For example, in two-dimensional polar coordinates, there will be no ϕ-derivative. But these angular derivatives are unchanged by the blunting and drop out in the difference in kinetic energy. So the decrease in kinetic energy becomes, after substituting for ψ' and simplifying:

$$I_T - I'_T = \frac{\hbar^2}{2m}\iint 2\left(\frac{\partial|\psi|}{\partial r}\right)\left(\frac{\partial\varphi}{\partial r}\right)\mathrm{d}A_n\mathrm{d}r - \frac{\hbar^2}{2m}\iint \left(\frac{\partial\varphi}{\partial r}\right)^2\mathrm{d}A_n\mathrm{d}r$$

Since φ does not depend on the angular coordinates, that can be written

$$I_T - I'_T = \frac{\hbar^2}{2m} \int 2 \left[\int \left(\frac{\partial |\psi|}{\partial r} \right) \frac{\mathrm{d}A_n}{A_n} \right] \left(\frac{\partial \varphi}{\partial r} \right) A_n \mathrm{d}r - \frac{\hbar^2}{2m} \int \left(\frac{\partial \varphi}{\partial r} \right)^2 A_n \mathrm{d}r$$

The expression between square brackets is just the r-derivative of φ. So the decrease in kinetic energy becomes, substituting in for A_n,

$$I_T - I'_T = \frac{\hbar^2}{2m} \int \left(\frac{\partial \varphi}{\partial r} \right)^2 n C_n r^{n-1} \mathrm{d}r$$

Note that the kinetic energy does decrease. The right hand side is positive. And if the maximum potential $V_{\max}$ in the vicinity of the point is negative, the potential energy decreases too. So that cannot be a ground state. It follows that the ground state wave function cannot become zero when $V_{\max}$ is negative or zero. (Do recall that the potential V was redefined. In terms of the original potential, there cannot be a zero if the potential is less than the expectation value of energy.)

But how about positive $V_{\max}$? Here the factor r^{n-1} in the kinetic energy integral is a problem in more than one dimension. In particular, if almost all the changes in φ occur at small r, the factor r^{n-1} will make the kinetic energy change small. Therefore there is no assurance that the kinetic energy decrease exceeds the potential energy increase. So a ground state cannot immediately be dismissed like in one dimension.

The solution is a trick. You might say that only a mathematician would think up a trick like that. However, the author insists that he is an aerospace engineer, not a mathematician. The first thing to note that there is a constraint on how much φ can change in the *outer* half of the sphere, for $r \geqslant \delta/2$. There the factor r^{n-1} is at least $\delta^{n-1}/2^{n-1}$. So the kinetic energy decrease is at least

$$I_T - I'_T \geqslant \frac{\hbar^2}{2m} \frac{n C_n}{2^{n-1}} \delta^{n-1} \int_{\delta/2}^{\delta} \left(\frac{\partial \varphi}{\partial r} \right)^2 \mathrm{d}r$$

Now the remaining integral can be estimated by the Cauchy-Schwartz inequality as before. Comparing this with the maximum possible increase in potential energy will give a limit on the maximum change in φ in the outer half of the sphere. In particular

$$\varepsilon - \varepsilon_{\mathrm{mid}} \leqslant \sqrt{\frac{3\, 2^{n-1}}{n} \frac{m V_{\max}}{\hbar^2}}\, \delta\, \varepsilon$$

where $\varepsilon_{\mathrm{mid}}$ denotes the value of φ at the midpoint $r = \delta/2$.

If the above inequality is not satisfied, the kinetic energy decrease would exceed the potential energy increase and it cannot be a ground state. Note

that if the sphere is chosen small, the relative decrease in φ is small too. For example, suppose you choose a sphere, call it sphere 1, with a radius

$$\delta_1 \leq \frac{1}{4}\sqrt{\frac{n}{3\,2^{n-1}}\frac{\hbar^2}{mV_{\max}}}$$

In that case,

$$\varepsilon_1 - \varepsilon_{\text{mid},1} \leqslant \tfrac{1}{4}\varepsilon_1$$

That means that φ can at most decrease by 25% going from the outside surface to the midpoint:

$$\varepsilon_{\text{mid},1} \geqslant \tfrac{3}{4}\varepsilon_1$$

You might say, "Why not?" And indeed, there would be nothing wrong with the idea that almost all the change would occur in the inner half of the sphere. But the idea is now to drive the mathematics into a corner from which eventually there is no escape. Suppose that you now define the inner half of sphere 1 to be a sphere 2. So the radius δ_2 of this sphere is half that of sphere 1, and its value of ε is

$$\varepsilon_2 = \varepsilon_{\text{mid},1} \geqslant \tfrac{3}{4}\varepsilon_1$$

(If there are φ values in the second sphere that exceed ε_2, you need to further reduce δ_2 to get rid of them. But all that does is reduce the possible changes in φ even more.) In this second sphere, the allowed relative decrease in its outer half is a factor 2 smaller than in sphere 1, because δ is a factor two smaller:

$$\varepsilon_{\text{mid},2} \geqslant \tfrac{7}{8}\varepsilon_2$$

Now take the midpoint as the radius of a sphere 3. Then

$$\varepsilon_3 = \varepsilon_{\text{mid},2} \geqslant \tfrac{7}{8}\varepsilon_2 \geqslant \tfrac{3}{4}\tfrac{7}{8}\varepsilon_1$$

Keep doing this and for sphere number i you get

$$\varepsilon_i = \tfrac{3}{4}\tfrac{7}{8}\tfrac{15}{16}\tfrac{31}{32}\cdots\tfrac{2^i-1}{2^i}\varepsilon_1$$

This must become zero for infinite i, because the sphere radii contract to zero and φ is zero at $r = 0$. But it does not! The allowed changes are simply too small to reach zero. Just take the logarithm:

$$\ln\varepsilon_i = \ln(1-\tfrac{1}{4}) + \ln(1-\tfrac{1}{8}) + \ln(1-\tfrac{1}{16}) + \ln(1-\tfrac{1}{32}) + \ldots + \ln\varepsilon_1$$

If ε_i becomes zero, its logarithm must become minus infinity. But the infinite sum does *not* become infinite. Just use the Taylor series approximation $\ln(1-x) \approx -x$:

$$\ln(1-\tfrac{1}{4}) + \ln(1-\tfrac{1}{8}) + \ln(1-\tfrac{1}{16}) + \ln(1-\tfrac{1}{32}) + \ldots = -[\tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{16} + \tfrac{1}{32} + \ldots]$$

The sum within square brackets is a geometric series that has a finite limit, not an infinite one.

So there is a contradiction. At some stage the decrease in kinetic energy *must* exceed the increase in potential energy. At that stage, the energy can be reduced by applying the blunting. So the assumed wave function cannot be a ground state.

You might still object that a Taylor series approximation is not exact. But in the region of interest

$$\ln(1-x) \geqslant -\frac{\ln(1-\frac{1}{4})}{-\frac{1}{4}} x$$

and the additional ratio is just a constant, about 1.15, that does not make a difference.

Woof.

D.23 Solution of the hydrogen molecule

To find the approximate solution for the hydrogen molecule, the key is to be able to find the expectation energy of the approximate wave functions $a\psi_\mathrm{l}\psi_\mathrm{r} + b\psi_\mathrm{r}\psi_\mathrm{l}$.

First, for given a/b, the individual values of a and b can be computed from the normalization requirement

$$a^2 + b^2 + 2ab\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle^2 = 1 \tag{D.11}$$

where the value of the overlap integral $\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle$ was given in derivation {D.21}.

The inner product

$$\langle a\psi_\mathrm{l}\psi_\mathrm{r} + b\psi_\mathrm{r}\psi_\mathrm{l}|H|a\psi_\mathrm{l}\psi_\mathrm{r} + b\psi_\mathrm{r}\psi_\mathrm{l}\rangle_6$$

is a six-dimensional integral, but when multiplied out, a lot of it can be factored into products of three-dimensional integrals whose values were given in derivation {D.21}. Cleaning up the inner product, and using the normalization condition, you can get:

$$\langle E\rangle = 2E_1 - \frac{e^2}{4\pi\epsilon_0}\left[A_1 + 2ab\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle^2 A_2\right]$$

using the abbreviations

$$A_1 = 2\langle\psi_\mathrm{l}|r_\mathrm{r}^{-1}\psi_\mathrm{l}\rangle - \frac{1}{d} - \langle\psi_\mathrm{l}\psi_\mathrm{r}|r_{12}^{-1}\psi_\mathrm{l}\psi_\mathrm{r}\rangle$$

$$A_2 = \frac{2\langle\psi_\mathrm{l}|r_\mathrm{l}^{-1}\psi_\mathrm{r}\rangle}{\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle} - 2\langle\psi_\mathrm{l}|r_\mathrm{r}^{-1}\psi_\mathrm{l}\rangle - \frac{\langle\psi_\mathrm{l}\psi_\mathrm{r}|r_{12}^{-1}\psi_\mathrm{r}\psi_\mathrm{l}\rangle}{\langle\psi_\mathrm{l}|\psi_\mathrm{r}\rangle^2} + \langle\psi_\mathrm{l}\psi_\mathrm{r}|r_{12}^{-1}\psi_\mathrm{l}\psi_\mathrm{r}\rangle$$

Values for several of the inner products in these expressions are given in derivation {D.21}. Unfortunately, these involving the distance $r_{12} = |\vec{r}_1 - \vec{r}_2|$ between the electrons cannot be done analytically. And one of the two cannot even be reduced to a three-dimensional integral, and needs to be done in six dimensions. (It can be reduced to five dimensions, but that introduces a nasty singularity and sticking to six dimensions seems a better idea.) So, it gets really elaborate, because you have to ensure numerical accuracy for singular, high-dimensional integrals. Still, it can be done with some perseverance.

In any case, the basic idea is still to print out expectation energies, easy to obtain or not, and to examine the print-out to see at what values of a/b and d the energy is minimal. That will be the ground state.

The results are listed in the main text, but here are some more data that may be of interest. At the 1.62 a_0 nuclear spacing of the ground state, the antisymmetric state $a/b = -1$ has a positive energy of 7 eV above separate atoms and is therefore unstable.

The nucleus to electron attraction energies are 82 eV for the symmetric state, and 83.2 eV for the antisymmetric state, so the antisymmetric state has the lower potential energy, like in the hydrogen molecular ion case, and unlike what you read in some books. The symmetric state has the lower energy because of lower kinetic energy, not potential energy.

Due to electron cloud merging, for the symmetric state the electron to electron repulsion energy is 3 eV lower than you would get if the electrons were point charges located at the nuclei. For the antisymmetric state, it is 5.8 eV lower.

As a consequence, the antisymmetric state also has less potential energy with respect to these repulsions. Adding it all together, the symmetric state has quite a lot less kinetic energy than the antisymmetric one.

D.24 Hydrogen molecule ground state and spin

The purpose of this note is to verify that the inclusion of spin does not change the spatial form of the ground state of the hydrogen molecule. The lowest expectation energy $\langle E \rangle = \langle \psi_{\rm gs} | H \psi_{\rm gs} \rangle$, characterizing the correct ground state, only occurs if all spatial components $\psi_{\pm\pm}$ of the ground state with spin,

$$\psi_{\rm gs} = \psi_{++}\uparrow\uparrow + \psi_{+-}\uparrow\downarrow + \psi_{-+}\downarrow\uparrow + \psi_{--}\downarrow\downarrow,$$

are proportional to the no-spin spatial ground state $\psi_{\rm gs,0}$.

The reason is that the assumed Hamiltonian (5.3) does not involve spin at all, only spatial coordinates, so, for example,

$$(H\psi_{++}\uparrow\uparrow) \equiv H\left(\psi_{++}(\vec{r}_1,\vec{r}_2)\uparrow(S_{z1})\uparrow(S_{z2})\right) = (H\psi_{++})\uparrow\uparrow$$

and the same for the other three terms in $H\psi_{\rm gs}$. So the expectation value of energy becomes

$$\begin{aligned}\langle E\rangle = \langle \psi_{++}\uparrow\uparrow + \psi_{+-}\uparrow\downarrow + \psi_{-+}\downarrow\uparrow + \psi_{--}\downarrow\downarrow \\ | \, (H\psi_{++})\uparrow\uparrow + (H\psi_{+-})\uparrow\downarrow + (H\psi_{-+})\downarrow\uparrow + (H\psi_{--})\downarrow\downarrow\rangle\end{aligned}$$

Because of the orthonormality of the spin states, this multiplies out into inner products of matching spin states as

$$\langle E\rangle = \langle\psi_{++}|H\psi_{++}\rangle + \langle\psi_{+-}|H\psi_{+-}\rangle + \langle\psi_{-+}|H\psi_{-+}\rangle + \langle\psi_{--}|H\psi_{--}\rangle.$$

In addition, the wave function must be normalized, $\langle\psi_{\rm gs}|\psi_{\rm gs}\rangle = 1$, or

$$\langle\psi_{++}|\psi_{++}\rangle + \langle\psi_{+-}|\psi_{+-}\rangle + \langle\psi_{-+}|\psi_{-+}\rangle + \langle\psi_{--}|\psi_{--}\rangle = 1.$$

Now when ψ_{++}, ψ_{+-}, ψ_{-+}, and ψ_{--} are each proportional to the no-spin spatial ground state $\psi_{\rm gs,0}$ with the lowest energy $E_{\rm gs}$, their individual contributions to the energy will be given by $\langle\psi_{\pm\pm}|H\psi_{\pm\pm}\rangle = E_{\rm gs}\,\langle\psi_{\pm\pm}|\psi_{\pm\pm}\rangle$, the lowest possible. Then the total energy $\langle E\rangle$ will be $E_{\rm gs}$. Anything else will have more energy and can therefore not be the ground state.

It should be pointed out that to a more accurate approximation, spin causes the electrons to be somewhat magnetic, and that produces a slight dependence of the energy on spin; compare addendum {A.39}. This note ignored that, as do most other derivations in this book.

D.25 Number of boson states

For identical bosons, the number is $I + N - 1$ choose I. To see that think of the I bosons as being inside a series of N single particle-state "boxes." The idea is as illustrated in figure D.2; the circles are the bosons and the thin lines separate the boxes. In the picture as shown, each term in the group of states has one boson in the first single-particle function, three bosons in the second, three bosons in the third, etcetera.



Figure D.2: Bosons in single-particle-state boxes.

Each picture of this type corresponds to exactly one system state. To figure out how many different pictures there are, imagine there are numbers written from 1 to I on the bosons and from $I+1$ to $I+N-1$ on the separators between the boxes. There are then $(I + N - 1)!$ ways to arrange that total of $I + N - 1$ objects. (There are $I+N-1$ choices for which object to put first, times $I+N-2$ choices for which object to put second, etcetera.) However, the $I!$ different ways

to order the subset of boson numbers do not produce different pictures if you erase the numbers again, so divide by $I!$. The same way, the different ways to order the subset of box separator numbers do not make a difference, so divide by $(N-1)!$.

For example, if $I = 2$ and $N = 4$, you get $5!/2!3!$ or 10 system states.

D.26 Density of states

This note derives the density of states for particles in a box.

Consider the wave number space, as shown to the left in figure 6.1. Each point represents one spatial state. The first question is how many points have a wave number vector whose length $\underline{k}$ is less than some given value k. Since the length of the wave number vector is the distance from the origin in wave number state, the points with $\underline{k} < k$ form an octant of a sphere with radius k. In fact, you can think of this problem as finding the number of red points in figure 6.11.

Now the octant of the sphere has a "volume" (in wave number space, not a physical volume)

$$\text{octant volume:} \quad \tfrac{1}{8}\tfrac{4}{3}\pi k^3$$

Conversely, every wave number point is the top-left front corner of a little block of "volume"

$$\text{single state volume:} \quad \Delta k_x \Delta k_y \Delta k_z$$

where Δk_x, Δk_y, and Δk_z are the spacings between the points in the x, y, and z directions respectively. To find the approximate number of points inside the octant of the sphere, take the ratio of the two "volumes:"

$$\text{number of spatial states inside:} \quad \frac{\pi k^3}{6 \Delta k_x \Delta k_y \Delta k_z}$$

Now the spacings between the points are given in terms of the sides ℓ_x, ℓ_y, and ℓ_z of the box containing the particles as, (6.3),

$$\Delta k_x = \frac{\pi}{\ell_x} \qquad \Delta k_y = \frac{\pi}{\ell_y} \qquad \Delta k_z = \frac{\pi}{\ell_z}$$

Plug this into the expression for the number of points in the octant to get:

$$\text{number of spatial states inside:} \quad \frac{\mathcal{V}}{6\pi^2} k^3 \tag{D.12}$$

where $\mathcal{V}$ is the (physical) volume of the box $\ell_x \ell_y \ell_z$. Each wave number point corresponds to one spatial state, but if the spin of the particles is s then each spatial state still has $2s+1$ different spin values. Therefore multiply by $2s+1$ to get the number of states.

To get the density of states on a wave number basis, take the derivative with respect to k. The number of states $\mathrm{d}N$ in a small wave number range $\mathrm{d}k$ is then:

$$\mathrm{d}N = \mathcal{V}\mathcal{D}_k \,\mathrm{d}k \qquad \mathcal{D}_k = \frac{2s+1}{2\pi^2}k^2$$

The factor $\mathcal{D}_k$ is the density of states on a wave number basis.

To get the density of states on an energy basis, simply eliminate k in terms of the single-particle energy E^{p} using $E^{\mathrm{p}} = \hbar k^2/2m$. That gives:

$$\mathrm{d}N = \mathcal{V}\mathcal{D}\,\mathrm{d}E^{\mathrm{p}} \qquad \mathcal{D} = \frac{2s+1}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E^{\mathrm{p}}}$$

The used expression for the kinetic energy E^{p} is only valid for nonrelativistic speeds.

The above arguments fail in the presence of confinement. Recall that each state is the top-left front corner of a little block in wave number space of volume $\Delta k_x \Delta k_y \Delta k_z$. The number of states with wave number $\underline{k}$ less than some given value k was found by computing how many such little block volumes are contained within the octant of the sphere of radius k.

The problem is that a wave number $\underline{k}$ is only inside the sphere octant if all of its little block is inside. Even if 99% of its block is inside, the state itself will still be outside, not 99% in. That makes no difference if the states are densely spaced in wave number space, like in figure 6.11. In that case almost all little blocks are fully inside the sphere. Only a thin layer of blocks near the surface of the sphere are partially outside it.

However, confinement in a given direction makes the corresponding spacing in wave number space large. And that changes things.

In particular, if the y-dimension ℓ_y of the box containing the particles is small, then $\Delta k_y = \pi/\ell_y$ is large. That is illustrated in figure 6.12. In this case, there are no states inside the sphere at all if k is less than Δk_y. Regardless of what (D.12) claims. In the range $\Delta k_y < k < 2\Delta k_y$, illustrated by the red sphere in figure 6.12, the red sphere gobbles up a number of states from the plate $\underline{k}_y = \Delta k_y$. This number of states can be estimated as

$$\frac{\frac{1}{4}\pi(k_x^2 + k_z^2)}{\Delta k_x \Delta k_z}$$

since the top of this ratio is the area of the quarter circle of states and the bottom is the rectangular area occupied per state.

This expression can be cleaned up by noting that

$$k_x^2 + k_z^2 = k^2 - k_y^2 = k^2 - (n_y \Delta k_y)^2$$

with $n_y = 1$ for the lowest plate. Substituting for Δk_x, Δk_y, and Δk_z in terms of the box dimensions then gives

$$\text{spatial states per plate:} \quad \frac{A}{4\pi}\left[k^2 - \left(n_y \frac{\pi}{\ell_y}\right)^2\right] \qquad \text{if} \quad \Big[\ldots\Big] > 0 \qquad \text{(D.13)}$$

Here $A = \ell_x \ell_z$ is the area of the quantum well and $n_y = 1$ is the plate number. For nonrelativistic speeds k^2 is proportional to the energy E^{p}. Therefore the density of states, which is the derivative of the number of states with respect to energy, is constant.

In the range $2\pi/\ell_y < k < 3\pi/\ell_y$ a second quarter circle of states gets added. To get the number of additional states in that circle, use $n_y = 2$ for the plate number in (D.13). For still larger values of k, just keep summing plates as long as the expression between the square brackets in (D.13) remains positive.

If the z-dimension of the box is also small, like in a quantum wire, the states in wave number space separate into individual lines, figure 6.13. There are now no states until the sphere of radius k hits the line that is closest to the origin, having quantum numbers $n_y = n_z = 1$. Beyond that value of k, the number of states on the line that is within the sphere is

$$\frac{\sqrt{k^2 - (n_y \Delta k_y)^2 - (n_z \Delta k_z)^2}}{\Delta k_x}$$

since the top is the length of the line inside the sphere and the bottom the spacing of the states on the line. Cleaning up, that gives

$$\text{spatial states per line:} \quad \frac{\ell}{\pi}\left[k^2 - \left(n_y \frac{\pi}{\ell_y}\right)^2 - \left(n_z \frac{\pi}{\ell_z}\right)^2\right]^{1/2} \qquad \text{if} \quad \Big[\ldots\Big] > 0 \tag{D.14}$$

with $\ell = \ell_x$ the length of the quantum wire. For still larger values of k sum over all values of n_y and n_z for which the argument of the square root remains positive.

For nonrelativistic speeds, k^2 is proportional to the energy. Therefore the above number of states is proportional to the square root of the amount of energy above the one at which the line of states is first hit. Differentiating to get the density of states, the square root becomes an reciprocal square root.

If the box is small in all three directions, figure 6.14, the number of states simply becomes the number of points inside the sphere:

$$\text{spatial states per point:} \quad 1 \qquad \left[k^2 - \left(n_x \frac{\pi}{\ell_x}\right)^2 - \left(n_y \frac{\pi}{\ell_y}\right)^2 - \left(n_z \frac{\pi}{\ell_z}\right)^2\right] > 0 \tag{D.15}$$

In other words, to get the total number of states inside, simply add a 1 for each set of natural numbers n_x, n_y, and n_z for which the expression in brackets is positive. The derivative with respect to energy, the density of states, becomes a series of delta functions at the energies at which the states are hit.

D.27 Radiation from a hole

To find how much blackbody radiation is emitted from a small hole in a box, first imagine that all photons move in the direction normal to the hole with the

speed of light c. In that case, in a time interval $\mathrm{d}t$, a cylinder of photons of volume $Ac\mathrm{d}t$ would leave through the hole, where A is the hole area. To get the electromagnetic energy in that cylinder, simply multiply by Planck's blackbody spectrum ρ. That gives the surface radiation formula except for an additional factor $\frac{1}{4}$. Half of that factor is due to the fact that on average only half of the photons will have a velocity component in the direction normal to the hole that is towards the hole. The other half will have a velocity component in that direction that is away from the hole. In addition, because the photons move in all directions, the average velocity component of the photons that move towards the hole is only half the speed of light.

More rigorously, assume that the hole is large compared to $c\mathrm{d}t$. The fraction of photons with velocity directions within in a spherical element $\sin\theta\mathrm{d}\theta\mathrm{d}\phi$ will be $\sin\theta\mathrm{d}\theta\mathrm{d}\phi/4\pi$. The amount of these photons that exits will be those in a skewed cylinder of volume $Ac\cos\theta\mathrm{d}t$. To get the energy involved multiply by ρ. So the energy leaving in this small range of velocity directions is

$$\rho Ac\mathrm{d}t\cos\theta\frac{\sin\theta\mathrm{d}\theta\mathrm{d}\phi}{4\pi}$$

Integrate over all ϕ and θ up to 90 degrees to get $\frac{1}{4}\rho Ac\mathrm{d}t$ for the total energy that exits.

Note also from the above expression that the amount of energy leaving per unit time, unit area, and unit solid angle is

$$\frac{\rho c}{4\pi}\cos\theta$$

where θ is the angle from the normal to the hole.

D.28 Kirchhoff's law

Suppose you have a material in thermodynamic equilibrium at a given temperature that has an emissivity at a given frequency that exceeds the corresponding absorptivity. Place it in a closed box. Since it emits more radiation at the given frequency than it absorbs from the surrounding blackbody radiation, the amount of radiation at that frequency will go up. That violates Plank's blackbody spectrum, because it remains a closed box. The case that the emissivity is less than the absorptivity goes similarly.

Note some of the implicit assumptions made in the argument. First, it assumes linearity, in the sense that emission or absorption at one frequency does not affect that at another, that absorption does not affect emission, and that the absorptivity is independent of the amount absorbed. It assumes that the surface is separable from the object you are interested in. Transparent materials require special consideration, but the argument that a layer of such material must emit the same fraction of blackbody radiation as it absorbs remains valid.

The argument also assumes the validity of Plank's blackbody spectrum. However you can make do without. Kirchhoff did. He (at first) assumed that there are gage materials that absorb and emit only in a narrow range of frequencies, and that have constant absorptivity $a_{\rm g}$ and emissivity $e_{\rm g}$ in that range. Place a plate of that gage material just above a plate of whatever material is to be examined. Insulate the plates from the surrounding. Wait for thermal equilibrium.

Outside the narrow frequency range, the material being examined will have to absorb the same radiation energy that it emits, since the gage material does not absorb nor emit outside the range. In the narrow frequency range, the radiation energy $\dot{E}$ going up to the gage plate must equal the energy coming down from it again, otherwise the gage plate would continue to heat up. If B is the blackbody value for the radiation in the narrow frequency range, then the energy going down from the gage plate consists of the radiation that the gage plate emits plus the fraction of the incoming radiation that it reflects instead of absorbs:

$$\dot{E} = e_{\rm g} B + (1 - a_{\rm g})\dot{E} \quad \Longrightarrow \quad \dot{E}/B = e_{\rm g}/a_{\rm g}$$

Similarly for the radiation going up from the material being examined:

$$\dot{E} = eB + (1 - a)\dot{E} \quad \Longrightarrow \quad \dot{E}/B = e/a$$

By comparing the two results, $e/a = e_{\rm g}/a_{\rm g}$. Since you can examine any material in this way, all materials must have the same ratio of emissivity to absorptivity in the narrow range. Assuming that gage materials exist for every frequency range, at any frequency e/a must be the same for all materials. So it must be the blackbody value 1.

No, this book does not know where to order these gage materials, [38]. And the same argument cannot be used to show that the absorptivity must equal emissivity in each individual direction of radiation, since direction is not preserved in reflections.

D.29 The thermionic emission equation

This note derives the thermionic emission equation for a typical metal following [42, p. 364ff]. The derivation is semi-classical.

To simplify the analysis, it will be assumed that the relevant electrons in the interior of the metal can be modeled as a free-electron gas. In other words, it will be assumed that in the interior of the metal the forces from surrounding particles come from all directions and so tend to average out.

(The free-electron gas assumption is typically qualitatively reasonable for the valence electrons of interest if you define the zero of the kinetic energy of the gas to be at the bottom of the conduction band. You can also reduce errors

by replacing the true mass of the electron by some suitable "effective mass." But the zero of the energy drops out in the final expression, and the effective mass of typical simple metals is not greatly different from the true mass. See chapter 6.22.3 for more on these issues.)

Assume that the surface through which the electrons escape is normal to the x-direction. Then the classical expression for the current of escaping electrons is

$$j = \rho e v_x$$

where ρ is the number of electrons per unit volume that is capable of escaping and v_x is their velocity in the x-direction. Note that the current above is supposed to be the current *inside* the metal of the electrons that will escape.

An electron can only escape if its energy E^{p} exceeds

$$E^{\mathrm{p}}_{\mathrm{esc}} = \mu + e\varphi_{\mathrm{w}}$$

where μ is the Fermi level, because the work function φ_{w} is defined that way. The number of electrons per unit volume in an energy range $\mathrm{d}E^{\mathrm{p}}$ above $E^{\mathrm{p}}_{\mathrm{esc}}$ can be found as

$$e^{-(e\varphi_{\mathrm{w}}+E^{\mathrm{p}}-E^{\mathrm{p}}_{\mathrm{esc}})/k_{\mathrm{B}}T}\ \frac{2}{4\pi^2}\left(\frac{2m_{\mathrm{e}}}{\hbar^2}\right)^{3/2}\sqrt{E^{\mathrm{p}}}\,\mathrm{d}E^{\mathrm{p}}$$

That is because the initial exponential is a rewritten Maxwell-Boltzmann distribution (6.21) that gives the number of electrons per state, while the remainder is the number of states in the energy range according to the density of states (6.6).

Normally, the typical thermal energy $k_{\mathrm{B}}T$ is very small compared to the minimum energy $e\varphi_{\mathrm{w}}$ above the Fermi level needed to escape. Then the exponential of the Maxwell-Boltzmann distribution is very small. That makes the amount of electrons with sufficient energy to escape very small. In addition, with increasing energy above $E^{\mathrm{p}}_{\mathrm{esc}}$ the amount of electrons very quickly becomes much smaller still. Therefore only a very small range of energies above the minimum energy $E^{\mathrm{p}}_{\mathrm{esc}}$ gives a contribution.

Further, even if an electron has in principle sufficient energy to escape, it can only do so if enough of its momentum is in the x-direction. Only momentum that is in the x-direction can be used to overcome the nuclei that pull it back towards the surface when it tries to escape. Momentum in the other two directions only produces motion parallel to the surface. So only a fraction, call it f_{esc}, of the electrons that have in principle enough energy to escape can actually do so. A bit of geometry shows how much. All possible end points of the momentum vectors with a magnitude p form a spherical surface with area $4\pi p^2$. But only a small circle on that surface around the x-axis, with an approximate radius of $\sqrt{p^2 - p^2_{\mathrm{esc}}}$, has enough x-momentum for the electron to escape, so

$$f_{\mathrm{esc}} \approx \frac{\pi\sqrt{p^2 - p^2_{\mathrm{esc}}}^2}{4\pi p^2} \approx \frac{E^{\mathrm{p}} - E^{\mathrm{p}}_{\mathrm{esc}}}{4E^{\mathrm{p}}}$$

where the final equality applies since the kinetic energy is proportional to the square momentum.

Since the velocity for the escaping electrons is mostly in the x-direction, $E^\mathrm{p} \approx \frac{1}{2} m_\mathrm{e} v_x^2$, which can be used to express v_x in terms of energy.

Putting it all together, the current density becomes

$$j = \int_{E^\mathrm{p}=E^\mathrm{p}_\mathrm{esc}}^{\infty} e^{-(e\varphi_\mathrm{w}+E^\mathrm{p}-E^\mathrm{p}_\mathrm{esc})/k_\mathrm{B}T} \frac{2}{4\pi^2} \left(\frac{2m_\mathrm{e}}{\hbar^2}\right)^{3/2} \sqrt{E^\mathrm{p}}\, \frac{E^\mathrm{p}-E^\mathrm{p}_\mathrm{esc}}{4E^\mathrm{p}} \left(\frac{2E^\mathrm{p}}{m_\mathrm{e}}\right)^{1/2} \mathrm{d}E^\mathrm{p}$$

Rewriting in terms of a new integration variable $u = (E^\mathrm{p} - E^\mathrm{p}_\mathrm{esc})/k_\mathrm{B}T$ gives the thermionic emission equation.

If an external electric field $\mathcal{E}_\mathrm{ext}$ helps the electrons escape, it lowers the energy that the electrons need to do so. Consider the potential energy in the later stages of escape, at first still without the additional electric field. When the electron looks back at the metal surface that it is escaping from, it sees a positron mirror image of itself inside the metal. Of course, there is not really a positron inside the metal; rearrangement of the surface electrons of the metal create this illusion. The surface electrons rearrange themselves to make the total component of the electric field in the direction parallel to the surface zero. Indeed, they have to keep moving until they do so, since the metal has negligible electrical resistance in the direction parallel to the surface. Now it just so happens that a positron mirror image of the electron has exactly the same effect as this rearrangement. The escaping electron pushes the surface electrons away from itself; that force has a repulsive component along the surface. The positron mirror image however attracts the surface electrons towards itself, exactly cancelling the component of force along the surface exerted by the escaping electron.

The bottom line is that it seems to the escaping electron that it is pulled back not by surface charges, but by a positron mirror image of itself. Therefore, including now an additional external electrical field, the total potential in the later stages of escape is:

$$V = -\frac{e^2}{16\pi\epsilon_0 d} - e\mathcal{E}_\mathrm{ext} d + \text{constant}$$

where d is the distance from the surface. The first term is the attracting force due to the positron image, while the second is due to the external electric field. The constant depends on where the zero of energy is defined. Note that only half the energy of attraction between the electron and the positron image should be assigned to the electron; the other half can be thought of as "work" on the image. If that is confusing, just write down the force on the electron and integrate it to find its potential energy.

If there is no external field, the maximum potential energy that the electron must achieve occurs at infinite distance d from the metal surface. If there is an electric field, it lowers the maximum potential energy, and it now occurs

somewhat closer to the surface. Setting the derivative of V with respect to d to zero to identify the maximum, and then evaluating V at that location shows that the external field lowers the maximum potential energy that must be achieved to escape by $\sqrt{e^3\mathcal{E}/4\pi\epsilon_0}$.

D.30 Number of conduction band electrons

This note finds the number of electrons in the conduction band of a semiconductor, and the number of holes in the valence band.

By definition, the density of states $\mathcal{D}$ is the number of single-particle states per unit energy range and unit volume. The fraction of electrons in those states is given by ι_e. Therefore the number of electrons in the conduction band per unit volume is given by

$$i_e = \int_{E_c^p}^{E_{top}^p} \mathcal{D}\iota_e \, \mathrm{d}E^p$$

where E_c^p is the energy at the bottom of the conduction band and E_{top}^p that at the top of the band.

To compute this integral, for ι_e the Maxwell-Boltzmann expression (6.33) can be used, since the number of electrons per state is invariably small. And for the density of states the expression (6.6) for the free-electron gas can be used if you substitute in a suitable effective mass of the electrons and replace $\sqrt{E^p}$ by $\sqrt{E^p - E_c^p}$.

Also, because ι_e decreases extremely rapidly with energy, only a very thin layer at the bottom of the conduction band makes a contribution to the number of electrons. The integrand of the integral for i_e is essentially zero above this layer. Therefore you can replace the upper limit of integration with infinity without changing the value of i_e. Now use a change of integration variable to $u = \sqrt{(E^p - E_c^p)/k_B T}$ and an integration by parts to reduce the integral to the one found under "!" in the notations section. The result is as stated in the text.

For holes, the derivation goes the same way if you use ι_h from (6.34) and integrate over the valence band energies.

D.31 Integral Schrödinger equation

In this note, the integral Schrödinger equation is derived from the partial differential equation version.

First the time-independent Schrödinger equation is rewritten in the form

$$\left(\nabla^2 + k^2\right)\psi = f \qquad k = \frac{\sqrt{2mE}}{\hbar} \quad f = \frac{2mV}{\hbar^2}\psi \tag{D.16}$$

The left equation is known as the "Helmholtz equation."

The Helmholtz equation is not at all specific to quantum mechanics. In general it describes basic wave propagation at a frequency related to the value of the constant k. The right hand side f describes the amount of wave motion that is created at a given location. Quantum mechanics is somewhat weird in that f involves the unknown wave function ψ that you want to find. In simpler applications, f is a given function.

The general solution to the Helmholtz equation can be written as

$$\boxed{\left(\nabla^2 + k^2\right)\psi = f \quad \Longrightarrow \quad \psi = \psi_0 - \int_{\text{all } \vec{r}\,'} \frac{e^{\mathrm{i}k|\vec{r}\,'-\vec{r}|}}{4\pi|\vec{r}\,'-\vec{r}|} f(\vec{r}\,')\,\mathrm{d}^3\vec{r}\,'} \tag{D.17}$$

Here ψ_0 is any solution of the *homogeneous* Helmholtz equation, the equation without f.

To see why this is the solution of the Helmholtz equation requires a bit of work. First consider the solution of the Helmholtz equation for the special case that f is a delta function at the origin:

$$\left(\nabla^2 + k^2\right) G = \delta^3(\vec{r})$$

The solution G to this problem is called the "Green's function of the Helmholtz equation.

The Green's function can be found relatively easily. Away from the origin G is a solution of the homogeneous Helmholtz equation, because the delta function is everywhere zero except at the origin. In terms of quantum mechanics, the homogeneous Helmholtz equation means a particle in free space, $V = 0$. Possible solutions for G are then spherical harmonics times spherical Hankel functions of the first and second kinds, {A.6}. However, Hankel functions of the first kind are preferred for physical reasons; they describe waves that propagate away from the region of wave generation to infinity. Hankel functions of the second kind describe waves that come in from infinity. Incoming waves, if any, are usually much more conveniently described using the homogeneous solution ψ_0.

Further, since the problem for G is spherically symmetric, the solution should not depend on the angular location. The spherical harmonic must be the constant Y_0^0. That makes the correct solution a multiple of the spherical Hankel function $h_0^{(1)}$, which means proportional to $e^{\mathrm{i}kr}/r$. You can easily check by direct substitution that this does indeed satisfy the homogeneous Helmholtz equation away from the origin in spherical coordinates.

To get the correct constant of proportionality, integrate the Helmholtz equation for G above over a small sphere around the origin. In the right hand side use the fact that the integral of a delta function is by definition equal to 1. In the left hand side, use the divergence theorem to avoid having to try to integrate the singular second order derivatives of G at the origin. That shows that the complete Green's function is

$$G(\vec{r}) = -\frac{e^{\mathrm{i}kr}}{4\pi r} \qquad r = |\vec{r}|$$

(You might worry about the mathematical justification for these manipulations. Singular functions like G are not proper solutions of partial differential equations. However, the real objective is to find the limiting solution G when a slightly smoothed delta function becomes truly singular. The described manipulations are justified in this limiting process.)

The next step is to solve the Helmholtz equation for an arbitrary right hand side f, rather than a delta function. To do so, imagine the region subdivided into infinitely many infinitesimal volume elements $\mathrm{d}\vec{r}\,'$. In each volume element, approximate the function f by a delta function spike $\delta(\vec{r}-\vec{r}\,')f(\vec{r}\,')\,\mathrm{d}\vec{r}\,'$. Such a spike integrates to the same value as f does over the volume element. Each spike produces a solution given by

$$G(\vec{r}-\vec{r}\,')f(\vec{r}\,')\,\mathrm{d}\vec{r}\,'$$

Integrate over all volume elements to get the solution of the Helmholtz equation (D.17). Substitute in what f is for the Schrödinger equation to get the integral Schrödinger equation.

D.32 Integral conservation laws

This section derives the integral conservation laws given in addendum {A.14}.

The rules of engagement are as follows:

- The Cartesian axes are numbered using an index i, with $i = 1$, 2, and 3 for x, y, and z respectively.
- Also, r_i indicates the coordinate in the i direction, x, y, or z.
- Derivatives with respect to a coordinate r_i are indicated by a simple subscript i.
- Time derivatives are indicated by a subscript t.
- A bare $\int$ integral sign is assumed to be an integration over all space, or over the entire box for particles in a box. The $\mathrm{d}^3\vec{r}$ is normally omitted for brevity and to be understood.
- A superscript * indicates a complex conjugate.

First it will be shown that according to the Schrödinger equation $\int |\Psi|^2$ is constant. The Schrödinger equation in free space is

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi$$

Taking the right hand term to the other side and writing it in index notation gives

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} + \sum_i \frac{\hbar^2}{2m}\Psi_{ii} = 0$$

Multiply the left hand side by $\Psi^*/\mathrm{i}\hbar$ and add the complex conjugate of the same equation to get

$$\Psi^* \frac{\partial \Psi}{\partial t} + \Psi \frac{\partial \Psi^*}{\partial t} + \sum_i \frac{\hbar}{2m\mathrm{i}} (\Psi^* \Psi_{ii} - \Psi \Psi^*_{ii}) = 0$$

To show that the integral $\int |\Psi|^2$ is constant, it must be shown that its time derivative is zero. Now the first two terms above are the time derivative of $|\Psi|^2 = \Psi^*\Psi$. So integrated over all space, they give the time derivative that must be shown to be zero. And the equation above shows that it is indeed zero provided that the remaining sum in it integrates to zero over all space.

The constant is not important in showing that this is true, so just examine for any i

$$\int (\Psi^* \Psi_{ii} - \Psi \Psi^*_{ii})$$

This equals

$$\int (\Psi^* \Psi_i - \Psi \Psi^*_i)_i$$

as can be seen by differentiating out the parenthetical expression with respect to r_i. The above integrand can be integrated with respect to r_i. It will then be zero for a periodic box since the expression in parenthesis is the same at the upper and lower limits of integration. It will also be zero for an impenetrable container, since Ψ will then be zero on the surface of the container. It will also be zero in an infinite region provided that Ψ and its derivatives vanish at large distances.

There is another way to see that $\int |\Psi|^2$ is constant. First recall that any solution of the Schrödinger equation takes the form

$$\Psi = \sum_n c_n e^{-\mathrm{i}E_n t} \psi_n(\vec{r})$$

Here the ψ_n are the energy eigenfunctions. Then

$$\int |\Psi|^2 = \int \Psi^* \Psi = \int \sum_{\underline{n}} c^*_{\underline{n}} e^{\mathrm{i}E_{\underline{n}} t} \psi_{\underline{n}}(\vec{r}) \sum_n c_n e^{-\mathrm{i}E_n t} \psi_n(\vec{r})$$

Now because of orthonormality of the eigenfunctions, the integration only produces a nonzero result when $\underline{n} = n$, and then the product of the eigenfunctions integrates to 1. So

$$\int |\Psi|^2 = \sum_n c^*_n c_n$$

That does not depend on time, and the normalization requirement makes it 1.

This also clarifies what goes wrong with the Klein-Gordon equation. For the Klein-Gordon equation

$$\Psi = \sum_n c_n e^{-\mathrm{i}E_n t}\psi_n(\vec{r}) + \sum_n d_n e^{\mathrm{i}E_n t}\psi_n(\vec{r})$$

The first sum are the particle states and the second sum the antiparticle states. That gives:

$$\int |\Psi|^2 = \sum_n \left(c_n^* c_n + d_n^* d_n + c_n^* d_n e^{2\mathrm{i}E_n t} + d_n^* c_n e^{-2\mathrm{i}E_n t}\right)$$

The final two terms in the sum oscillate in time. So the integral is no longer constant.

The exception is if there are only particle states (no d_n) or only antiparticle states (no c_n). In those two cases, the integral is constant. In general

$$\Psi = \Psi_1 + \Psi_2 \qquad \Psi_1 = \sum_n c_n e^{-\mathrm{i}E_n t}\psi_n(\vec{r}) \quad \Psi_2 = \sum_n d_n e^{\mathrm{i}E_n t}\psi_n(\vec{r})$$

where the integrated square magnitudes of Ψ_1 and Ψ_2 are constant.

Next it will be shown that the rearranged Klein-Gordon equation

$$\frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2} - \sum_i \Psi_{ii} + \left(\frac{mc^2}{\hbar c}\right)^2 \Psi = 0$$

preserves the sum of integrals

$$\int \left|\frac{1}{c}\frac{\partial\Psi}{\partial t}\right|^2 + \int \sum_i |\Psi_i|^2 + \int \left|\frac{mc^2}{\hbar c}\Psi\right|^2$$

To do so it suffices to show that the sum of the time derivatives of the three integrals is zero. That can be done by multiplying the Klein-Gordon equation by $\partial\Psi^*/\partial t$, adding the complex conjugate of the obtained equation, and integrating over all space. Each of the three terms in the Klein-Gordon equation will then give one of the three needed time derivatives. So their sum will indeed be zero.

To check that, look at what each term in the Klein-Gordon equation produces separately. The first term gives

$$\int \frac{1}{c^2}\left(\frac{\partial\Psi^*}{\partial t}\frac{\partial^2\Psi}{\partial t^2} + \frac{\partial\Psi}{\partial t}\frac{\partial^2\Psi^*}{\partial t^2}\right)$$

or taking one time derivative outside the integral, that is

$$\frac{1}{c^2}\frac{\mathrm{d}}{\mathrm{d}t}\int \frac{\partial\Psi^*}{\partial t}\frac{\partial\Psi}{\partial t}$$

That is the first needed time derivative, since a number times its complex conjugate is the square magnitude of that number.

The second term in the Klein-Gordon equation produces

$$-\sum_i \int \left(\frac{\partial \Psi^*}{\partial t} \Psi_{ii} + \frac{\partial \Psi}{\partial t} \Psi^*_{ii} \right)$$

That equals

$$-\sum_i \int \left(\frac{\partial \Psi^*}{\partial t} \Psi_i + \frac{\partial \Psi}{\partial t} \Psi^*_i \right)_i + \sum_i \frac{\mathrm{d}}{\mathrm{d}t} \int \Psi^*_i \Psi_i$$

as can be seen by differentiating out the parenthetical expression in the first integral with respect to r_i and bringing the time derivative in the second term inside the integral. The first integral above is zero for a periodic box, for an impenetrable container, and for infinite space for the same reasons as given in the derivation for the Schrödinger equation. The second term above is the needed second time derivative.

The final of the three terms in the Klein-Gordon equation produces

$$\left(\frac{mc^2}{\hbar c} \right)^2 \int \left(\frac{\partial \Psi^*}{\partial t} \Psi + \frac{\partial \Psi}{\partial t} \Psi^* \right)$$

That equals

$$\left(\frac{mc^2}{\hbar c} \right)^2 \frac{\mathrm{d}}{\mathrm{d}t} \int \Psi^* \Psi$$

as can be seen by bringing the time derivative inside the integral. This is the last of the three needed time derivatives.

D.33 Quantum field derivations

This derivation will find the properties of a system described by a Hamiltonian of the form:

$$H = E^{\mathrm{p}} \left(\widehat{P}^2 + \widehat{Q}^2 \right) + E_{\mathrm{ref}} \tag{1}$$

Here $\widehat{P}$ and $\widehat{Q}$ are Hermitian operators with commutator

$$\left[\widehat{P}, \widehat{Q} \right] = -\tfrac{1}{2}\mathrm{i} \tag{2}$$

and E^{p} and E_{ref} are constants with units of energy.

First note that the commutator (2) directly implies the uncertainty relationship, chapter 4.5.2 (4.46):

$$\sigma_P \sigma_Q \geqslant \tfrac{1}{4} \tag{3}$$

Also note that the evolution equations for the expectation values of P and Q follow directly from chapter 7.2 (7.4). The commutator appearing in it is readily worked out using the commutator (2) and the rules of chapter 4.5.4. Since energy eigenstates are stationary, according to the evolution equations in such states the expectation values of P and Q will have to be zero.

The equality of the $\widehat{P}$ and $\widehat{Q}$ terms in the Hamiltonian is a simple matter of symmetry. Nothing changes if you swap $\widehat{P}$ and $\widehat{Q}$, adding a minus sign for one. Then unavoidably the two terms in the Hamiltonian must be equal; it is shown below that the eigenfunctions are unique.

The commutator (2) also implies that $\widehat{P}$, $\widehat{Q}$, and all their combinations, do not commute with the Hamiltonian. So they are not conserved quantities of the system. However, there are two combinations,

$$\widehat{a} \equiv \widehat{P} - \mathrm{i}\widehat{Q} \qquad \widehat{a}^\dagger \equiv \widehat{P} + \mathrm{i}\widehat{Q} \tag{4}$$

whose commutator with the Hamiltonian gives back a multiple of the same thing:

$$[H, \widehat{a}] = -E^{\mathrm{p}} \widehat{a} \qquad [H, \widehat{a}^\dagger] = E^{\mathrm{p}} \widehat{a}^\dagger$$

In other words, $\widehat{a}$ and $\widehat{a}^\dagger$ are commutator eigenoperators of the Hamiltonian. The above relations are readily checked using the given commutator (2) and the rules of chapter 4.5.4.

To see why that is important, multiply both sides of the eigenvalue problems above with a system energy eigenfunction of energy E:

$$[H, \widehat{a}]\psi_E = -E^{\mathrm{p}} \widehat{a}\psi_E \qquad [H, \widehat{a}^\dagger]\psi_E = E^{\mathrm{p}} \widehat{a}^\dagger \psi_E$$

After writing out the definitions of the commutators, recognizing $H\psi_E$ as $E\psi_E$, and rearranging, that gives

$$H(\widehat{a}\psi_E) = (E - E^{\mathrm{p}})(\widehat{a}\psi_E) \qquad H(\widehat{a}^\dagger \psi_E) = (E + E^{\mathrm{p}})(\widehat{a}^\dagger \psi_E)$$

These results can be compared to the definition of an energy eigenfunction. Then it is seen that $\widehat{a}\psi_E$ is an energy eigenfunction with one unit E^{p} less energy than ψ_E. And $\widehat{a}^\dagger \psi_E$ is an energy eigenfunction with one unit E^{p} more energy than ψ_E. So apparently $\widehat{a}$ and $\widehat{a}^\dagger$ act as annihilation and creation operators of quanta of energy E^{p}. They act as shown to the left in figure A.6.

There are however two important caveats for these statements. If $\widehat{a}\psi_E$ or $\widehat{a}^\dagger \psi_E$ is zero, it is not an energy eigenfunction. Eigenfunctions must be nonzero. Also, even if the states $\widehat{a}\psi_E$ or $\widehat{a}^\dagger \psi_E$ are not zero, they will not normally be normalized states.

To get a better understanding of these issues, it is helpful to first find the Hamiltonian in terms of $\widehat{a}$ and $\widehat{a}^\dagger$. There are two equivalent forms,

$$H = E^{\mathrm{p}}\widehat{a}\widehat{a}^\dagger - \tfrac{1}{2}E^{\mathrm{p}} + E_{\mathrm{ref}} \qquad H = E^{\mathrm{p}}\widehat{a}^\dagger\widehat{a} + \tfrac{1}{2}E^{\mathrm{p}} + E_{\mathrm{ref}} \tag{5}$$

These expressions can be verified by plugging in the definitions (4) of $\widehat{a}$ and $\widehat{a}^\dagger$ and using the commutator (2). (Note that subtracting the two expressions gives the commutator of $\widehat{a}$ and $\widehat{a}^\dagger$ to be 1.)

Now look at the first Hamiltonian first. If $\widehat{a}^\dagger\psi_E$ would be zero for some state ψ_E, then that state would have energy $E_{\mathrm{ref}} - \frac{1}{2}E^{\mathrm{p}}$. But that is not possible. If you look at the original Hamiltonian (1), the energy must at least be E_{ref}; square Hermitian operators are nonnegative.

(To be more precise, if you square a Hermitian operator, you square the eigenvalues, making them nonnegative. It is said that the square operator is "positive definite," or, if there are zero eigenvalues, positive semi-definite. And such an operator produces nonnegative expectation values. And the expectation values of the operators in the Hamiltonian do add up to the total energy; just take an inner product of the Hamiltonian eigenvalue problem with the wave function. See chapter 4.4 for more information on expectation values.)

It follows that $\widehat{a}^\dagger\psi_E$ is never zero. This operator can be applied indefinitely to find states of higher and higher energy. So there is no maximum energy.

But there is a possibility that $\widehat{a}\psi_E$ is zero. As the second form of the Hamiltonian in (5) shows, that requires that the energy of state ψ_E equals

$$E_0 = \tfrac{1}{2}E^{\mathrm{p}} + E_{\mathrm{ref}}$$

Now if you start from *any* energy state ψ_E and apply $\widehat{a}$ sufficiently many times, you must eventually end up at this energy level. If not, you could go on lowering the energy forever. That would be inconsistent with the fact that the energy cannot be lower than E_{ref}. It follows that the above energy is the lowest energy that a state can have. So it is the ground state energy.

And any other energy must be a whole multiple of E^{p} higher than the ground state energy. Otherwise you could not end up at the ground state energy by applying $\widehat{a}$. Therefore, the energy eigenstates can be denoted more meaningfully by $|i\rangle$ rather than ψ_E. Here i is the number of quanta E^{p} that the energy is above the ground state level.

Now assume that the ground state is unique. In that case, there is one unique energy eigenfunction at each energy level. That is a consequence of the fact that if you go down a unit in energy with $\widehat{a}$ and then up a unit again with $\widehat{a}^\dagger$, (or vice versa), you must end up not just at the same energy, but at the same state. Otherwise the state would not be an eigenfunction of the Hamiltonian in one of the forms given in (5). Repeated application shows that if you go down any number of steps, and then up the same number of steps, you end up at the same state. Since every state must end up at the unique ground state, every state must be the result of applying $\widehat{a}^\dagger$ to the ground state sufficiently many times. There is just one such state for each energy level.

If there are two independent ground states, applying $\widehat{a}^\dagger$ on each gives two separate sets of energy eigenstates. And similar if there are still more ground states. Additional symbols will need to be added to the kets to keep the different families apart.

It was already mentioned that the states produced by the operators $\widehat{a}$ and $\widehat{a}^\dagger$ are usually not normalized. For example, the state $\widehat{a}|i\rangle$ will have a square magnitude given by the inner product

$$|\widehat{a}|i\rangle|^2 = \Big\langle \widehat{a}|i\rangle \Big| \widehat{a}|i\rangle \Big\rangle$$

Now if you take $\widehat{a}$ or $\widehat{a}^\dagger$ to the other side of an inner product, it will change into the other one; the i in the definitions (4) will change sign. So the square magnitude of $\widehat{a}|i\rangle$ becomes

$$|\widehat{a}|i\rangle|^2 = \Big\langle |i\rangle \Big| \widehat{a}^\dagger\widehat{a}|i\rangle \Big\rangle$$

From the second form of the Hamiltonian in (5), it is seen that $\widehat{a}^\dagger\widehat{a}$ gives the number of energy quanta i. And since the state $|i\rangle$ is normalized, the square magnitude of $\widehat{a}|i\rangle$ is therefore i. That means that

$$\widehat{a}|i\rangle = c\sqrt{i}|i{-}1\rangle$$

where c is some number of magnitude 1. Similarly

$$\widehat{a}^\dagger|i\rangle = d\sqrt{i+1}|i{-}1\rangle$$

But note that you can always change the definition of an energy eigenfunction by a constant of magnitude 1. That allows you, while going up from $|0\rangle$ using $\widehat{a}^\dagger$, to redefine each state so that d is 1. And if d is always one, then so is c. Otherwise $\widehat{a}^\dagger\widehat{a}$ would not be i.

In the ground state, the expectation values of P and Q are zero, while the expectation values of P^2 and Q^2 are equal to the minimum $\frac{1}{4}$ allowed by the uncertainty relation (3). The derivations of these statements are the same as those for the harmonic oscillator ground state in {D.13}.

D.34 The adiabatic theorem

Consider the Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = H\Psi$$

If the Hamiltonian is independent of time, the solution can be written in terms of the Hamiltonian energy eigenvalues $E_{\vec{n}}$ and eigenfunctions $\psi_{\vec{n}}$ as

$$\Psi = \sum_{\vec{n}} c_{\vec{n}}(0)e^{\mathrm{i}\theta_{\vec{n}}}\psi_{\vec{n}} \qquad \theta_{\vec{n}} = -\frac{1}{\hbar}E_{\vec{n}}t$$

Here $\vec{n}$ stands for the quantum numbers of the eigenfunctions and the $c_{\vec{n}}(0)$ are arbitrary constants.

However, the Hamiltonian varies with time for the systems of interest here. Still, at any given time its eigenfunctions form a complete set. So it is still possible to write the wave function as a sum of them, say like

$$\Psi = \sum_{\vec{n}} \bar{c}_{\vec{n}} e^{\mathrm{i}\theta_{\vec{n}}} \psi_{\vec{n}} \qquad \theta_{\vec{n}} = -\frac{1}{\hbar} \int E_{\vec{n}} \,\mathrm{d}t \tag{D.18}$$

But the coefficients $\bar{c}_{\vec{n}}$ can no longer be assumed to be constant like the $c_{\vec{n}}(0)$. They may be different at different times.

To get an equation for their variation, plug the expression for Ψ in the Schrödinger equation. That gives:

$$\mathrm{i}\hbar \sum_{\vec{n}} \bar{c}'_{\vec{n}} e^{\mathrm{i}\theta_{\vec{n}}} \psi_{\vec{n}} - \mathrm{i}\hbar \sum_{\vec{n}} \bar{c}_{\vec{n}} \frac{\mathrm{i}}{\hbar} E_{\vec{n}} e^{\mathrm{i}\theta_{\vec{n}}} \psi_{\vec{n}} + \mathrm{i}\hbar \sum_{\vec{n}} \bar{c}_{\vec{n}} e^{\mathrm{i}\theta_{\vec{n}}} \psi'_{\vec{n}} = H \sum_{\vec{n}} \bar{c}_{\vec{n}} e^{\mathrm{i}\theta_{\vec{n}}} \psi_{\vec{n}}$$

where the primes indicate time derivatives. The middle sum in the left hand side and the right hand side cancel against each other since by definition $\psi_{\vec{n}}$ is an eigenfunction of the Hamiltonian with eigenvalue $E_{\vec{n}}$. For the remaining two sums, take an inner product with an arbitrary eigenfunction $\langle \psi_{\underline{\vec{n}}} |$:

$$\mathrm{i}\hbar \bar{c}'_{\underline{\vec{n}}} e^{\mathrm{i}\theta_{\underline{\vec{n}}}} + \mathrm{i}\hbar \sum_{\vec{n}} \bar{c}_{\vec{n}} e^{\mathrm{i}\theta_{\vec{n}}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle = 0$$

In the first sum only the term $\vec{n} = \underline{\vec{n}}$ survived because of the orthonormality of the eigenfunctions. Divide by $\mathrm{i}\hbar e^{\mathrm{i}\theta_{\underline{\vec{n}}}}$ and rearrange to get

$$\bar{c}'_{\underline{\vec{n}}} = -\sum_{\vec{n}} e^{\mathrm{i}(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \bar{c}_{\vec{n}} \tag{D.19}$$

This is still exact.

However, the purpose of the current derivation is to address the adiabatic approximation. The adiabatic approximation assumes that the entire evolution takes place very slowly over a large time interval T. For such an evolution, it helps to consider all quantities to be functions of the scaled time variable t/T. Variables change by a finite amount when t changes by a finite fraction of T, so when t/T changes by a finite amount. This implies that the time derivatives of the slowly varying quantities are normally small, of order $1/T$.

Consider now first the case that there is no degeneracy, in other words, that there is only one eigenfunction $\psi_{\vec{n}}$ for each energy $E_{\vec{n}}$. If the Hamiltonian changes slowly and regularly in time, then so do the energy eigenvalues and eigenfunctions. In particular, the time derivatives of the eigenfunctions in (D.19) are small of order $1/T$. It then follows from the entire equation that the time derivatives of the coefficients are small of order $1/T$ too.

(Recall that the square magnitudes of the coefficients give the probability for the corresponding energy. So the magnitude of the coefficients is bounded by 1. Also, for simplicity it will be assumed that the number of eigenfunctions in the system is finite. Otherwise the sums over $\vec{n}$ might explode. This book routinely assumes that it is "good enough" to approximate an infinite system by a large-enough finite one. That makes life a lot easier, not just here but also in other derivations like {D.18}.)

It is convenient to split up the sum in (D.19):

$$\bar{c}'_{\underline{\vec{n}}} = -\langle \psi_{\underline{\vec{n}}} | \psi'_{\underline{\vec{n}}} \rangle \bar{c}_{\underline{\vec{n}}} - \sum_{\vec{n} \neq \underline{\vec{n}}} e^{\mathrm{i}(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \bar{c}_{\vec{n}} \tag{D.20}$$

Under the stated conditions, the final sum can be ignored.

However, that is not because it is small due to the time derivative in it, as one reference claims. While the time derivative of $\psi_{\vec{n}}$ is indeed small of order $1/T$, it acts over a time that is large of order T. The sum can be ignored because of the exponential in it. As the definition of $\theta_{\vec{n}}$ shows, it varies on the normal time scale, rather than on the long time scale T. Therefore it oscillates many times on the long time scale; that causes opposite values of the exponential to largely cancel each other.

To show that more precisely, note that the formal solution of the full equation (D.20) is, [41, 19.2]:

$$\bar{c}_{\underline{\vec{n}}}(t) = e^{\mathrm{i}\gamma_{\underline{\vec{n}}}} \left[\bar{c}_{\underline{\vec{n}}}(0) - \sum_{\vec{n} \neq \underline{\vec{n}}} \int_{\bar{t}=0}^{t} e^{\mathrm{i}(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} e^{-\mathrm{i}\gamma_{\underline{\vec{n}}}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \bar{c}_{\vec{n}} \, \mathrm{d}\bar{t} \right] \qquad \gamma'_{\underline{\vec{n}}} = \mathrm{i} \langle \psi_{\underline{\vec{n}}} | \psi'_{\underline{\vec{n}}} \rangle \tag{D.21}$$

To check this solution, you can just plug it in. Note in doing so that the integrands are taken to be functions of $\bar{t}$, not t.

All the integrals are negligibly small because of the rapid variation of the first exponential in them. To verify that, rewrite them a bit and then perform an integration by parts:

$$\int_{\bar{t}=0}^{t} -\frac{\mathrm{i}}{\hbar}(E_{\vec{n}} - E_{\underline{\vec{n}}}) e^{\mathrm{i}(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \quad \frac{\hbar e^{-\mathrm{i}\gamma_{\underline{\vec{n}}}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \bar{c}_{\vec{n}}}{\mathrm{i}(E_{\vec{n}} - E_{\underline{\vec{n}}})} \, \mathrm{d}\bar{t} =$$

$$e^{\mathrm{i}(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \frac{\hbar e^{-\mathrm{i}\gamma_{\underline{\vec{n}}}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \bar{c}_{\vec{n}}}{\mathrm{i}(E_{\vec{n}} - E_{\underline{\vec{n}}})} \Bigg|_{\bar{t}=0}^{t} - \int_{\bar{t}=0}^{t} e^{\mathrm{i}(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \left(\frac{\hbar e^{-\mathrm{i}\gamma_{\underline{\vec{n}}}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \bar{c}_{\vec{n}}}{\mathrm{i}(E_{\vec{n}} - E_{\underline{\vec{n}}})} \right)' \mathrm{d}\bar{t}$$

The first term in the right hand side is small of order $1/T$ because the time derivative of the wave function is. The integrand in the second term is small of order $1/T^2$ because of the two time derivatives. So integrated over an order T time range, it is small of order $1/T$ like the first term. It follows that the integrals in (D.21) become zero in the limit $T \to \infty$.

And that means that in the adiabatic approximation

$$\bar{c}_{\vec{n}} = c_{\vec{n}}(0)e^{\mathrm{i}\gamma_{\vec{n}}} \qquad \gamma_{\vec{n}} = \mathrm{i}\int \langle \psi_{\vec{n}} | \psi'_{\vec{n}} \rangle \, \mathrm{d}t$$

The underbar used to keep $\underline{\vec{n}}$ and $\vec{n}$ apart is no longer needed here since only one set of quantum numbers appears. This expression for the coefficients can be plugged in (D.18) to find the wave function Ψ. The constants $c_{\vec{n}}(0)$ depend on the initial condition for Ψ. (They also depend on the choice of integration constants for $\theta_{\vec{n}}$ and $\gamma_{\vec{n}}$, but normally you take the phases zero at the initial time).

Note that $\gamma_{\vec{n}}$ is real. To verify that, differentiate the normalization requirement to get

$$\langle \psi_{\vec{n}} | \psi_{\vec{n}} \rangle = 1 \quad \Longrightarrow \quad \langle \psi'_{\vec{n}} | \psi_{\vec{n}} \rangle + \langle \psi_{\vec{n}} | \psi'_{\vec{n}} \rangle = 0$$

So the sum of the inner product plus its complex conjugate are zero. That makes it purely imaginary, so $\gamma_{\vec{n}}$ is real.

Since both $\gamma_{\vec{n}}$ and $\theta_{\vec{n}}$ are real, it follows that the magnitudes of the coefficients of the eigenfunctions do not change in time. In particular, if the system starts out in a single eigenfunction, then it stays in that eigenfunction.

So far it has been assumed that there is no degeneracy, at least not for the considered state. However it is no problem if at a finite number of times, the energy of the considered state crosses some other energy. For example, consider a three-dimensional harmonic oscillator with three time varying spring stiffnesses. Whenever any two stiffnesses become equal, there is significant degeneracy. Despite that, the given adiabatic solution still applies. (This does assume that you have chosen the eigenfunctions to change smoothly through degeneracy, as perturbation theory says you can, {D.79}.)

To verify that the solution is indeed still valid, cut out a time interval of size δT around each crossing time. Here δ is some number still to be chosen. The parts of the integrals in (D.21) outside of these intervals have magnitudes $\varepsilon(T, \delta)$ that become zero when $T \to \infty$ for the same reasons as before. The parts of the integrals corresponding to the intervals can be estimated as no more than some finite multiple of δ. The reason is that the integrands are of order $1/T$ and they are integrated over ranges of size δT. All together, that is enough to show that the complete integrals are less than say 1%; just take δ small enough that the intervals contribute no more than 0.5% and then take T large enough that the remaining integration range contributes no more than 0.5% too. Since you can play the same game for 0.1%, 0.01% or any arbitrarily small amount, the conclusion is that for infinite T, the contribution of the integrals becomes zero. So in the limit $T \to \infty$, the adiabatic solution applies.

Things change if some energy levels are permanently degenerate. Consider an harmonic oscillator for which at least two spring stiffnesses are permanently equal. In that case, you need to solve for all coefficients at a given energy

level $E_{\underline{\vec{n}}}$ together. To figure out how to do that, you will need to consult a book on mathematics that covers systems of ordinary differential equations. In particular, the coefficient $\bar{c}_{\underline{\vec{n}}}$ in (D.21) gets replaced by a vector of coefficients with the same energy. The scalar $\gamma_{\underline{\vec{n}}}$ becomes a matrix with indices ranging over the set of coefficients in the vector. Also, $e^{\mathrm{i}\gamma_{\underline{\vec{n}}}}$ gets replaced by a "fundamental solution matrix," a matrix consisting of independent solution vectors. And $e^{-\mathrm{i}\gamma_{\underline{\vec{n}}}}$ is the inverse matrix. The sum no longer includes any of the coefficients of the considered energy.

More recent derivations allow the spectrum to be continuous, in which case the nonzero energy gaps $E_{\underline{\vec{n}}}-E_{\vec{n}}$ can no longer be assumed to be larger than some nonzero amount. And unfortunately, assuming the system to be approximated by a finite one helps only partially here; an accurate approximation will produce very closely spaced energies. Such problems are well outside the scope of this book.

D.35 The evolution of expectation values

To verify the stated formulae for the evolution of expectation values, just write the definition of expectation value, $\langle\Psi|A\Psi\rangle$, differentiate to get

$$\langle\Psi_t|A\Psi\rangle + \langle\Psi|A\Psi_t\rangle + \langle\Psi|A_t\Psi\rangle$$

and replace Ψ_t by $H\Psi/\mathrm{i}\hbar$ on account of the Schrödinger equation. Note that in the first inner product, the i appears in the left part, hence comes out as its complex conjugate $-\mathrm{i}$.

D.36 Photon wave function derivations

The rules of engagement are as follows:

- The Cartesian axes are numbered using an index i, with $i = 1$, 2, and 3 for x, y, and z respectively.
- Also, r_i indicates the coordinate in the i direction, x, y, or z.
- Derivatives with respect to a coordinate r_i are indicated by a simple subscript i.
- If the quantity being differentiated is a vector, a comma is used to separate the vector index from differentiation ones.
- Index $\bar{\imath}$ is the number immediately following i in the cyclic sequence $\ldots 123123\ldots$ and $\bar{\bar{\imath}}$ is the number immediately preceding i.
- Time derivatives are indicated by a subscript t.
- A bare $\int$ integral sign is assumed to be an integration over all space, or over the entire box for particles in a box. The $\mathrm{d}^3\vec{r}$ is normally omitted for brevity and to be understood.
- A superscript * indicates a complex conjugate.

D.36.1 Rewriting the energy integral

As given in the text, the energy in an electromagnetic field in free space that satisfies the Coulomb-Lorenz gauge is, writing out the square magnitudes and individual components,

$$E = \tfrac{1}{2}\epsilon_0 \int \left(\left| \frac{\partial \vec{A}}{\partial t} \right|^2 + c^2 \left| \nabla \vec{A} \right|^2 \right) = \tfrac{1}{2}\epsilon_0 \sum_i \int \left(A^*_{i,t} A_{i,t} + c^2 \sum_{j=1}^{3} A^*_{i,j} A_{i,j} \right)$$

However, a bit more general expression is desirable. If only the Lorenz condition is satisfied, there may also be an electrostatic potential φ. In that case, a more general expression for the energy is:

$$E = \tfrac{1}{2}\epsilon_0 \Big[\sum_{i=1}^{3} \int \left(A^*_{i,t} A_{i,t} + c^2 \sum_{j=1}^{3} A^*_{i,j} A_{i,j} \right) - \int \left(\frac{1}{c^2} \varphi^*_t \varphi_t + \sum_{j=1}^{3} \varphi^*_j \varphi_j \right) \Big] \quad (1)$$

The minus sign for the φ terms appears because this is really a dot product of relativistic four-vectors. The zeroth components in such a dot product acquire a minus sign, chapter 1.2.4 and 1.3.2. In derivation {D.32} it was shown that each of the four integrals is constant. That is because each component satisfies the Klein-Gordon equation. So their sum is constant too.

The claim to verify now is that the same energy can be obtained from integrating the electric and magnetic fields as

$$E = \tfrac{1}{2}\epsilon_0 \int \left(|\vec{\mathcal{E}}|^2 + c^2 |\vec{\mathcal{B}}|^2 \right) = \tfrac{1}{2}\epsilon_0 \sum_i \int \left(\mathcal{E}^*_i \mathcal{E}_i + c^2 \mathcal{B}^*_i \mathcal{B}_i \right) \quad (2)$$

Since $\vec{\mathcal{E}} = -\partial \vec{A}/\partial t - \nabla \varphi$ and $\vec{\mathcal{B}} = \nabla \times \vec{A}$:

$$\mathcal{E}_i = -A_{i,t} - \varphi_i \qquad \mathcal{B}_i = A_{\bar{i},\bar{i}} - A_{\bar{i},\bar{i}}$$

From now on, it will be understood that there is a summation over i and j and that everything has a $\frac{1}{2}\epsilon_0$. Therefore these will no longer be shown.

Start with the electric field integral. It is, using the above expressions and multiplying out,

$$\int A^*_{i,t} A_{i,t} + A^*_{i,t} \varphi_i + \varphi^*_i A_{i,t} + \varphi^*_i \varphi_i$$

The first term already gives the vector-potential time derivatives in (1). That leaves the final three terms. Perform an integration by parts on the first two. It will always be assumed that the potentials vanish at infinity or that the system is in a periodic box. In that case there are no boundary terms in an integration by parts. So the three terms become

$$\int -A^*_{i,it} \varphi - \varphi^* A_{i,it} + \varphi^*_i \varphi_i$$

However, the divergence $A_{i,i}$ is according to the Lorenz condition equal to $-\varphi_t/c^2$, so

$$\int \frac{1}{c^2}\varphi^*_{tt}\varphi + \frac{1}{c^2}\varphi^*\varphi_{tt} + \varphi^*_i\varphi_i$$

Using the Klein-Gordon equation, $\varphi_{tt}/c^2 = \varphi_{ii}$, and then another integration by parts on the first two terms and renotating i by j gives the $\varphi^*_j\varphi_j$ terms in (1).

Now consider the integral of $|\vec{\mathcal{B}}|^2$ in (2). You get, multiplying out,

$$c^2 \int \left(A^*_{\bar{i},i} - A^*_{i,\bar{i}}\right)\left(A_{\bar{i},i} - A_{i,\bar{i}}\right) = c^2 \int \left(A^*_{\bar{i},i}A_{\bar{i},i} - A^*_{\bar{i},i}A_{i,\bar{i}} - A^*_{i,\bar{i}}A_{\bar{i},i} + A^*_{i,\bar{i}}A_{i,\bar{i}}\right)$$

Now the first and last terms in the right hand side summed over i produce all terms $A^*_{i,j}A_{i,j}$ in (1) in which i and j are different. That leaves the middle terms. An integration by parts yields

$$c^2 \int \left(A^*_{\bar{i},\bar{i}i}A_{\bar{i}} + A^*_{\bar{i},i\bar{i}}A_{\bar{i}}\right)$$

Renotate the indices cyclically to get

$$c^2 \int \left(A^*_{\bar{i},\bar{i}i}A_i + A^*_{\bar{i},\bar{i}i}A_i\right)$$

(If you want, you can check that this is the same by writing out all three terms in the sum.) This is equivalent to

$$c^2 \int \left(A^*_{i,i} + A^*_{\bar{i},\bar{i}} + A^*_{\bar{i},\bar{i}}\right)_i A_i - A^*_{i,ii}A_i$$

as you can see from differentiating and multiplying out. The final term gives after integration by parts the $A^*_{i,j}A_{i,j}$ terms in (1) in which i and j are equal. That leaves the first part. The term in parentheses is the divergence $-\varphi_t/c^2$, so the first part is

$$\int -\varphi^*_{it}A_i$$

Perform an integration by parts

$$\int \varphi^*_t A_{i,i}$$

Recognizing once more the divergence, this gives the final $-\varphi^*_t\varphi_t/c^2$ term in (1)

D.36.2 Angular momentum states

The rules of engagement listed at the start of this section apply. In addition:

- The quantum numbers ℓ and m_ℓ will be renotated by l and m, while k stays k. That is easier to type.
- The quantum numbers are not shown unless needed. For example, j stands for j_l.
- A bar on a quantity, like in $\bar{Y}Y$, means the complex conjugate. In addition, the (unlisted) quantum numbers of spherical harmonic $\bar{Y}$ may in general be different from those of Y and are indicated by bars too.
- The symbols f and g are used as generic scalar functions. They often stand in particular for the jY scalar modes.
- An integral in spherical coordinates takes the form $\int \dots r^2 \,\mathrm{d}r\mathrm{d}\Omega$ where $\mathrm{d}\Omega = \sin\theta \,\mathrm{d}\theta\mathrm{d}\phi$.

D.36.2.1 About the scalar modes

The scalar modes are the jY.

It will be assumed that the j are zero at the large radius $r_{\max}$ at which the domain is assumed to terminate. That makes the scalar modes a complete set; *any* scalar function f can be written as a combination of them. (That is because they are the eigenfunctions of the Laplacian inside the sphere, and the zero boundary condition on the sphere surface $r = r_{\max}$ makes the Laplacian Hermitian. This will not be explicitly proved since it is very standard.)

The Bessel function j of the scalar modes satisfy the ordinary differential equation, {A.6}

$$r^2 j'' + 2rj' = l(l+1)j - k^2 r^2 j \tag{3}$$

The following integral is needed (note that j is real):

$$\int_0^{r_{\max}} j^2 r^2 \,\mathrm{d}r \sim \frac{r_{\max}}{2k^2} \tag{4}$$

This is valid for large $kr_{\max}$, which applies since $r_{\max}$ is large and the k values of interest are finite. The above result comes from the integral of the square two-dimensional Bessel functions J, and a recurrence relation, [41, 27.18,88], using $j_l(kr) = J_{l+\frac{1}{2}}(kr)\sqrt{\pi/2kr}$, [1, p 437, 10.1.1], and the asymptotic behavior of the Bessel function you get from {A.6} (A.19). To get the leading asymptotic term, each time you have to differentiate the trigonometric function. And where the trigonometric function in j_l is zero at $r_{\max}$ because of the boundary condition, the one in j_{l+1} has magnitude 1.

The spherical harmonics are orthonormal on the unit sphere, {D.14.4}

$$\int \bar{Y}Y \,\mathrm{d}\Omega = \delta_{\bar{l}l}\delta_{\bar{m}m} \tag{5}$$

In other words, the integral is only 1 if $l = \bar{l}$ and $m = \bar{m}$ and otherwise it is zero. Further

$$\int \left(\bar{Y}_\theta Y_\theta + \frac{1}{\sin^2\theta} \bar{Y}_\phi Y_\phi \right) \mathrm{d}\Omega = l(l+1)\delta_{\bar{l}l}\delta_{\bar{m}m} \tag{6}$$

D.36.2.2 Basic observations and eigenvalue problem

For any function f

$$\vec{r} \times \nabla f = -\nabla \times (\vec{r} f) \tag{7}$$

This follows from writing out the right hand side

$$-(r_{\bar{\imath}} f)_{\bar{\imath}} + (r_{\bar{\imath}} f)_{\bar{\bar{\imath}}} = -r_{\bar{\imath}} f_{\bar{\imath}} + r_{\bar{\imath}} f_{\bar{\bar{\imath}}}$$

the latter since $\bar{\imath}$ and $\bar{\bar{\imath}}$ are different indices.

The electric modes

$$\nabla \times \vec{r} \times \nabla f$$

are solenoidal because $\nabla \cdot \nabla \times \ldots$ gives zero. The magnetic modes

$$\vec{r} \times \nabla f$$

are solenoidal for the same reason, after noting (7) above.

The Laplacian commutes with the operators in front of the scalar functions in the electric and magnetic modes. That can be seen for the magnetic ones from

$$(r_{\bar{\imath}} f_{\bar{\bar{\imath}}} - r_{\bar{\bar{\imath}}} f_{\bar{\imath}})_{jj} = r_{\bar{\imath}} f_{jj\bar{\bar{\imath}}} - r_{\bar{\bar{\imath}}} f_{jj\bar{\imath}} + 2r_{\bar{\imath},j} f_{j\bar{\bar{\imath}}} - 2r_{\bar{\bar{\imath}},j} f_{j\bar{\imath}} = r_{\bar{\imath}} f_{jj\bar{\bar{\imath}}} - r_{\bar{\bar{\imath}}} f_{jj\bar{\imath}} + 2f_{\bar{\bar{u}}} - 2f_{\bar{\bar{u}}}$$

and the final two terms cancel. And the Laplacian also commutes with the additional $\nabla\times$ in the electric modes since differentiations commute.

From this it follows that the energy eigenvalue problem is satisfied because by definition of the scalar modes $-\nabla^2 jY = k^2 jY$. In addition,

$$\nabla \times \nabla \times \vec{r} \times \nabla f = k^2 \vec{r} \times \nabla f \tag{8}$$

because $\nabla \times \nabla\times = -\nabla^2$ for a solenoidal function, (D.1).

D.36.2.3 Spherical form and net angular momentum

In spherical coordinates, the magnetic mode is

$$\vec{A}^{\mathrm{M}} = \vec{r} \times \nabla jY = j \left[\hat{\imath}_\phi Y_\theta - \hat{\imath}_\theta \frac{1}{\sin\theta} Y_\phi \right] \tag{9}$$

and then the electric mode is

$$\vec{A}^{\rm E} = \nabla \times \vec{r} \times \nabla jY = -\hat{\imath}_r l(l+1)\frac{j}{r}Y - \frac{(rj)'}{r}\left[\hat{\imath}_\theta Y_\theta + \hat{\imath}_\phi \frac{1}{\sin\theta} Y_\phi\right] \tag{10}$$

from [41, 20.74,76,82] and for the r component of $\vec{A}^{\rm E}$ the eigenvalue problem of chapter 4.2.3.

Now note that the ϕ dependence of Y is through a simple factor $e^{{\rm i}m\phi}$, chapter 4.2.3. Therefore it is seen that if the coordinate system is rotated over an angle γ around the z-axis, it produces a factor $e^{{\rm i}m\gamma}$ in the vectors. First of all that means that the azimuthal quantum number of net angular momentum is m, {A.19}. But it also means that, {A.19},

$$\widehat{J}_z \vec{r} \times \nabla f = \vec{r} \times \nabla L_z f \qquad \widehat{J}_z \nabla \times \vec{r} \times \nabla f = \nabla \times \vec{r} \times \nabla L_z f$$

because either way the vector gets multiplied by $m\hbar$ for the modes. And if it is true for all the modes, then it is true for any function f. Since the z-axis is not special for general f, the same must hold for the x and y angular momentum operators. From that it follows that the modes are also eigenfunctions of net square angular momentum, with azimuthal quantum number l.

At the cut-off $r = r_{\max}$, $j = 0$, which gives:

$$\text{At } r_{\max}\text{:} \qquad \vec{A}^{\rm M} = 0 \qquad \vec{A}^{\rm E} = -j'\left[\hat{\imath}_\theta Y_\theta + \hat{\imath}_\phi \frac{1}{\sin\theta} Y_\phi\right] \tag{11}$$

Also needed is, differentiating (10):

$$\text{At } r_{\max}\text{:} \qquad \frac{\partial \vec{A}^{\rm E}}{\partial r} = -\hat{\imath}_r l(l+1)\frac{j'}{r}Y + \frac{j'}{r}\left[\hat{\imath}_\theta Y_\theta + \hat{\imath}_\phi \frac{1}{\sin\theta} Y_\phi\right] \tag{12}$$

which used (3) to get rid of the second order derivative of j.

D.36.2.4 Orthogonality and normalization

Whether the modes are orthogonal, and whether the Laplacian is Hermitian, is not obvious because of the weird boundary conditions at $r_{\max}$.

In general the important relations here

$$\begin{aligned}
&\int (\bar{k}^2 - k^2) A_i A_i \\
&\quad = \int \bar{A}_{i,jj} A_i - \bar{A}_i A_{i,jj} \\
&\quad = \int (\bar{A}_{i,j} A_i - \bar{A}_i A_{i,j})_j \\
&\quad = \int_S (\bar{A}_{i,j} A_i - \bar{A}_i A_{i,j}) \frac{\partial r_j}{\partial r} \, {\rm d}S
\end{aligned} \tag{13}$$

where S is the surface of the sphere $r = r_{\max}$. The second last line can be verified by differentiating out and the last line is the divergence theorem.

The first and second line in (13) show that the Laplacian is Hermitian if all unequal modes are orthogonal (or have equal k values, but orthogonality should be shown anyway.). For unequal k values orthogonality may be shown by showing that the final surface integral is zero.

It is convenient to show right away that the electric and magnetic modes are always mutually orthogonal:

$$\int (r_{\bar{\imath}} \bar{f}_{\bar{\bar{\imath}}} - r_{\bar{\bar{\imath}}} \bar{f}_{\bar{\imath}}) A_i^{\mathrm{E}} = \int (r_{\bar{\imath}} \bar{f} A_i^{\mathrm{E}})_{\bar{\bar{\imath}}} - (r_{\bar{\bar{\imath}}} \bar{f} A_i^{\mathrm{E}})_{\bar{\imath}} - r_{\bar{\imath}} \bar{f} A_{i,\bar{\bar{\imath}}}^{\mathrm{E}} + r_{\bar{\bar{\imath}}} \bar{f} A_{i,\bar{\imath}}^{\mathrm{E}}$$

The first two terms in the right hand side can be integrated in the $\bar{\bar{\imath}}$, respectively $\bar{\imath}$ direction and are then zero because $\bar{f}$ is zero on the spherical surface $r = r_{\max}$. The final two terms summed over i can be renotated by shifting the summation index one unit down, respectively up in the cyclic sequence to give

$$\sum_i - \int \bar{f} r_i (A_{\bar{\bar{\imath}},\bar{\imath}}^{\mathrm{E}} - A_{\bar{\imath},\bar{\bar{\imath}}}^{\mathrm{E}}) = \sum_i - \int \bar{f} \vec{r} \cdot \nabla \times \vec{A}^{\mathrm{E}} = -k^2 \int \bar{f} \vec{r} \cdot \vec{r} \times \nabla f$$

the latter because of the form of $\vec{A}^{\mathrm{E}}$, the fact that $\nabla \times \nabla \times = -\nabla^2$ for a solenoidal vector, and the energy eigenvalue problem established for $\vec{A}^{\mathrm{M}}$. The final term is zero because $\vec{r} \cdot \vec{r} \times$ is.

Next consider the orthogonality of the magnetic modes for different quantum numbers. For $\bar{l} \neq l$ or $\bar{m} \neq m$, the orthogonality follows from (9) and (6). For $\bar{k} \neq k$, the orthogonality follows from the final line in (13) since the magnetic modes are zero at $r_{\max}$, (11).

Finally the electric modes. For $\bar{l} \neq l$ or $\bar{m} \neq m$, the orthogonality follows from (10), (5), and (6). For $\bar{k} \neq k$, the orthogonality follows from the final line in (13). To see that, recognize that $A_{i,j} \partial r_j / \partial r$ is the radial derivative of $\vec{A}$; therefore using (11) and (12), the integrand vanishes.

The integral of the absolute square integral of a magnetic mode is, using (9), (6), and (4),

$$\int \vec{A}^{\mathrm{M}*} \cdot \vec{A}^{\mathrm{M}} = l(l+1) \frac{r_{\max}}{2k^2}$$

The integral of the absolute square integral of an electric mode is, using (10), (5), and (6),

$$l^2(l+1)^2 \int_0^{r_{\max}} j^2 \, \mathrm{d}r + l(l+1) \int_0^{r_{\max}} (rj)'(rj)' \, \mathrm{d}r$$

Apply an integration by parts on the second integral,

$$l^2(l+1)^2 \int_0^{r_{\max}} j^2 \, \mathrm{d}r - l(l+1) \int_0^{r_{\max}} jr(rj)'' \, \mathrm{d}r$$

and then use (3) to get

$$\int \vec{A}^{\mathrm{M}*} \cdot \vec{A}^{\mathrm{M}} = k^2 l(l+1) \frac{r_{\max}}{2k^2}$$

The normalizations given in the text follow.

D.36.2.5 Completeness

Because of the condition $\nabla \cdot \vec{A} = 0$, you would generally speaking expect two different types of modes described by scalar functions. The electric and magnetic modes seem to fit that bill. But that does not mean that there could not be say a few more special modes. What is needed is to show completeness. That means to show that any smooth vector field satisfying $\nabla \cdot \vec{A} = 0$ can be written as a sum of the electric and magnetic modes, and nothing else.

This author does not know any simple way to do that. It would be automatic without the solenoidal condition; you would just take each Cartesian component to be a combination of the scalar modes jY satisfying a zero boundary condition at $r_{\max}$. Then completeness would follow from the fact that they are eigenfunctions of the Hermitian Laplacian. Or from more rigorous arguments that you can find in mathematical books on partial differential equations. But how to do something similar here is not obvious, at least not to this author.

What will be done is show that any reasonable solenoidal vector can be written in the form

$$\vec{A} = \vec{r} \times \nabla f + \nabla \times \vec{r} \times \nabla g$$

where f and g are scalar functions. Completeness then follows since the modes jY provide a complete description of any arbitrary function f and g.

But to show the above does not seem easy either, so what will be actually shown is that any vector without radial component can be written in the form

$$\vec{v} = \vec{r} \times \nabla f + \hat{\imath}_r \times \vec{r} \times \nabla g$$

That is sufficient because the Fourier transform of $\vec{A}$ does not have a radial component, so it will be of this form. And the inverse Fourier transform of $\vec{v}$ is of the form $\vec{A}$, compare any book on Fourier transforms and (7).

The proof that $\vec{v}$ must be of the stated form is by construction. Note that automatically, the radial component of the two terms is zero. Writing out the gradients in spherical coordinates, [41, 20.74,82], multiplying out the cross products and equating components gives at any arbitrary radius r

$$-\frac{\partial f}{\partial \phi} - \sin\theta \frac{\partial g}{\partial \theta} = v_\theta \sin\theta \qquad \sin\theta \frac{\partial f}{\partial \theta} - \frac{\partial g}{\partial \phi} = v_\phi \sin\theta$$

Now decompose this in Fourier modes $e^{\mathrm{i}m\phi}$ in the ϕ direction:

$$-\mathrm{i}m f_m - \sin\theta \frac{\partial g_m}{\partial \theta} = v_{\theta m} \sin\theta \qquad \sin\theta \frac{\partial f_m}{\partial \theta} - \mathrm{i}m g_m = v_{\phi m} \sin\theta$$

For $m = 0$, f_0 and g_0 follow by integration. Note that the integrands are periodic of period 2π and antisymmetric about the z-axis. That makes f and g periodic of period 2π too,

For $m \neq 0$ make a coordinate transform from θ to

$$t = \int \mathrm{d}\theta / \sin\theta = \ln \tan \tfrac{1}{2}\theta$$

Note that $-\infty < t < \infty$. If anybody is actually reading this, send me an email. The system becomes after cleaning up

$$-\mathrm{i}m f_m - \frac{\partial g_m}{\partial t} = v_{\theta m}/\cosh t \qquad \frac{\partial f_m}{\partial t} - \mathrm{i}m g_m = v_{\phi m}/\cosh t$$

It is now easiest to solve the above equations for each of the two right hand sides separately. Here the first right hand side will be done, the second right hand side goes similarly.

From the two equations it is seen that f_m must satisfy

$$\frac{\partial^2 f_m}{\partial t^2} - m^2 f_m = \frac{-2m\mathrm{i}v_{\theta m}}{e^t + e^{-t}}$$

and $\mathrm{i}m g_m$ must the derivative of f_m. The solution satisfying the required regularity at $\pm\infty$ is, [41, 19.8],

$$f_m = \int_t^\infty \mathrm{i}v_{\theta m} \frac{e^{-m(\tau-t)}}{e^\tau + e^{-\tau}}\,\mathrm{d}\tau + \int_{-\infty}^t \mathrm{i}v_{\theta m} \frac{e^{m(\tau-t)}}{e^\tau + e^{-\tau}}\,\mathrm{d}\tau$$

That finishes the construction, but you may wonder about potential non-exponential terms in the first integral at $-\infty$ and the second integral at ∞. Those would produce weak logarithmic singularities in the physical f and g. You could simply guess that the two right hand sides will combine so that these terms drop out. After all, there is nothing special about the chosen direction of the z-axis. If you choose a different axis, it will show no singularities at the old z-axis, and the solution is unique.

For more confidence, you can check the cancellation explicitly for the leading order, $m = 1$ terms. But there is a better way. If the right hand sides are zero within a nonzero angular distance $\Delta\theta_1$ from the z-axis, there are no singularities. And it is easy to split off a part of $\vec{v}$ that is zero within $\Delta\theta_1$ of the axis and then changes smoothly to the correct $\vec{v}$ in an angular range from $\Delta\theta_1$ to $\Delta\theta_2$ from the axis. The remainder of $\vec{v}$ can than be handled by using say the x-axis as the axis of the spherical coordinate system.

D.36.2.6 Density of states

The spherical Bessel function is for large arguments proportional to $\sin(kr)/r$ or $\cos(kr)/r$. Either way, the zeros are spaced $\Delta k\, r_{\max} = \pi$ apart. So there is one state $\Delta N = 1$ in an interval $\Delta E = \hbar \Delta k c = \hbar \pi c / r_{\max}$. The ratio gives the stated density of states.

D.36.2.7 Parity

Parity is what happens to the sign of the wave function under a parity transformation. A parity transformation inverts the positive direction of all three Cartesian axes, replacing any position vector $\vec{r}$ by $-\vec{r}$. The parity of something is 1 or even if it does not change, and -1 or odd if it changes sign. Under a parity transformation, the operators $\vec{r}\times$ and $\nabla\times$ flip over the parity of what they act on. On the other hand, $\nabla j_j Y_j^{m_j}$ has the same parity as $j_j Y_j^{m_j}$; the spatial components flip over, but so do the unit vectors that multiply them. And the parity of $j_j Y_j^{m_j}$ is even if j is even and odd if j is odd. The stated parities follow.

D.36.2.8 Orbital angular momentum of the states

In principle a state of definite net angular momentum j and definite spin 1 may involve orbital angular momentum $l = j-1$, $l = j$ and $l = j-1$, chapter 7.4.2. But states of definite parity restrict that to either only odd values or only even values, {A.20}. To get the stated parities, $l = j$ for magnetic states and $l = j-1$ or $l = j+1$ for electric ones.

woof.

D.37 Forces by particle exchange derivations

D.37.1 Classical energy minimization

The energy minimization including a selecton is essentially the same as the one for only spoton and foton field. That one has been discussed in chapter A.22.1 and in detail in {A.2}. So only the key differences will be listed here.

The energy to minimize is now

$$\frac{\epsilon_1}{2}\int (\nabla\varphi)^2 \,\mathrm{d}^3\vec{r} - \int \varphi(\vec{r})\Big(s_\mathrm{p}\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{p}) + s_2\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{e})\Big)\,\mathrm{d}^3\vec{r}$$

So the only real difference in the variational analysis is

$$s_\mathrm{p}\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{p}) \quad\rightarrow\quad s_\mathrm{p}\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{p}) + s_\mathrm{e}\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{p})$$

That means that the Poisson equation now becomes

$$-\nabla^2\varphi(\vec{r}) = \frac{s_\mathrm{p}}{\epsilon_1}\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{p}) + \frac{s_\mathrm{e}}{\epsilon_1}\delta_\varepsilon^3(\vec{r}-\vec{r}_\mathrm{e})$$

Since the Poisson equation is linear, the solution is $\varphi = \varphi^\mathrm{p} + \varphi^\mathrm{e}$. Here φ^p is the foton field (A.107) produced by the spoton as before, and φ^e is a similar expression, but using the selecton sarge and distance from the selecton:

$$\varphi^\mathrm{p} = \frac{s_\mathrm{p}}{4\pi\epsilon_1|\vec{r}-\vec{r}_\mathrm{p}|} \qquad \varphi^\mathrm{e} = \frac{s_\mathrm{e}}{4\pi\epsilon_1|\vec{r}-\vec{r}_\mathrm{e}|}$$

The energy lowering is now

$$-\tfrac{1}{2}\int \Big(\varphi^{\rm p}(\vec{r})+\varphi^{\rm e}(\vec{r})\Big)\Big(s_{\rm p}\delta^3_\varepsilon(\vec{r}-\vec{r}_{\rm p})+s_{\rm e}\delta^3_\varepsilon(\vec{r}-\vec{r}_{\rm e})\Big)\,{\rm d}^3\vec{r}$$

Multiplying out, you get, of course, the energy lowerings for the spoton and selecton in isolation. But you also get two additional interaction terms between these sarges. These two terms are equal; the selecton field $\varphi^{\rm e}$ evaluated at the position of the spoton times spoton sarge is the same as the spoton field $\varphi^{\rm p}$ at the selecton times selecton sarge. So it is seen that each term contributes half to the Koulomb energy as claimed in the text.

The foton field energy is still half of the particle-field interaction energies and of opposite sign. That is why the energy change is half of what you would expect from the interaction of the particles with each other's field: the other half is offset by changes in field energy.

D.37.2 Quantum energy minimization

This derivation includes the selecton in the spoton-fotons system analyzed in {A.22.3}. Since the analysis is essentially unchanged, only the key differences will be highlighted.

If an selecton is added to the system, the system wave function becomes

$$\psi_{\varphi{\rm pe}} = C_0\psi_{\rm p}\psi_{\rm e}|0\rangle + C_1\psi_{\rm p}\psi_{\rm e}|1\rangle + \ldots \qquad |C_0|^2+|C_1|^2+\ldots = 1$$

The demon can hold the selecton in its other hand. The Hamiltonian will now of course include a term for the selecton in isolation, as well as an interaction with the foton field. These are completely analogous to the corresponding spoton terms.

So the energy to be minimized for the ground state becomes

$$E = E_{\rm p}+E_{\rm e}+|C_1|^2\hbar\omega - 2\frac{\varepsilon_k}{\sqrt{2}k}\Big|s_{\rm p}\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}\psi_{\rm p}\rangle + s_{\rm e}\langle\psi_{\rm e}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm e}}\psi_{\rm e}\rangle\Big||C_1|\cos(\alpha+\beta)$$

If this is minimized as in {A.22.3}, the energy is

$$E = E_{\rm p}+E_{\rm e}-\frac{1}{2\epsilon_1\mathcal{V}k^2}|s_{\rm p}\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}|\psi_{\rm p}\rangle + s_{\rm e}\langle\psi_{\rm e}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm e}}|\psi_{\rm e}\rangle|^2$$

The square absolute value of a quantity can be found as the product of that quantity times its complex conjugate. That gives the same energy lowering as for the lone spoton, and a similar term for a lone selecton. However, there is an additional term

$$-\frac{s_{\rm p}s_{\rm e}}{2\epsilon_1\mathcal{V}k^2}\Big(\langle\psi_{\rm p}|e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}|\psi_{\rm p}\rangle\langle\psi_{\rm e}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm e}}|\psi_{\rm e}\rangle + \langle\psi_{\rm e}|e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm e}}|\psi_{\rm e}\rangle\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}|\psi_{\rm p}\rangle\Big)$$

If you write out the inner product integrals over the selecton coordinates explicitly, this becomes

$$-\int s_{\rm e}\psi_{\rm e}^*(\vec{r}_{\rm e})\frac{s_{\rm p}}{2\epsilon_1\mathcal{V}k^2}\left[\langle\psi_{\rm p}|e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}|\psi_{\rm p}\rangle e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm e}}+\langle\psi_{\rm p}|e^{{\rm i}\vec{k}\cdot\vec{r}_{\rm p}}|\psi_{\rm p}\rangle e^{-{\rm i}\vec{k}\cdot\vec{r}_{\rm e}}\right]\psi_{\rm e}(\vec{r}_{\rm e})\,{\rm d}^3\vec{r}_{\rm e}$$

Summed over all $\vec{k}$, the second term inside the square brackets gives the same answer as the first; that is because opposite $\vec{k}$ values appear equally in the summation. Looking at the first term, the summation over $\vec{k}$ produces again the spoton potential $\varphi_{\rm cl}^{\rm p}$, but now evaluated at the position of the selecton. That then shows the additional energy lowering to be

$$-\int s_{\rm e}\psi_{\rm e}^*(\vec{r}_{\rm e})\varphi_{\rm cl}^{\rm p}(\vec{r}_{\rm e})\psi_{\rm e}(\vec{r}_{\rm e})\,{\rm d}^3\vec{r}$$

Except for the differences in notation, that is the same selecton-spoton interaction energy as found in {A.22.1}.

D.37.3 Rewriting the Lagrangian

The rules of engagement are as follows:

- The Cartesian axes are numbered using an index i, with $i = 1$, 2, and 3 for x, y, and z respectively.
- Also, r_i indicates the coordinate in the i direction, x, y, or z.
- Derivatives with respect to a coordinate r_i are indicated by a simple subscript i.
- If the quantity being differentiated is a vector, a comma is used to separate the vector index from differentiation ones.
- Index $\bar{\imath}$ is the number immediately following i in the cyclic sequence $\ldots 123123\ldots$ and $\bar{\bar{\imath}}$ is the number immediately preceding i.
- If i is already been used for something else, j can be used the same way.
- Time derivatives are indicated by a subscript t.

Consider first the square magnetic field:

$$\mathcal{B}^2=\sum_i(A_{\bar{\bar{\imath}},\bar{\imath}}-A_{\bar{\imath},\bar{\bar{\imath}}})^2$$

Expanding out the square, that is equivalent to

$$\mathcal{B}^2=\sum_i(A_{\bar{\bar{\imath}},\bar{\imath}}^2+A_{\bar{\imath},\bar{\bar{\imath}}}^2+A_{i,i}^2-A_{i,i}A_{i,i}-A_{\bar{\imath},\bar{\bar{\imath}}}A_{\bar{\bar{\imath}},\bar{\imath}}-A_{\bar{\bar{\imath}},\bar{\imath}}A_{\bar{\imath},\bar{\bar{\imath}}})$$

The summation indices can now be cyclically redefined to give an equivalent sum over i equal to

$$\mathcal{B}^2=\sum_i(A_{i,\bar{\bar{\imath}}}^2+A_{i,\bar{\imath}}^2+A_{i,i}^2-A_{i,i}A_{i,i}-A_{i,\bar{\imath}}A_{\bar{\imath},i}-A_{i,\bar{\bar{\imath}}}A_{\bar{\bar{\imath}},i})$$

The terms can be combined in sets of three as

$$\mathcal{B}^2 = A_{i,j}^2 - A_{i,j}A_{j,i}$$

Here summation over i and j is now understood.

The square electric field is

$$\mathcal{E}^2 = (-\varphi_i - A_{i,t})^2 = A_{i,t}^2 + 2A_{i,t}\varphi_i + \varphi_i^2$$

All together, that gives

$$\begin{aligned}\mathcal{E}^2 - c^2\mathcal{B}^2 &= A_{i,t}^2 - c^2A_{i,j}^2 - \frac{1}{c^2}\varphi_t^2 + \varphi_i^2 \\ &\quad + \frac{1}{c^2}(\varphi_t + c^2A_{i,i})(\varphi_t + c^2A_{j,j}) \\ &\quad + 2A_{i,t}\varphi_i - 2A_{i,i}\varphi_t + c^2A_{i,j}A_{j,i} - c^2A_{i,i}A_{j,j}\end{aligned}$$

as can be verified by multiplying out and simplifying. The right hand side in the first line is the self-evident electromagnetic Lagrangian density, except for the factor $\epsilon_0/2$. The second line is the square of the Lorentz condition quantity. The final line can be written as a sum of pure derivatives:

$$2(A_i\varphi_i)_t - 2(A_i\varphi_t)_i + c^2(A_iA_{j,i})_j - c^2(A_iA_{j,j})_i$$

Pure derivatives do not produce changes in the action, as the changes in the potentials disappear on the boundaries of integration.

D.37.4 Coulomb potential energy

The Coulomb potential energy between charged particles is typically derived in basic physics. But it can also easily be verified from the conventional electromagnetic energy (A.143). In the steady case, there is only the electric field, due to the Coulomb potential. The energy may then be written as

$$\frac{\epsilon_0}{2}\int \mathcal{E}_{\rm C}^2\,\mathrm{d}^3\vec{r} = \frac{\epsilon_0}{2}\int (\nabla\varphi_{\rm C})^2\,\mathrm{d}^3\vec{r} = -\frac{\epsilon_0}{2}\int \varphi_{\rm C}\nabla^2\varphi_{\rm C}\,\mathrm{d}^3\vec{r} = \frac{1}{2}\int \varphi_{\rm C}(\vec{r};t)\rho(\vec{r};t)\,\mathrm{d}^3\vec{r}$$

where the first equality comes from the definition of the electric field, the second from integration by parts and the third one from the first Maxwell equation. Substitution of the Coulomb potential in terms of the charge distribution as given in {A.22.8},

$$\varphi_{\rm C}(\vec{r};t) = \int_{\text{all }\underline{\vec{r}}} \frac{\rho(\underline{\vec{r}};t)}{4\pi\epsilon_0|\vec{r}-\underline{\vec{r}}|}\,\mathrm{d}^3\underline{\vec{r}}$$

now gives the Koulomb potential energy $V_{\rm C}$ for a continuous charge distribution:

$$V_{\rm C} = \tfrac{1}{2}\int_{\text{all }\vec{r}}\int_{\text{all }\vec{r}} \frac{\rho(\vec{r};t)\rho(\underline{\vec{r}};t)}{4\pi\epsilon_0|\vec{r}-\underline{\vec{r}}|}\,\mathrm{d}^3\vec{r}\mathrm{d}^3\underline{\vec{r}}$$

For point charges, the charge distribution is by definition

$$\rho(\vec{r};t) = \sum_{i=1}^{I} q_i \delta^3(\vec{r} - \vec{r}_i)$$

Here δ^3 is the three-dimensional delta function, $\vec{r}_i = \vec{r}_i(t)$ the position of point charge i, and q_i its charge.

Recall that the delta function picks out the value at $\vec{r}_i$ from whatever it is integrated against. Using this twice on the Coulomb potential energy above,

$$V_{\rm C} = \tfrac{1}{2} \sum_{i=1}^{I} \int_{\text{all } \underline{\vec{r}}} \frac{q_i \rho(\underline{\vec{r}};t)}{4\pi\epsilon_0 |\vec{r}_i - \underline{\vec{r}}|} \, {\rm d}^3\underline{\vec{r}} = \tfrac{1}{2} \sum_{i=1}^{I} \sum_{\substack{\underline{i}=1 \\ \underline{i}\neq i}}^{I} \frac{q_i q_{\underline{i}}}{4\pi\epsilon_0 |\vec{r}_i - \vec{r}_{\underline{i}}|}$$

That is the Coulomb potential energy $V_{\rm C}$ for point charges.

Note again that physically all the energy is inside the electromagnetic field. There is no energy of interaction of the charged particles with the field. If equal charges move closer together, they increase the energy in the electromagnetic field. That requires work.

D.38 Time-dependent perturbation theory

The equations to be solved are

$$\mathrm{i}\hbar\dot{c}_1 = \langle E_1 \rangle c_1 + H_{12} c_2 \qquad \mathrm{i}\hbar\dot{c}_2 = H_{21} c_1 + \langle E_2 \rangle c_2$$

To simplify the use of perturbation theory, it is convenient to use a trick that gets rid of half the terms in these equations. The trick is to define new coefficients $\bar{c}_1$ and $\bar{c}_2$ by

$$\bar{c}_1 = c_1 e^{\mathrm{i}\int \langle E_1 \rangle \, {\rm d}t/\hbar} \qquad \bar{c}_2 = c_2 e^{\mathrm{i}\int \langle E_2 \rangle \, {\rm d}t/\hbar} \tag{D.22}$$

The new coefficients $\bar{c}_1$ and $\bar{c}_2$ are physically just as good as c_1 and c_2. For one, the probabilities are given by the square magnitudes of the coefficients, and the square magnitudes of $\bar{c}_1$ and $\bar{c}_2$ are exactly the same as those of c_1 and c_2. That is because the exponentials have magnitude one. Also, the initial conditions are unchanged, assuming that you choose the integration constants so that the integrals are initially zero.

The evolution equations for $\bar{c}_1$ and $\bar{c}_2$ are

$$\boxed{\mathrm{i}\hbar\dot{\bar{c}}_1 = H_{12} e^{-\mathrm{i}\int E_{21} \, {\rm d}t/\hbar} \bar{c}_2 \qquad \mathrm{i}\hbar\dot{\bar{c}}_2 = H_{21} e^{\mathrm{i}\int E_{21} \, {\rm d}t/\hbar} \bar{c}_1} \tag{D.23}$$

with $E_{21} = \langle E_2 \rangle - \langle E_1 \rangle$. Effectively, the two energy expectation values have been turned into zero. However, the matrix element is now time-dependent, if

it was not already. To check the above evolution equations, just plug in the definition of the coefficients.

It will from now on be assumed that the original Hamiltonian coefficients are independent of time. That makes the difference in expectation energies E_{21} constant too.

Now the formal way to perform time-dependent perturbation theory is to assume that the matrix element H_{21} is small. Write H_{21} as εH^0_{21} where ε is a scale factor. Then you can find the behavior of the solution in the limiting process $\varepsilon \to 0$ by expanding the solution in powers of ε. The definition of the scale factor ε is not important. You might identify it with a small physical parameter in the matrix element. But in fact you can take H^0_{21} the same as H_{21} and ε as an additional mathematical parameter with no meaning for the physical problem. In that approach, ε disappears when you take it to be 1 in the final answer.

But because the problem here is so trivial, there is really no need for a formal time-dependent perturbation expansion. In particular, by assumption the system stays close to state ψ_1, so the coefficient $\bar{c}_2$ must remain small. Then the evolution equations above show that $\bar{c}_1$ will hardly change. That allows it to be treated as a constant in the evolution equation for $\bar{c}_2$. That then allows $\bar{c}_2$ to be found by simple integration. The integration constant follows from the condition that c_2 is zero at the initial time. That then gives the result cited in the text.

It may be noted that for the analysis to be valid, $H_{21}t/\hbar$ must be small. That ensures that $\bar{c}_2$ is correspondingly small according to its evolution equation. And then the change in $\bar{c}_1$ from its original value is small of order $(H_{21}t/\hbar)^2$ according to its evolution equation. So the assumption that it is about constant in the equation for $\bar{c}_2$ is verified. The error will be of order $(H_{21}t/\hbar)^3$.

To be sure, this does not verify that this error in $\bar{c}_2$ decays to zero when $E_{21}t/2\hbar$ tends to infinity. But it does, as can be seen from the exact solution,

$$|c_2|^2 = \left(\frac{|H_{21}|t}{\hbar}\right)^2 \frac{\sin^2(\tilde{E}_{21}t/2\hbar)}{(\tilde{E}_{21}t/2\hbar)^2} \qquad \tilde{E}_{21} \equiv \sqrt{E_{21}^2 + |H_{21}|^2}$$

By splitting it up into ranges $|E_{21}|t/\hbar$ no larger than $|H_{21}|t/\hbar$ and $|E_{21}|t/\hbar$ no larger than 1, you can see that the error is never larger than order $(H_{21}t/\hbar)^2$ for $|E_{21}|t/\hbar$ no larger than 1. And it is of order $(H_{21}t/\hbar)^2/(|E_{21}|t/\hbar)^2$ outside that range.

Finally, consider the case that the state cannot just transition to one state ψ_2 but to a large number N of them, each with its own coefficient $\bar{c}_2$. In that case, the individual contributions of all these states add up to change $\bar{c}_1$. And $\bar{c}_1$ must definitely stay approximately constant for the above analysis to be valid. Fortunately, if you plug the approximate expressions for the $\bar{c}_2$ into the evolution equation for $\bar{c}_1$, you can see that $\bar{c}_1$ stays approximately constant as

long as the sum of all the transition probabilities does. So as long as there is little probability of *any* transition at time t, time-dependent perturbation theory should be OK.

D.39 Selection rules

This note derives the selection rules for electric dipole transitions between two hydrogen states $\psi_{\rm L}$ and $\psi_{\rm H}$. Some selection rules for forbidden transitions are also derived. The derivations for forbidden transitions use some more advanced results from later chapters. It may be noted that in any case, the Hamiltonian assumes that the velocity of the electrons is small compared to the speed of light.

According to chapter 4.3, the hydrogen states take the form $\psi_{\rm L} = \psi_{n_{\rm L} l_{\rm L} m_{\rm L}}\updownarrow$ and $\psi_{\rm H} = \psi_{n_{\rm H} l_{\rm H} m_{\rm H}}\updownarrow$. Here $1 \leqslant n$, $0 \leqslant l \leqslant n$ and $|m| \leqslant l$ are integer quantum numbers. The final $\updownarrow$ represents the electron spin state, up or down.

As noted in the text, allowed electric dipole transitions must respond to at least one component of a constant ambient electric field. That means that they must have a nonzero value for at least one electrical dipole moment,

$$\langle \psi_{\rm L} | r_i | \psi_{\rm H} \rangle \neq 0$$

where r_i can be one of $r_1 = x$, $r_2 = y$, or $r_3 = z$ for the three different components of the electric field.

The trick in identifying when these inner products are zero is based on taking inner products with cleverly chosen commutators. Since the hydrogen states are eigenfunctions of $\widehat{L}_z$, the following commutator is useful

$$\langle \psi_{\rm L} | [r_i, \widehat{L}_z] | \psi_{\rm H} \rangle = \langle \psi_{\rm L} | r_i \widehat{L}_z - \widehat{L}_z r_i | \psi_{\rm H} \rangle$$

For the $r_i \widehat{L}_z$ term in the right hand side, the operator $\widehat{L}_z$ acts on $\psi_{\rm H}$ and produces a factor $m_{\rm H}\hbar$, while for the $\widehat{L}_z r_i$ term, $\widehat{L}_z$ can be taken to the other side of the inner product and then acts on $\psi_{\rm L}$, producing a factor $m_{\rm L}\hbar$. So:

$$\langle \psi_{\rm L} | [r_i, \widehat{L}_z] | \psi_{\rm H} \rangle = (m_{\rm H} - m_{\rm L}) \hbar \langle \psi_{\rm L} | r_i | \psi_{\rm H} \rangle \tag{D.24}$$

The final inner product is the dipole moment of interest. Therefore, if a suitable expression for the commutator in the left hand side can be found, it will fix the dipole moment.

In particular, according to chapter 4.5.4 $[z, \widehat{L}_z]$ is zero. That means according to equation (D.24) above that the dipole moment $\langle \psi_{\rm L} | z | \psi_{\rm H} \rangle$ in the right hand side will have to be zero too, unless $m_{\rm H} = m_{\rm L}$. So the first conclusion is that the z-component of the electric field does not do anything unless $m_{\rm H} = m_{\rm L}$. One down, two to go.

For the x and y components, from chapter 4.5.4

$$[x, \widehat{L}_z] = -\mathrm{i}\hbar y \qquad [y, \widehat{L}_z] = \mathrm{i}\hbar x$$

Plugging that into (D.24) produces

$$-\mathrm{i}\hbar\langle\psi_\mathrm{L}|y|\psi_\mathrm{H}\rangle = (m_\mathrm{H} - m_\mathrm{L})\hbar\langle\psi_\mathrm{L}|x|\psi_\mathrm{H}\rangle \quad \mathrm{i}\hbar\langle\psi_\mathrm{L}|x|\psi_\mathrm{H}\rangle = (m_\mathrm{H} - m_\mathrm{L})\hbar\langle\psi_\mathrm{L}|y|\psi_\mathrm{H}\rangle$$

From these equations it is seen that the y dipole moment is zero if the x one is, and vice-versa. Further, plugging the y dipole moment from the first equation into the second produces

$$\mathrm{i}\hbar\langle\psi_\mathrm{L}|x|\psi_\mathrm{H}\rangle = \frac{(m_\mathrm{H} - m_\mathrm{L})^2\hbar^2}{-\mathrm{i}\hbar}\langle\psi_\mathrm{L}|x|\psi_\mathrm{H}\rangle$$

and if the x dipole moment is nonzero, that requires that $(m_\mathrm{H} - m_\mathrm{L})^2$ is one, so $m_\mathrm{H} = m_\mathrm{L} \pm 1$. It follows that dipole transitions can only occur if $m_\mathrm{H} = m_\mathrm{L}$, through the z component of the electric field, or if $m_\mathrm{H} = m_\mathrm{L} \pm 1$, through the x and y components.

To derive selection rules involving the azimuthal quantum numbers l_H and l_L, the obvious approach would be to try the commutator $[r_i, \widehat{L}^2]$ since $\widehat{L}^2$ produces $l(l+1)\hbar^2$. However, according to chapter 4.5.4, (4.68), this commutator will bring in the $\widehat{\vec{r}} \times \widehat{\vec{L}}$ operator, which cannot be handled. The commutator that works is the second of (4.73):

$$[[r_i, \widehat{L}^2], \widehat{L}^2] = 2\hbar^2(r_i\widehat{L}^2 + \widehat{L}^2 r_i)$$

where by the definition of the commutator

$$[[r_i, \widehat{L}^2], \widehat{L}^2] = (r_i\widehat{L}^2 - \widehat{L}^2 r_i)\widehat{L}^2 - \widehat{L}^2(r_i\widehat{L}^2 - \widehat{L}^2 r_i) = r_i\widehat{L}^2\widehat{L}^2 - 2\widehat{L}^2 r_i\widehat{L}^2 + \widehat{L}^2\widehat{L}^2 r_i$$

Evaluating $\langle\psi_\mathrm{L}|[[r_i, \widehat{L}^2], \widehat{L}^2]|\psi_\mathrm{H}\rangle$ according to each of the two equations above and equating the results gives

$$2\hbar^2[l_\mathrm{H}(l_\mathrm{H}+1) + l_\mathrm{L}(l_\mathrm{L}+1)]\langle\psi_\mathrm{L}|r_i|\psi_\mathrm{H}\rangle = \hbar^2[l_\mathrm{H}(l_\mathrm{H}+1) - l_\mathrm{L}(l_\mathrm{L}+1)]^2\langle\psi_\mathrm{L}|r_i|\psi_\mathrm{H}\rangle$$

For $\langle\psi_\mathrm{L}|r_i|\psi_\mathrm{H}\rangle$ to be nonzero, the numerical factors in the left and right hand sides must be equal,

$$2[l_\mathrm{H}(l_\mathrm{H}+1) + l_\mathrm{L}(l_\mathrm{L}+1)] = [l_\mathrm{H}(l_\mathrm{H}+1) - l_\mathrm{L}(l_\mathrm{L}+1)]^2$$

The right hand side is obviously zero for $l_\mathrm{H} = l_\mathrm{L}$, so $l_\mathrm{H} - l_\mathrm{L}$ can be factored out of it as

$$[l_\mathrm{H}(l_\mathrm{H}+1) - l_\mathrm{L}(l_\mathrm{L}+1)]^2 = (l_\mathrm{H} - l_\mathrm{L})^2(l_\mathrm{H} + l_\mathrm{L} + 1)^2$$

and the left hand side can be written in terms of these same factors as

$$2[l_\mathrm{H}(l_\mathrm{H}+1) + l_\mathrm{L}(l_\mathrm{L}+1)] = (l_\mathrm{H} - l_\mathrm{L})^2 + (l_\mathrm{H} + l_\mathrm{L} + 1)^2 - 1$$

Equating the two results and simplifying gives

$$[(l_{\rm H} - l_{\rm L})^2 - 1][(l_{\rm H} + l_{\rm L} + 1)^2 - 1] = 0$$

The second factor is only zero if $l_{\rm H} = l_{\rm L} = 0$, but then $\langle \psi_{\rm L} | r_i | \psi_{\rm H} \rangle$ is still zero because both states are spherically symmetric. It follows that the first factor will have to be zero for dipole transitions to be possible, and that means that $l_{\rm H} = l_{\rm L} \pm 1$.

The spin is not affected by the perturbation Hamiltonian, so the dipole moment inner products are still zero unless the spin magnetic quantum numbers m_s are the same, both spin-up or both spin-down. Indeed, if the electron spin is not affected by the electric field to the approximations made, then obviously it cannot change. That completes the selection rules as given in chapter 7.4.4 for electric dipole transitions.

Now consider the effect of the magnetic field on transitions. For such transitions to be possible, the matrix element formed with the magnetic field must be nonzero. Like the electric field, the magnetic field can be approximated as spatially constant and quasi-steady. The perturbation Hamiltonian of a constant magnetic field is according to chapter 13.4

$$H_1 = \frac{e}{2m_{\rm e}} \vec{\mathcal{B}} \cdot \left(\widehat{\vec{L}} + 2\widehat{\vec{S}} \right)$$

Note that now electron spin must be included in the discussion.

According to this perturbation Hamiltonian, the perturbation coefficient $H_{\rm HL}$ for the z-component of the magnetic field is proportional to

$$\langle \psi_{\rm L} | \widehat{L}_z + 2\widehat{S}_z | \psi_{\rm H} \rangle$$

and that is zero because $\psi_{\rm H}\updownarrow$ is an eigenfunction of both operators and orthogonal to $\psi_{\rm L}\updownarrow$. So the z-component of the magnetic field does not produce transitions to different states.

However, the x-component (and similarly the y-component) produces a perturbation coefficient proportional to

$$\langle \psi_{\rm L} | \widehat{L}_x | \psi_{\rm H} \rangle + 2 \langle \psi_{\rm L} | \widehat{S}_x | \psi_{\rm H} \rangle$$

According to chapter 12.11, the effect of $\widehat{L}_x$ on a state with magnetic quantum number $m_{\rm H}$ is to turn it into a linear combination of two similar states with magnetic quantum numbers $m_{\rm H} + 1$ and $m_{\rm H} - 1$. Therefore, for the first inner product above to be nonzero, $m_{\rm L}$ will have to be either $m_{\rm H} + 1$ or $m_{\rm H} - 1$. Also the orbital azimuthal momentum numbers l will need to be the same, and so will the spin magnetic quantum numbers m_s. And the principal quantum numbers n, for that matter; otherwise the radial parts of the wave fuctions are orthogonal.

The magnetic field simply wants to rotate the orbital angular momentum vector in the hydrogen atom. That does not change the energy, in the absence of an average ambient magnetic field. For the second inner product, the spin magnetic quantum numbers have to be different by one unit, while the orbital magnetic quantum numbers must now be equal. So, all together

$$l_{\rm H} = l_{\rm L} \qquad m_{\rm H} = m_{\rm L} \text{ or } m_{\rm L} \pm 1 \qquad m_{s,{\rm H}} = m_{s,{\rm L}} \text{ or } m_{s,{\rm L}} \pm 1$$

and either the orbital or the spin magnetic quantum numbers must be unequal. That are the selection rules as given in chapter 7.4.4 for magnetic dipole transitions. Since the energy does not change in these transitions, Fermi's golden rule would have the decay rate zero. Fermi's analysis is not exact, but such transitions should be very rare.

The logical way to proceed to electric quadrupole transitions would be to expand the electric field in a Taylor series in terms of y:

$$\vec{\mathcal{E}} = \hat{k}\mathcal{E}_{\rm f} \cos\Big(\omega(t - y/c) - \alpha\Big) \approx \hat{k}\mathcal{E}_{\rm f} \cos(\omega t - \alpha) + \hat{k}\frac{\omega}{c}\mathcal{E}_{\rm f} \sin(\omega t - \alpha) y$$

The first term is the constant electric field of the electric dipole approximation, and the second would then give the electric quadrupole approximation. However, an electric field in which $\mathcal{E}_z$ is a multiple of y is not conservative, so the electrostatic potential does no longer exist.

It is necessary to retreat to the so-called vector potential $\vec{A}$. It is then simplest to chose this potential to get rid of the electrostatic potential altogether. In that case the typical electromagnetic wave is described by the vector potential

$$\vec{A} = -\hat{k}\frac{1}{\omega}\mathcal{E}_{\rm f} \sin\Big(\omega(t - y/c) - \alpha\Big) \qquad \vec{\mathcal{E}} = -\frac{\partial \vec{A}}{\partial t} \quad \vec{\mathcal{B}} = \nabla \times \vec{A}$$

In terms of the vector potential, the perturbation Hamiltonian is, chapter 13.1 and 13.4, and assuming a weak field,

$$H_1 = \frac{e}{2m_{\rm e}}(\vec{A} \cdot \widehat{\vec{p}} + \widehat{\vec{p}}\vec{A}) + \frac{e}{m_{\rm e}}\widehat{\vec{S}} \cdot \vec{\mathcal{B}}$$

Ignoring the spatial variation of $\vec{A}$, this expression produces an Hamiltonian coefficient

$$H_{\rm HL} = -\frac{e}{m_{\rm e}\omega}\mathcal{E}_{\rm f} \sin(\omega t - \alpha)\langle \psi_{\rm L} | \widehat{p}_z | \psi_{\rm H} \rangle$$

That should be same as for the electric dipole approximation, since the field is now completely described by $\vec{A}$, but it is not quite. The earlier derivation assumed that the electric field is quasi-steady. However, $\widehat{p}_z$ is equal to the commutator $\mathrm{i}m_{\rm e}[H_0, z]/\hbar$ where H_0 is the unperturbed hydrogen atom Hamiltonian. If that is plugged in and expanded, it is found that the expressions are equivalent, provided that the perturbation frequency is close to the frequency of the

photon released in the transition, and that that frequency is sufficiently rapid that the phase shift from sine to cosine can be ignored. Those are in fact the normal conditions.

Now consider the second term in the Taylor series of $\vec{A}$ with respect to y. It produces a perturbation Hamiltonian

$$\frac{e}{m_{\mathrm{e}}}\frac{1}{c}\mathcal{E}_{\mathrm{f}}\cos(\omega t-\alpha)y\widehat{p}_z$$

The factor $y\widehat{p}_z$ can be trivially rewritten to give

$$\frac{e}{2m_{\mathrm{e}}}\frac{1}{c}\mathcal{E}_{\mathrm{f}}\cos(\omega t-\alpha)(y\widehat{p}_z-z\widehat{p}_y)+\frac{e}{2m_{\mathrm{e}}}\frac{1}{c}\mathcal{E}_{\mathrm{f}}\cos(\omega t-\alpha)(y\widehat{p}_z+z\widehat{p}_y)$$

The first term has already been accounted for in the magnetic dipole transitions discussed above, because the factor within parentheses is $\widehat{L}_x$. The second term is the electric quadrupole Hamiltonian for the considered wave.

As second terms in the Taylor series, both Hamiltonians will be much smaller than the electric dipole one. The factor that they are smaller can be estimated from comparing the first and second term in the Taylor series. Note that c/ω is proportional to the wave length λ of the electromagnetic wave. Also, the additional position coordinate in the operator scales with the atom size, call it R. So the factor that the magnetic dipole and electric quadrupole matrix elements are smaller than the electric dipole one is R/λ. Since transition probabilities are proportional to the square of the corresponding matrix element, it follows that, all else being the same, magnetic dipole and electric quadrupole transitions are slower than electric dipole ones by a factor $(R/\lambda)^2$. (But note the earlier remark on the problem for the hydrogen atom that the energy does not change in magnetic dipole transitions.)

The selection rules for the electric quadrupole Hamiltonian can be narrowed down with a bit of simple reasoning. First, since the hydrogen eigenfunctions are complete, applying any operator on an eigenfunction will always produce a linear combination of eigenfunctions. Now reconsider the derivation of the electric dipole selection rules above from that point of view. It is then seen that z only produces eigenfunctions with the same values of m and the values of l exactly one unit different. The operators x and y change both m and l by exactly one unit. And the components of linear momentum do the same as the corresponding components of position, since $\widehat{p}_i = \mathrm{i}m_{\mathrm{e}}[H_0, r_i]/\hbar$ and H_0 does not change the eigenfunctions, just their coefficients. Therefore $y\widehat{p}_z + z\widehat{p}_y$ produces only eigenfunctions with azimuthal quantum number l either equal to l_{H} or to $l_{\mathrm{H}} \pm 2$, depending on whether the two unit changes reinforce or cancel each other. Furthermore, it produces only eigenfunctions with m equal to $m_{\mathrm{H}} \pm 1$. However, $x\widehat{p}_y + y\widehat{p}_x$, corresponding to a wave along another axis, will produce values of m equal to m_{H} or to $m_{\mathrm{H}} \pm 2$. Therefore the selection rules become:

$$l_{\mathrm{H}} = l_{\mathrm{L}} \text{ or } l_{\mathrm{L}} \pm 2 \qquad m_{\mathrm{H}} = m_{\mathrm{L}} \text{ or } m_{\mathrm{L}} \pm 1 \text{ or } m_{\mathrm{L}} \pm 2 \qquad m_{s,\mathrm{H}} = m_{s,\mathrm{L}}$$

That are the selection rules as given in chapter 7.4.4 for electric quadrupole transitions. These arguments apply equally well to the magnetic dipole transition, but there the possibilities are narrowed down much further because the angular momentum operators only produce a couple of eigenfunctions. It may be noted that in addition, electric quadrupole transitions from $l_{\rm H} = 0$ to $l_{\rm L} = 0$ are not possible because of spherical symmetry.

D.40 Quantization of radiation derivations

This gives various derivations for the addendum of the same name.

It is to be shown first that

$$\int_{\rm all} c^2 (\vec{\mathcal{B}}^{\rm n}_\gamma)^2 \,{\rm d}^3\vec{r} = -\int_{\rm all} (\vec{\mathcal{E}}^{\rm n}_\gamma)^2 \,{\rm d}^3\vec{r} \tag{1}$$

To see that, note from (A.157) that

$$c\vec{\mathcal{B}}^{\rm n}_\gamma = \frac{1}{{\rm i}k} \nabla \times \vec{\mathcal{E}}^{\rm n}_\gamma$$

so the left-hand integral becomes

$$\int_{\rm all} c^2 (\vec{\mathcal{B}}^{\rm n}_\gamma)^2 \,{\rm d}^3\vec{r} = -\frac{1}{k^2} \int_{\rm all} (\nabla \times \vec{\mathcal{E}}^{\rm n}_\gamma) \cdot (\nabla \times \vec{\mathcal{E}}^{\rm n}_\gamma) \,{\rm d}^3\vec{r}$$

Now the curl, $\nabla\times$, is Hermitian, {D.10}, so the second curl can be pushed in front of the first curl. Then curl curl acts as $-\nabla^2$ because $\vec{\mathcal{E}}^{\rm n}_\gamma$ is solenoidal and the standard vector identity (D.1). And the eigenvalue problem turns $-\nabla^2$ into k^2.

Note incidentally that the additional surface integral in {D.10} is zero even for the photon modes of definite angular momentum, {A.21.7}, because for them either $\vec{\mathcal{E}}^{\rm n}_\gamma$ is zero on the surface or $\nabla \times \vec{\mathcal{E}}^{\rm n}_\gamma$ is. Also note that the integrals become equal instead of opposite if you push complex conjugates on the first factors in the integrands.

Now the Hamiltonian can be worked out. Using Using (A.152) and (A.162), it is

$$H = \tfrac{1}{4}\epsilon_0 \int_{\rm all} \left[(\widehat{a}\vec{\mathcal{E}}^{\rm n}_\gamma + \widehat{a}\vec{\mathcal{E}}^{\rm n*}_\gamma)^2 + (\widehat{a}c\vec{\mathcal{B}}^{\rm n}_\gamma + \widehat{a}^\dagger c\vec{\mathcal{B}}^{\rm n*}_\gamma)^2 \right] {\rm d}^3\vec{r}$$

When that is multiplied out and integrated, the $(\widehat{a})^2$ and $(\widehat{a}^\dagger)^2$ terms drop out because of (1). The remaining multiplied-out terms in the Hamiltonian produce the stated Hamiltonian after noting the wave function normalization (A.158).

The final issue is to identify the relationships between the coefficients D_0, D_1 and C as given in the text. The most important question is here under what circumstances $2|D_1|$ and $4|C|^2$ can get very close to the larger value $2D_0$.

The coefficient D_1 was defined as

$$2D_1 = \sum_i c^*_{i-1} c_{i+1} \sqrt{i}\sqrt{i+1}$$

To estimate this, consider the infinite-dimensional vectors $\vec{a}$ and $\vec{b}$ with coefficients

$$a_i \equiv c_{i-1}\sqrt{i} \qquad b_i \equiv c_{i+1}\sqrt{i+1}$$

Note that $2D_1$ above is the inner product of these two vectors. And an inner product is less in magnitude than the product of the lengths of the vectors involved.

$$|2D_1| = |\langle \vec{a} \big| \vec{b} \rangle| \leqslant |\vec{a}||\vec{b}| = \sqrt{\left[\sum_i |c_{i-1}|^2 i\right]\left[\sum_i |c_{i+1}|^2 (i+1)\right]}$$

By changing the notations for the summation indices, (letting $i-1 \to i$ and $i+1 \to i$), the sums become the expectation values of $i+1$, respectively i. So

$$|2D_1| \leqslant \sqrt{(\langle i \rangle + 1)(\langle i \rangle)} = \sqrt{\langle i \rangle^2 + \langle i \rangle} < \sqrt{\langle i \rangle^2 + \langle i \rangle + \tfrac{1}{4}} = \sqrt{(\langle i \rangle + \tfrac{1}{2})^2} = 2D_0$$

The final equality is by the definition of D_0. The second inequality already implies that $|D_1|$ is always smaller than D_0. However, if the expectation value of i is large, it does not make much of a difference.

In that case, the bigger problem is the inner product between the vectors $\vec{a}$ and $\vec{b}$. Normally it is smaller than the product of the lengths of the vectors. For it to become equal, the two vectors have to be proportional. The coefficients of $\vec{b}$ must be some multiple, call it $B^2 e^{2\mathrm{i}\beta}$, of those of $\vec{a}$:

$$c_{i+1}\sqrt{i+1} \approx B^2 e^{2\mathrm{i}\beta} c_{i-1}\sqrt{i}$$

For larger values of i the square roots are about the same. Then the above relationship requires an exponential decay of the coefficients. For small values of i, obviously the above relation cannot be satisfied. The needed values of c_i for negative i do not exist. To reduce the effect of this "start-up" problem, significant coefficients will have to exist for a considerable range of i values.

In addition to the above conditions, the coefficient $4|C|^2$ has to be close to $2D_0$. Here the coefficient C was defined as

$$\sqrt{2}C = \sum_i c^*_{i-1} c_i \sqrt{i}$$

Using the same manipulations as for D_1, but with

$$a_i \equiv c_{i-1}\sqrt{\sqrt{i}} \qquad b_i \equiv c_{i+1}\sqrt{\sqrt{i}}$$

gives

$$2|C|^2 \leqslant \left[\sum_i |c_{i-1}|^2 \sqrt{i}\right] \left[\sum_i |c_i|^2 \sqrt{i}\right] = \left\langle \sqrt{i+1} \right\rangle \left\langle \sqrt{i} \right\rangle$$

To bound this further, define

$$f(x) = \left\langle \sqrt{i + \tfrac{1}{2} + x} \right\rangle$$

By expanding the square root in a Taylor series,

$$f(-\tfrac{1}{2}) < f(0) - \Delta f \qquad f(\tfrac{1}{2}) < f(0) + \Delta f$$

where Δf is the expectation value of the linear term in the Taylor series; the inequalities express that a square root function has a negative second order derivative. Multiplying these two expressions shows that

$$f(-\tfrac{1}{2}) f(\tfrac{1}{2}) < f^2(0) \qquad \Longrightarrow \qquad \left\langle \sqrt{i+1} \right\rangle \left\langle \sqrt{i} \right\rangle < \left\langle \sqrt{i + \tfrac{1}{2}} \right\rangle^2$$

Since it has already been shown that the expectation value of i must be large, this inequality will be almost an equality, anyway.

In any case,

$$2|C|^2 < \left\langle \sqrt{i + \tfrac{1}{2}} \right\rangle^2$$

This is less than

$$\left\langle \sqrt{i + \tfrac{1}{2}}^2 \right\rangle = 2D_0$$

The big question is now how much it is smaller. To answer that, use the shorthand

$$\sqrt{i + \tfrac{1}{2}} \equiv x_i = x + x_i'$$

where x is the expectation value of the square root and x_i' is the deviation from the average. Then, noting that the expectation value of x_i' is zero,

$$2D_0 = \left\langle (x + x_i')^2 \right\rangle = \langle x \rangle^2 + \left\langle (x_i')^2 \right\rangle$$

The second-last term is the bound for $2|C|^2$ as obtained above. So, the only way that $2|C|^2$ can be close to $2D_0$ is if the final term is relatively small. That means that the deviation from the expectation square root must be relatively small. So the coefficients c_i can only be significant in some limited range around an average value of i. In addition, for the vectors $\vec{a}$ and $\vec{b}$ in the earlier estimate for C to be almost proportional,

$$c_{i-1} \sqrt{\sqrt{i}} \approx A e^{\mathrm{i}\alpha} c_i \sqrt{\sqrt{i}}$$

where $Ae^{\mathrm{i}\alpha}$ is some constant. That again means an exponential dependence, like for the condition on D_1. And $Ae^{\mathrm{i}\alpha}$ will have to be approximately $Be^{\mathrm{i}\beta}$. And A will have to be about 1, because otherwise start and end effects will dominate the exponential part. That gives the situation as described in the text.

D.41 Derivation of the Einstein B coefficients

The purpose of this note is to derive the Einstein B coefficients of chapter 7.8. They determine the transition rates between the energy states of atoms. For simplicity it will be assumed that there are just two atomic energy eigenstates involved, a lower energy one $\psi_{\rm L}$ and an higher energy one $\psi_{\rm H}$. It is further assumed that the atoms are subject to incoherent ambient electromagnetic radiation. The energy in the ambient radiation is $\rho(\omega)$ per unit volume and unit frequency range. Finally it is assumed that the atoms suffer frequent collisions with other atoms. The typical time between collisions will be indicated by $t_{\rm c}$. It is small compared to the typical decay time of the states, but large compared to the frequency of the relevant electromagnetic field.

Unlike what you may find elsewhere, it will not be assumed that the atoms are either fully in the high or fully in the low energy state. That is a highly unsatisfactory assumption for many reasons. For one thing it assumes that the atoms know what you have selected as z-axis. In the derivation below, the atoms are allowed to be in a linear combination of the states $\psi_{\rm L}$ and $\psi_{\rm H}$, with coefficients $c_{\rm L}$ and $c_{\rm H}$.

Since both the electromagnetic field and the collisions are random, a statistical rather than a determinate treatment is needed. In it, the probability that a randomly chosen atom can be found in the lower energy state $\psi_{\rm L}$ will be indicated by $P_{\rm L}$. Similarly, the probability that an atom can be found in the higher energy state $\psi_{\rm H}$ will be indicated by $P_{\rm H}$. For a single atom, these probabilities are given by the square magnitudes of the coefficients $c_{\rm L}$ and $c_{\rm H}$ of the energy states. Therefore, $P_{\rm L}$ and $P_{\rm H}$ will be defined as the averages of $|c_{\rm L}|^2$ respectively $|c_{\rm H}|^2$ over all atoms.

It is assumed that the collisions are globally elastic in the sense that they do not change the average energy picture of the atoms. In other words, they do not affect the average probabilities of the eigenfunctions $\psi_{\rm L}$ and $\psi_{\rm H}$. However, they are assumed to leave the wave function of an individual atom immediately after a collision in some state $c_{\rm L,0}\psi_{\rm L} + c_{\rm H,0}\psi_{\rm H}$ in which $c_{\rm L,0}$ and $c_{\rm H,0}$ are quite random, especially with respect to their phase. What is now to be determined in this note is how, until the next collision, the wave function of the atom will develop under the influence of the electromagnetic field and how that changes the average probabilities $|c_{\rm L}|^2$ and $|c_{\rm H}|^2$.

The evolution equations of the coefficients $\bar c_{\rm L}$ and $\bar c_{\rm H}$, in between collisions, were given in chapter 7.7.2 (7.42). They are in terms of modified variables $\bar c_{\rm L}$ and $\bar c_{\rm H}$. However, these variables have the same square magnitudes and initial conditions as $c_{\rm L}$ and $c_{\rm H}$. So it really does not make a difference.

Further, because the equations are linear, the solution for the coefficients $\bar c_{\rm L}$ and $\bar c_{\rm H}$ can be written as a sum of two contributions, one proportional to the initial value $\bar c_{\rm L,0}$ and the other to $\bar c_{\rm H,0}$:

$$\bar c_{\rm L} = \bar c_{\rm L,0}\bar c_{\rm L}^{\rm L} + \bar c_{\rm H,0}\bar c_{\rm L}^{\rm H} \qquad \bar c_{\rm H} = \bar c_{\rm L,0}\bar c_{\rm H}^{\rm L} + \bar c_{\rm H,0}\bar c_{\rm H}^{\rm H}$$

Here $(\bar{c}^{\rm L}_{\rm L}, \bar{c}^{\rm L}_{\rm H})$ is the solution that starts out from the lower energy state $(\bar{c}^{\rm L}_{\rm L}, \bar{c}^{\rm L}_{\rm H}) = (1, 0)$ while $(\bar{c}^{\rm H}_{\rm L}, \bar{c}^{\rm H}_{\rm H})$ is the solution that starts out from the higher energy state $(\bar{c}^{\rm H}_{\rm L}, \bar{c}^{\rm H}_{\rm H}) = (0, 1)$.

Now consider what happens to the probability of an atom to be in the excited state in the time interval between collisions:

$$|\bar{c}_{\rm H}|^2 - |\bar{c}_{\rm H,0}|^2 = (\bar{c}_{\rm H,0} + \bar{c}_{\rm L,0}\Delta\bar{c}^{\rm L}_{\rm H} + \bar{c}_{\rm H,0}\Delta\bar{c}^{\rm H}_{\rm H})^*(\bar{c}_{\rm H,0} + \bar{c}_{\rm L,0}\Delta\bar{c}^{\rm L}_{\rm H} + \bar{c}_{\rm H,0}\Delta\bar{c}^{\rm H}_{\rm H}) - \bar{c}^*_{\rm H,0}\bar{c}_{\rm H,0}$$

Here $\Delta\bar{c}^{\rm L}_{\rm H}$ indicates the change in $\bar{c}^{\rm L}_{\rm H}$ in the time interval between collisions; in particular $\Delta\bar{c}^{\rm L}_{\rm H} = \bar{c}^{\rm L}_{\rm H}$ since this solution starts from the ground state with $\bar{c}^{\rm L}_{\rm H} = 0$. Similarly, the change $\Delta\bar{c}^{\rm H}_{\rm H}$ equals $\bar{c}^{\rm H}_{\rm H} - 1$ since this solution starts out from the excited state with $\bar{c}^{\rm H}_{\rm H} = 1$.

Because the typical time between collisions $t_{\rm c}$ is assumed small, so will be the changes $\Delta\bar{c}^{\rm L}_{\rm H}$ and $\Delta\bar{c}^{\rm H}_{\rm H}$ as given by the evolution equations (7.42). Note also that $\Delta\bar{c}^{\rm H}_{\rm H}$ will be quadratically small, since the corresponding solution starts out from $\bar{c}^{\rm H}_{\rm L} = 0$, so $\bar{c}^{\rm H}_{\rm L}$ is an additional small factor in the equation (7.42) for $\bar{c}^{\rm H}_{\rm H}$.

Therefore, if the change in probability $|\bar{c}_{\rm H}|^2$ above is multiplied out, ignoring terms that are cubically small or less, the result is, (remember that for a complex number c, $c + c^*$ is twice its real part):

$$|\bar{c}_{\rm H}|^2 - |\bar{c}_{\rm H,0}|^2 = 2\Re\left(\bar{c}^*_{\rm H,0}\bar{c}_{\rm L,0}\Delta\bar{c}^{\rm L}_{\rm H}\right) + |\bar{c}_{\rm L,0}|^2|\Delta\bar{c}^{\rm L}_{\rm H}|^2 + |\bar{c}_{\rm H,0}|^2 2\Re\left(\Delta\bar{c}^{\rm H}_{\rm H}\right)$$

Now if this is averaged over all atoms and time intervals between collisions, the first term in the right hand side will average away. The reason is that it has a random phase angle, for one since those of $\bar{c}_{\rm L,0}$ and $\bar{c}_{\rm H,0}$ are assumed to be random after a collision. For a number with a random phase angle, the real part is just as likely to be positive as negative, so it averages away. Also, for the final term, $2\Re(\Delta\bar{c}^{\rm H}_{\rm H})$ is the approximate change in $|\bar{c}^{\rm H}_{\rm H}|^2$ in the time interval, and that equals $-|\Delta\bar{c}^{\rm H}_{\rm L}|^2$ because of the normalization condition $|\bar{c}^{\rm H}_{\rm L}|^2 + |\bar{c}^{\rm H}_{\rm H}|^2 = 1$. So the relevant expression for the average change in probability becomes

$$|\bar{c}_{\rm H}|^2 - |\bar{c}_{\rm H,0}|^2 = |\bar{c}_{\rm L,0}|^2|\Delta\bar{c}^{\rm L}_{\rm H}|^2 - |\bar{c}_{\rm H,0}|^2|\Delta\bar{c}^{\rm H}_{\rm L}|^2$$

Summing the changes in the probabilities therefore means summing the changes in the square magnitudes of $\Delta\bar{c}^{\rm L}_{\rm H}$ and $\Delta\bar{c}^{\rm H}_{\rm L}$.

If the above expression for the average change in the probability of the high energy state is compared to (7.46), it is seen that the Einstein coefficient $B_{\rm L\to H}$ is the average change $|\Delta\bar{c}^{\rm L}_{\rm H}|^2$ per unit time. This is admittedly the same answer you would get if you assumed that the atoms are either in the low energy state or in the high energy state immediately after each collision. But as noted, that assumption is simply not reasonable.

Now the needed $\Delta\bar{c}^{\rm L}_{\rm H} = \bar{c}^{\rm L}_{\rm H}$ may be found from the second evolution equation (7.42). To do so, you can consider $\bar{c}^{\rm L}_{\rm L}$ to be 1. The reason is that it starts out as 1, and it never changes much because of the assumed short evolution time $t_{\rm c}$

compared to the typical transition time between states. That allows $\bar{c}_{\rm H}^{\rm L}$ to be found from a simple integration. And the second term in the modified Hamiltonian coefficient (7.44) can be ignored because of the additional assumption that $t_{\rm c}$ is still large compared to the frequency of the electromagnetic wave. That causes the exponential in the second term to oscillate rapidly and it does not integrate to a sizable contribution.

What is left is

$$\Delta\bar{c}_{\rm H}^{\rm L} = \frac{\mathcal{E}_{\rm f}}{\hbar}\langle\psi_{\rm L}|ez|\psi_{\rm H}\rangle e^{{\rm i}\alpha}\frac{e^{-{\rm i}(\omega-\omega_0)t}-1}{2(\omega-\omega_0)} \tag{D.25}$$

and $\Delta\bar{c}_{\rm L}^{\rm H}$ is given by a virtually identical expression. However, since it is assumed that the atoms are subject to incoherent radiation of all wave numbers $\vec{k}$ and polarizations p, the complete $\Delta\bar{c}_{\rm H}^{\rm L}$ will consist of the sum of all their contributions:

$$\Delta\bar{c}_{\rm H}^{\rm L} = \sum_{\vec{k},p}\Delta\bar{c}_{\rm H}^{\rm L}(\vec{k},p)$$

(This really assumes that the particles are in a very large periodic box so that the electromagnetic field is given by a Fourier series; in free space you would need to integrate over the wave numbers instead of sum over them.) The square magnitude is then

$$|\Delta\bar{c}_{\rm H}^{\rm L}|^2 = \sum_{\vec{k},p}\sum_{\underline{\vec{k}},\underline{p}}\Delta\bar{c}_{\rm H}^{\rm L,*}(\vec{k},p)\Delta\bar{c}_{\rm H}^{\rm L}(\underline{\vec{k}},\underline{p}) = \sum_{\vec{k},p}|\Delta\bar{c}_{\rm H}^{\rm L}(\vec{k},p)|^2$$

where the final equality comes from the assumption that the radiation is incoherent, so that the phases of different waves are uncorrelated and the corresponding products average to zero.

The bottom line is that square magnitudes must be summed together to find the total contribution of all waves. And the square magnitude of the contribution of a single wave is, according to (D.25) above,

$$|\Delta\bar{c}_{\rm H}^{\rm L}(\vec{k},p)|^2 = \left|\frac{\mathcal{E}_{\rm f}}{2\hbar}\langle\psi_{\rm L}|ez|\psi_{\rm H}\rangle\right|^2 t^2\left(\frac{\sin\left(\frac{1}{2}(\omega-\omega_0)t\right)}{\frac{1}{2}(\omega-\omega_0)t}\right)^2$$

Now broadband radiation is described in terms of an electromagnetic energy density $\rho(\omega)$; in particular $\rho(\omega)\,{\rm d}\omega$ gives the energy per unit volume due to the electromagnetic waves in an infinitesimal frequency range ${\rm d}\omega$ around a frequency ω. For a single wave, this energy equals $\frac{1}{2}\epsilon_0\mathcal{E}_{\rm f}^2$, chapter 13.2 (13.11). And the square amplitudes of different waves simply add up to the total energy; that is the so-called Parseval equality of Fourier analysis. So to sum the expression above over all the frequencies ω of the broadband radiation, make the substitution $\mathcal{E}_{\rm f}^2 = 2\rho(\omega)\,{\rm d}\omega/\epsilon_0$ and integrate:

$$|\Delta\bar{c}_{\rm H}^{\rm L}|^2 = \frac{|\langle\psi_{\rm L}|ez|\psi_{\rm H}\rangle|^2}{2\hbar^2\epsilon_0}t_{\rm c}^2\int_{\omega=0}^{\infty}\rho(\omega)\left(\frac{\sin\left(\frac{1}{2}(\omega-\omega_0)t_{\rm c}\right)}{\frac{1}{2}(\omega-\omega_0)t_{\rm c}}\right)^2{\rm d}\omega$$

If a change of integration variable is made to $u = \frac{1}{2}(\omega - \omega_0)t_c$, the integral becomes

$$|\Delta \bar{c}_H^L|^2 = \frac{|\langle \psi_L | ez | \psi_H \rangle|^2}{\hbar^2 \epsilon_0} t \int_{u=-\frac{1}{2}\omega_0 t_c}^{\infty} \rho(\omega_0 + 2(u/t_c)) \left(\frac{\sin u}{u} \right)^2 \, du$$

Recall that a starting assumption underlying these derivations was that $\omega_0 t_c$ was large. So the lower limit of integration can be approximated as $-\infty$.

Note that this is essentially the same analysis as the one for Fermi's golden rule, except for the presence of the given field strength ρ. However, here the mathematics can be performed more straightforwardly, using integration rather than summation.

Consider for a second the limiting process that the field strength ρ goes to zero, and that the atom is kept isolated enough that the collision time t_c can increase correspondingly. Then the term $2u/t_c$ in the argument of ρ will tend to zero. So only waves with the exact frequency $\omega = \omega_0$ will produce transitions in the limit of zero field strength. That confirms the basic claim of quantum mechanics that only the energy eigenvalues are measurable. In the absence of an electromagnetic field and other disturbances, the energy eigenvalues are purely the atomic ones. (Also recall that relativistic quantum mechanics adds that in reality, the electric field is never zero.)

In any case, while the term $2u/t_c$ may not be exactly zero, it is certainly small compared to ω_0 because of the assumption that $\omega_0 t_c$ is large. So the term may be ignored anyway. Then $\rho(\omega_0)$ is a constant in the integration and can be taken out. The remaining integral is in table books, [41, 18.36], and the result is

$$|\Delta \bar{c}_H^L|^2 = \frac{\pi |\langle \psi_L | ez | \psi_H \rangle|^2}{\hbar^2 \epsilon_0} \rho(\omega_0) t$$

This must still be averaged over all directions of wave propagation and polarization. That gives:

$$|\Delta \bar{c}_H^L|^2 = \frac{\pi |\langle \psi_L | e\vec{r} | \psi_H \rangle|^2}{3\hbar^2 \epsilon_0} \rho(\omega_0) t_c$$

where

$$|\langle \psi_L | e\vec{r} | \psi_H \rangle|^2 = |\langle \psi_L | ex | \psi_H \rangle|^2 + |\langle \psi_L | ey | \psi_H \rangle|^2 + |\langle \psi_L | ez | \psi_H \rangle|^2.$$

To see why, consider the electromagnetic waves propagating along any axis, not just the y-axis, and polarized in either of the other two axial directions. These waves will include ex and ey as well as ez in the transition probability, making the average as shown above. And of course, waves propagating in an oblique rather than axial direction are simply axial waves when seen in a rotated coordinate system and produce the same average.

The Einstein coefficient $B_{\mathrm{L}\to\mathrm{H}}$ is the average change per unit time, so the claimed (7.47) results from dividing by the time t_c between collisions. There is no need to do $B_{\mathrm{H}\to\mathrm{L}}$ separately from $\Delta\bar{c}^{\mathrm{L}}_{\mathrm{L}}$; it follows immediately from the symmetry property mentioned at the end of chapter 7.7.2 that it is the same.

D.42 Derivation of the Einstein A coefficients

Einstein did not really derive the spontaneous emission rate from relativistic quantum mechanics. That did not exist at the time. Instead Einstein used a dirty trick; he peeked at the solution.

To see how, consider a system of identical atoms that can be in a low energy state ψ_{L} or in an excited energy state ψ_{H}. The fraction of atoms in the low energy state is P_{L} and the fraction in the excited energy state is P_{H}. Einstein assumed that the fraction P_{H} of excited atoms would evolve according to the equation

$$\frac{\mathrm{d}P_{\mathrm{H}}}{\mathrm{d}t} = B_{\mathrm{L}\to\mathrm{H}}\rho(\omega_0)\ P_{\mathrm{L}} - B_{\mathrm{H}\to\mathrm{L}}\rho(\omega_0)\ P_{\mathrm{H}} - A_{\mathrm{H}\to\mathrm{L}}\ P_{\mathrm{H}}$$

where $\rho(\omega)$ is the ambient electromagnetic field energy density, ω_0 the frequency of the photon emitted in a transition from the high to the low energy state, and the A and B values are constants. This assumption agrees with the expression (7.46) given in chapter 7.8.

Then Einstein demanded that in an equilibrium situation, in which P_{H} is independent of time, the formula must agree with Planck's formula for the blackbody electromagnetic radiation energy. The equilibrium version of the formula above gives the energy density as

$$\rho(\omega_0) = \frac{A_{\mathrm{H}\to\mathrm{L}}/B_{\mathrm{H}\to\mathrm{L}}}{(B_{\mathrm{L}\to\mathrm{H}}P_{\mathrm{L}}/B_{\mathrm{H}\to\mathrm{L}}P_{\mathrm{H}}) - 1}$$

Equating this to Planck's blackbody spectrum as derived in chapter 6.8 (6.11) gives

$$\frac{A_{\mathrm{H}\to\mathrm{L}}/B_{\mathrm{H}\to\mathrm{L}}}{(B_{\mathrm{L}\to\mathrm{H}}P_{\mathrm{L}}/B_{\mathrm{H}\to\mathrm{L}}P_{\mathrm{H}}) - 1} = \frac{\hbar}{\pi^2 c^3}\frac{\omega_0^3}{e^{\hbar\omega_0/k_{\mathrm{B}}T} - 1}$$

The atoms can be modeled as distinguishable particles. Therefore the ratio $P_{\mathrm{H}}/P_{\mathrm{L}}$ can be found from the Maxwell-Boltzmann formula of chapter 6.14; that gives the ratio as $e^{-(E_{\mathrm{H}}-E_{\mathrm{L}})/k_{\mathrm{B}}T}$, or $e^{-\hbar\omega_0/k_{\mathrm{B}}T}$ in terms of the photon frequency. It then follows that for the two expressions for $\rho(\omega_0)$ to be equal,

$$B_{\mathrm{L}\to\mathrm{H}} = B_{\mathrm{H}\to\mathrm{L}} \qquad \frac{A_{\mathrm{H}\to\mathrm{L}}}{B_{\mathrm{H}\to\mathrm{L}}} = \frac{\hbar\omega_0^3}{\pi^2 c^3}$$

That $B_{\mathrm{L}\to\mathrm{H}}$ must equal $B_{\mathrm{H}\to\mathrm{L}}$ is a consequence of the symmetry property mentioned at the end of chapter 7.7.2. But it was not self-evident when Einstein wrote the paper; Einstein really invented stimulated emission here.

The valuable result for this book is the formula for the spontaneous emission rate $A_{\mathrm{H}\to\mathrm{L}}$. With $B_{\mathrm{H}\to\mathrm{L}}$ given by (7.47), it determines the spontaneous emission rate. So it has been obtained without using relativistic quantum mechanics. (Or at least not explicitly; there simply are no nonrelativistic photons.)

D.43 Multipole derivations

This derives the multipole matrix elements corresponding to a single particle in an atom or nucleus. These will normally still need to be summed over all particles.

Both a basis of linear momentum photon wave functions and of angular momentum ones are covered. For the angular momentum wave functions, the long wave length approximation will be made that kR is small. Here k is the photon wave number and R the typical size of atom or nucleus.

The derivations include a term due to an effect that was mentioned in the initial 1952 derivation by B. Stech, [44]. This effect is not mentioned in any textbook that the author is aware off. That seems to be unjustified. The term does not appear to be small for nuclei, but at the very least comparable to the usual electric multipole element given.

The rules of engagement are as follows:

- The considered particle will be indicated by a subscript i.
- The Cartesian axes are numbered using an index j, with $j = 1$, 2, and 3 for x_i, y_i, and z_i respectively.
- Also, $r_{i,j}$ indicates the coordinate in the j direction, x_i, y_i, or z_i.
- Derivatives with respect to a coordinate $r_{i,j}$ are indicated by a simple subscript j.
- If the quantity being differentiated is a vector, a comma is used to separate the vector index from differentiation ones.
- A bare $\int$ integral sign is assumed to be an integration over all nuclear coordinates.
- A superscript * indicates a complex conjugate.

The convoluted derivations in this note make use of a trick. Since "trick" sounds too tricky, it will be referred to as:

Lemma 1: This lemma allows you to get rid of derivatives on the wave function. The lemma assumes nonrelativistic particles. It is a generalization of a derivation of [16].

The lemma says that if i is the number of a particle in the atom or nucleus, and if F_i is any function of the position of that particle i, then

$$\left\langle \psi_{\mathrm{L}} \middle| (\nabla_i F_i)\cdot\nabla_i \middle| \psi_{\mathrm{H}} \right\rangle = \frac{m_i}{\hbar^2}\left\langle \psi_{\mathrm{L}} \middle| (E_{\mathrm{H}}-E_{\mathrm{L}})F_i + [V,F_i] \middle| \psi_{\mathrm{H}} \right\rangle - \tfrac{1}{2}\left\langle \psi_{\mathrm{L}} \middle| \nabla_i^2 F_i \middle| \psi_{\mathrm{H}} \right\rangle \tag{D.26}$$

Here ∇_i represents the vector of derivatives with respect to the coordinates of particle i, V is the potential, and $\psi_{\rm L}$ and $\psi_{\rm H}$ are the final and initial atomic or nuclear wave functions.

The energy difference can be expressed in terms of the energy $\hbar\omega_0$ of the nominal photon emitted in the transition,

$$\boxed{\left\langle \psi_{\rm L}\middle|(\nabla_i F_i)\cdot\nabla_i\middle|\psi_{\rm H}\right\rangle = \frac{m_i\omega_0}{\hbar}\left\langle \psi_{\rm L}\middle|\pm F_i + [V/\hbar\omega_0, F_i]\middle|\psi_{\rm H}\right\rangle - \tfrac{1}{2}\left\langle \psi_{\rm L}\middle|\nabla_i^2 F_i\middle|\psi_{\rm H}\right\rangle} \tag{D.27}$$

The $\pm$ allows for the possibility (in absorption) that $\psi_{\rm L}$ is actually the high energy state. The nominal photon frequency ω_0 is normally taken equal to the actual photon frequency ω.

Note that none of my sources includes the commutator in the first term, not even [16]. (The original 1952 derivation by [44] used a relativistic Dirac formulation, in which the term appears in a different place than here. The part in which it appears there is small without the term and is not worked out with it included.) The commutator is zero if the potential V only depends on the position coordinates of the particles. However, nuclear potentials include substantial momentum terms.

To prove the lemma, start with the left hand side

$$\langle\psi_{\rm L}|(\nabla_i F_i)\cdot\nabla_i|\psi_{\rm H}\rangle \equiv \int \psi_{\rm L}^* F_{i,j}\psi_{{\rm H},j}$$

where subscripts $j = 1$, 2, and 3 indicates the derivatives with respect to the three coordinates of particle i. Summation over j is to be understood. Average the above expression with what you get from doing an integration by parts:

$$\langle\psi_{\rm L}|(\nabla_i F_i)\cdot\nabla_i|\psi_{\rm H}\rangle = \tfrac{1}{2}\int \psi_{\rm L}^* F_{i,j}\psi_{{\rm H},j} - \tfrac{1}{2}\int (\psi_{\rm L}^* F_{i,j})_j\psi_{\rm H}$$

or differentiating out

$$\langle\psi_{\rm L}|(\nabla_i F_i)\cdot\nabla_i|\psi_{\rm H}\rangle = \tfrac{1}{2}\int \psi_{\rm L}^* F_{i,j}\psi_{{\rm H},j} - \tfrac{1}{2}\int \psi_{{\rm L},j}^* F_{i,j}\psi_{\rm H} - \tfrac{1}{2}\int \psi_{\rm L}^* F_{i,jj}\psi_{\rm H}$$

Combine the first two integrals

$$\langle\psi_{\rm L}|(\nabla_i F_i)\cdot\nabla_i|\psi_{\rm H}\rangle = \tfrac{1}{2}\int (\psi_{\rm L}^*\psi_{{\rm H},j} - \psi_{{\rm L},j}^*\psi_{\rm H})F_{i,j} - \tfrac{1}{2}\int \psi_{\rm L}^* F_{i,jj}\psi_{\rm H}$$

and do another integration by parts (I got this from [16], thanks):

$$\langle\psi_{\rm L}|(\nabla_i F_i)\cdot\nabla_i|\psi_{\rm H}\rangle = -\tfrac{1}{2}\int (\psi_{\rm L}^*\psi_{{\rm H},jj} - \psi_{{\rm L},jj}^*\psi_{\rm H})F_i - \tfrac{1}{2}\int \psi_{\rm L}^* F_{i,jj}\psi_{\rm H}$$

Now note the nonrelativistic eigenvalue problems for the two states

$$-\frac{\hbar^2}{2m_i}\nabla_i^2\psi_{\rm L} - \sum_{\underline{i}\neq i}\frac{\hbar^2}{2m_{\underline{i}}}\nabla_{\underline{i}}^2\psi_{\rm L} + V\psi_{\rm L} = E_{\rm L}\psi_{\rm L}$$

$$-\frac{\hbar^2}{2m_i}\nabla_i^2\psi_{\rm H} - \sum_{\underline{i}\neq i}\frac{\hbar^2}{2m_{\underline{i}}}\nabla_{\underline{i}}^2\psi_{\rm H} + V\psi_{\rm H} = E_{\rm H}\psi_{\rm H}$$

Here the sum is over the other particles in the nucleus. These two eigenvalue problems are used to eliminate the second order derivatives in the integral above. The terms involving the Laplacians with respect to the coordinates of the other particles then drop out. The reason is that F_i is just a constant with respect to those coordinates, and that Laplacians are Hermitian. Assuming that V is at least Hermitian, as it should, the V terms produce the commutator in the lemma. And the right hand sides give the energy-difference term. The result is the lemma as stated.

D.43.1 Matrix element for linear momentum modes

This requires in addition:
Lemma 2: This lemma allows you to express a certain combination of derivatives in terms of the angular momentum operator. It will be assumed that vector $\vec{A}_0$ is normal to vector $\vec{k}$.

In that case:

$$(\vec{k}\cdot\vec{r}_i)(\vec{A}_0\cdot\nabla_i) - (\vec{A}_0\cdot\vec{r}_i)(\vec{k}\cdot\nabla_i) = (\vec{k}\times\vec{A}_0)\cdot(\vec{r}_i\times\nabla_i) = \frac{\mathrm{i}}{\hbar}(\vec{k}\times\vec{A}_0)\cdot\widehat{\vec{L}}_i$$

The quickest way to prove this is to take the x-axis in the direction of $\vec{k}$, and the y-axis in the direction of $\vec{A}_0$. (The expression above is also true if the two vectors are nor orthogonal. You can see that using index notation. However, that will not be needed.) The final equality is just the definition of the angular momentum operator.

The objective is now to use these lemmas to work out the matrix element

$$H_{21,i} = -\frac{q_i}{m_i}\langle\psi_{\rm L}|\vec{A}_0 e^{-\mathrm{i}\vec{k}\cdot\vec{r}_i}\cdot\widehat{\vec{p}}_i|\psi_{\rm H}\rangle$$

where $\vec{k}$ is the constant wave number vector and $\vec{A}_0$ is some other constant vector normal to $\vec{k}$. Also $\vec{r}_i$ is the position of the considered particle, and $\widehat{\vec{p}}_i$ is the momentum operator $\hbar\nabla_i/\mathrm{i}$ based on these coordinates.

To reduce this, take the factor $\hbar/\mathrm{i}$ out of $\widehat{\vec{p}}_i$ and write the exponential in a Taylor series:

$$H_{21,i} = \sum_{n=0}^{\infty}\frac{\mathrm{i}q_i\hbar}{m_i}\langle\psi_{\rm L}|\frac{(-\mathrm{i}\vec{k}\cdot\vec{r}_i)^n}{n!}\vec{A}_0\cdot\nabla_i|\psi_{\rm H}\rangle$$

Take another messy factor out of the inner product:

$$H_{21,i} = \sum_{n=0}^{\infty} \frac{\mathrm{i}(-\mathrm{i})^n q_i \hbar}{m_i (n+1)!} \langle \psi_\mathrm{L} | (n+1)(\vec{k}\cdot\vec{r}_i)^n \vec{A}_0 \cdot \nabla_i | \psi_\mathrm{H} \rangle$$

For brevity, just consider the inner product by itself for now. It can trivially be rewritten as a sum of two terms, ([16], not me):

$$\langle \psi_\mathrm{L} | (\vec{k}\cdot\vec{r}_i)^{n-1} [(\vec{k}\cdot\vec{r}_i)\vec{A}_0 \cdot \nabla_i + n(\vec{A}_0\cdot\vec{r}_i)\vec{k}\cdot\nabla_i] | \psi_\mathrm{H} \rangle \quad (1)$$

$$+ \langle \psi_\mathrm{L} | (\vec{k}\cdot\vec{r}_i)^{n-1} n [(\vec{k}\cdot\vec{r}_i)\vec{A}_0 \cdot \nabla_i - (\vec{A}_0\cdot\vec{r}_i)\vec{k}\cdot\nabla_i] | \psi_\mathrm{H} \rangle \quad (2)$$

Now on the first inner product (1), lemma 1 can be applied with

$$F_i = (\vec{k}\cdot\vec{r}_i)^n (\vec{A}_0\cdot\vec{r}_i) \quad \Longrightarrow \quad \nabla_i^2 F_i = n(n-1)k^2 (\vec{k}\cdot\vec{r}_i)^{n-2} (\vec{A}_0\cdot\vec{r}_i)$$

(Recall that $\vec{A}_0$ and $\vec{k}$ are orthogonal. Also note that the Laplacian of F_i is of essentially the same form as F_i, just for a different value of n.) On the second inner product (2), lemma 2 can be applied.

Plugging these results back into the expression for the matrix element, renotating n into $\ell - 1$ for the first part of (1), into $\ell + 1$ for the second part, which can then be combined with the first part, and into ℓ for (2), and cleaning up gives the final result:

$$H_{21,i} = -\frac{q_i}{m_i} \langle \psi_\mathrm{L} | \vec{A}_0 e^{-\mathrm{i}\vec{k}\cdot\vec{r}_i} \cdot \widehat{\vec{p}}_i | \psi_\mathrm{H} \rangle \equiv \sum_{\ell=1}^{\infty} H_{21,i}^{\mathrm{E}\ell} + H_{21,i}^{\mathrm{M}\ell 1}$$

where

$$H_{21,i}^{\mathrm{E}\ell} = \mathrm{i} q_i k c A_0 \frac{(-\mathrm{i}k)^{\ell-1}}{\ell!} \langle \psi_\mathrm{L} | f r_{i,k}^{\ell-1} r_{i,\mathcal{E}} + [V/\hbar\omega, r_{i,k}^{\ell-1} r_{i,\mathcal{E}}] | \psi_\mathrm{H} \rangle$$

and

$$H_{21,i}^{\mathrm{M}\ell 1} = \mathrm{i} \frac{q_i}{m_i c} k c A_0 \frac{\ell(-\mathrm{i}k)^{\ell-1}}{(\ell+1)!} \langle \psi_\mathrm{L} | r_{i,k}^{\ell-1} \widehat{L}_{i,\mathcal{B}} | \psi_\mathrm{H} \rangle$$

Here A_0 is the magnitude of $\vec{A}_0$. Also $r_{i,k}$ is the component of the position $\vec{r}_i$ of particle i in the direction of motion of the electromagnetic wave. The direction of motion is the direction of $\vec{k}$. Similarly $r_{i,\mathcal{E}}$ is the component of $\vec{r}_i$ in the direction of the electric field. The electric field has the same direction as $\vec{A}_0$. Further, $L_{i,\mathcal{B}}$ is the component of the orbital angular momentum operator of particle i in the direction of the magnetic field. The magnetic field is in the same direction as $\vec{k} \times \vec{A}_0$. Finally, the factor f is

$$f = \pm 1 + \frac{\hbar\omega}{2 m_i c^2} \frac{\ell}{\ell + 2} \approx \pm 1$$

The approximation applies because normally the energy release in a transition is small compared to the rest mass energy of the particle. (And if it was not, the nonrelativistic electromagnetic interaction used here would not be valid in the first place.) For the emission process covered in {A.25}, the plus sign applies, $f = 1$.

The commutator is zero if the potential depends only on position. That is a valid approximation for electrons in atoms, but surely not for nuclei. For these it is a real problem, {N.14}.

For addendum {A.25}, the constant A_0 should be taken equal to $-\mathcal{E}_0/\sqrt{2}\mathrm{i}kc$. Note also that the interaction of the particle spin with the magnetic field still needs to be added to $H^{\mathrm{M}\ell 1}_{21,i}$. This interaction is unchanged from the naive approximation.

D.43.2 Matrix element for angular momentum modes

This subsection works out the details of the matrix element when angular momentum modes are used for the photon wave function.

The first matrix element to find is

$$H^{\mathrm{E}\ell 1}_{21,i} = -\frac{q_i}{m_i}\langle\psi_{\mathrm{L}}|A_0\vec{A}^{\mathrm{E}*}_{\gamma i}\cdot\widehat{\vec{p}}_i|\psi_{\mathrm{H}}\rangle$$

where, {A.21.7},

$$\vec{A}^{\mathrm{E}}_{\gamma i} = \nabla_i \times \vec{r}_i \times \nabla_i j_{\ell i} Y^m_{\ell i}$$

is the electric multipole vector potential at the location of particle i. This uses the short hand

$$j_{\ell i} \equiv j_\ell(kr_i) \qquad Y^m_{\ell i} \equiv Y^m_\ell(\theta_i,\phi_i)$$

where ℓ is the multipole order or photon angular momentum, k the photon wave number, j_ℓ a spherical Bessel function, and Y^m_ℓ a spherical harmonic.

Note that the electric multipole vector potential is closely related to the magnetic one:

$$\vec{A}^{\mathrm{E}}_{\gamma i} = \nabla_i \times \vec{A}^{\mathrm{M}}_{\gamma i} \qquad \vec{A}^{\mathrm{M}}_{\gamma i} = \vec{r}_i \times \nabla_i j_{\ell i} Y^m_{\ell i} = -\nabla_i \times j_{\ell i} Y^m_{\ell i}\vec{r}_i$$

The expression for the electric potential can be simplified for long photon wave lengths. Note first that

$$\nabla_i \times \vec{A}^{\mathrm{E}}_{\gamma i} = \nabla_i \times \nabla_i \times \vec{A}^{\mathrm{M}}_{\gamma i} = -\nabla_i^2 \vec{A}^{\mathrm{M}}_{\gamma i} = k^2 \vec{A}^{\mathrm{M}}_{\gamma i} = -k^2 \nabla_i \times j_{\ell i} Y^m_{\ell i}\vec{r}_i$$

where the second equality applied because the vector potentials are solenoidal and the standard vector identity (D.1), while the third equality is the energy eigenvalue problem, {A.21}. It follows that the electric vector potential is of the form

$$\vec{A}^{\mathrm{E}}_{\gamma i} = -k^2 j_{\ell i} Y^m_{\ell i}\vec{r}_i + \nabla_i F_i$$

because vector calculus says that if the curl of something is zero, it is the gradient of some scalar function F_i. Here

$$F_i = \int_{\underline{\vec{r}}_i=0}^{\vec{r}_i} [\vec{A}^{\mathrm{E}}_{\gamma i} + k^2 j_{\ell i} Y^m_{\ell i} \vec{r}_i] \cdot \mathrm{d}\underline{\vec{r}}_i$$

The direction of integration in the expression for F_i does not make a difference, so the simplest is to integrate radially outwards. The expression for $\vec{A}^{\mathrm{E}}_{\gamma i}$ was given in {D.36.2}. That gives

$$F_i = \int_{\underline{r}_i=0}^{r_i} [-l(l+1) + k^2 r^2] j_{\ell i} \frac{\mathrm{d}r}{r}\, Y^m_{\ell i}$$

Long photon wave length corresponds to small photon wave number k. All k^2 terms above can then be ignored and in addition the following approximation for the Bessel function applies, {A.6},

$$j_{\ell i} \approx \frac{(kr_i)^\ell}{(2\ell+1)!!}$$

This is readily integrated to find

$$F_i \approx -(\ell+1) \frac{(kr_i)^\ell}{(2\ell+1)!!} Y^m_{\ell i}$$

and $\vec{A}^{\mathrm{E}}_{\gamma i}$ is the gradient.

That allows lemma 1 to be used to find the electric matrix element.

$$\begin{aligned} H^{\mathrm{E}\ell 1}_{21,i} &= -\frac{q_i}{m_i} \langle \psi_{\mathrm{L}} | A_0 \vec{A}^{\mathrm{E}*}_{\gamma i} \cdot \widehat{\vec{p}}_i | \psi_{\mathrm{H}} \rangle \\ &\approx -\mathrm{i} q_i k c A_0 \frac{(\ell+1)k^\ell}{(2\ell+1)!!} \langle \psi_{\mathrm{L}} | r_i^\ell Y^{m*}_{\ell i} + [V/\hbar\omega, r_i^\ell Y^{m*}_{\ell i}] | \psi_{\mathrm{H}} \rangle \end{aligned}$$

This assumes ψ_{L} is indeed the lower-energy state. The value of A_0 (as defined here) to use in addendum {A.25} is $-\varepsilon^{\mathrm{E}}_k / \sqrt{2}\mathrm{i}kc$.

The commutator is again negligible for atoms, but a big problem for nuclei, {N.14}.

There is also a term due to the interaction of the spin with the magnetic field, given by the curl of $\vec{A}^{\mathrm{E}}_{\gamma i}$ as already found above,

$$H^{\mathrm{E}\ell 2}_{21,i} = -\frac{q_i}{m_i} \frac{g_i}{2} \langle \psi_{\mathrm{L}} | k^2 A_0 \vec{A}^{\mathrm{M}*}_{\gamma i} \cdot \widehat{\vec{S}}_i | \psi_{\mathrm{H}} \rangle = -\frac{q_i}{m_i} \frac{g_i}{2} \langle \psi_{\mathrm{L}} | k^2 A_0 (\vec{r}_i \times \nabla_i j_{\ell i} Y^{m*}_{\ell i}) \cdot \widehat{\vec{S}}_i | \psi_{\mathrm{H}} \rangle$$

Using the property of the scalar triple product that the factors can be interchanged if a minus sign is added, the matrix element becomes

$$H^{\mathrm{E}\ell 2}_{21,i} = \frac{q_i}{m_i} \frac{g_i}{2} k^2 A_0 \langle \psi_{\mathrm{L}} | (\nabla_i j_{\ell i} Y^{m*}_{\ell i}) \cdot (\vec{r}_i \times \widehat{\vec{S}}_i) | \psi_{\mathrm{H}} \rangle$$

(Note that ∇_i only acts on the $j_{\ell i} Y_{\ell i}^{m*}$; $\vec{A}_{\gamma i}^{\mathrm{M}*}$ is a function, not a differential operator.) In the long wave length approximation of the Bessel function, that becomes

$$H_{21,i}^{\mathrm{E}\ell 2} \approx q_i k c A_0 \frac{(\ell+1)k^\ell}{(2\ell+1)!!} \frac{\hbar\omega}{2(\ell+1)m_i c^2} \frac{g_i}{2} \langle \psi_{\mathrm{L}} | (\nabla_i r_i^\ell Y_{\ell i}^{m*}) \cdot (\vec{r}_i \times \frac{2}{\hbar} \widehat{\vec{S}}_i) | \psi_{\mathrm{H}} \rangle$$

The inner product should normally be of the same order as the one of $H_{21,i}^{\mathrm{E}\ell 1}$. However, the second fraction above is normally small; usually the photon energy is small compared to the rest mass energy of the particles. (And if it was not, the nonrelativistic electromagnetic interaction used here would not be valid in the first place.) So this second term will be ignored in addendum {A.25}.

The third matrix element to find is the magnetic multipole one

$$H_{21,i}^{\mathrm{M}\ell 1} = -\frac{q_i}{m_i} \langle \psi_{\mathrm{L}} | A_0 \vec{A}_{\gamma i}^{\mathrm{M}*} \cdot \widehat{\vec{p}}_i | \psi_{\mathrm{H}} \rangle$$

Note that in index notation

$$\vec{A}_{\gamma i}^{\mathrm{M}} \cdot \widehat{\vec{p}}_i = \sum_j r_{i,\bar{j}} (j_{\ell i} Y_{\ell i}^m)_{\bar{\bar{j}}} \widehat{p}_{i,j} - r_{i,\bar{\bar{j}}} (j_{\ell i} Y_{\ell i}^m)_{\bar{j}} \widehat{p}_{i,j}$$

where $\bar{j}$ follows j in the cyclic sequence $\ldots 123123 \ldots$ and $\bar{\bar{j}}$ precedes j. By a trivial renotation of the summation indices,

$$\vec{A}_{\gamma i}^{\mathrm{M}*} \cdot \widehat{\vec{p}}_i = \sum_j (j_{\ell i} Y_{\ell i}^{m*})_j r_{i,\bar{\bar{j}}} \widehat{p}_{i,\bar{j}} - (j_{\ell i} Y_{\ell i}^{m*})_j r_{i,\bar{j}} \widehat{p}_{i,\bar{\bar{j}}} = -(\nabla_i j_{\ell i} Y_{\ell i}^{m*}) \cdot \widehat{\vec{L}}_i$$

where $\widehat{\vec{L}}$ is the orbital angular momentum operator. Note that the parenthetical term commutes with this operator, something not mentioned in [33, p. 874].

It follows that

$$H_{21,i}^{\mathrm{M}\ell 1} = -\frac{q_i}{m_i} \langle \psi_{\mathrm{L}} | A_0 \vec{A}_{\gamma i}^{\mathrm{M}*} \cdot \widehat{\vec{p}}_i | \psi_{\mathrm{H}} \rangle = \frac{q_i}{m_i} A_0 \langle \psi_{\mathrm{L}} | (\nabla_i j_{\ell i} Y_{\ell i}^{m*}) \cdot \widehat{\vec{L}}_i | \psi_{\mathrm{H}} \rangle$$

or in the long wave length approximation

$$H_{21,i}^{\mathrm{M}\ell 1} = -\frac{q_i}{m_i} \langle \psi_{\mathrm{L}} | A_0 \vec{A}_{\gamma i}^{\mathrm{M}*} \cdot \widehat{\vec{p}}_i | \psi_{\mathrm{H}} \rangle \approx \frac{q_i}{m_i} A_0 \frac{k^\ell}{(2\ell+1)!!} \langle \psi_{\mathrm{L}} | (\nabla_i r_i^\ell Y_{\ell i}^{m*}) \cdot \widehat{\vec{L}}_i | \psi_{\mathrm{H}} \rangle$$

There is also a term due to the interaction of the spin with the magnetic field, given by the curl of $\vec{A}_{\gamma i}^{\mathrm{B}}$, which equals $\vec{A}_{\gamma i}^{\mathrm{E}}$,

$$H_{21,i}^{\mathrm{M}\ell 2} = -\frac{q_i}{m_i} \frac{g_i}{2} \langle \psi_{\mathrm{L}} | A_0 \vec{A}_{\gamma i}^{\mathrm{E}*} \cdot \widehat{\vec{S}}_i | \psi_{\mathrm{H}} \rangle$$

Using the same long wave length approximation for $\vec{A}_{\gamma i}^{\mathrm{E}}$ as before, that becomes

$$H_{21,i}^{\mathrm{M}\ell 2} \approx \frac{q_i}{m_i} A_0 \frac{(\ell+1)k^\ell}{(2\ell+1)!!} \frac{g_i}{2} \langle \psi_{\mathrm{L}} | (\nabla_i r_i^\ell Y_{\ell i}^{m*}) \cdot \widehat{\vec{S}}_i | \psi_{\mathrm{H}} \rangle$$

The orbital and spin matrix elements may be combined into one as

$$H^{\mathrm{M}\ell}_{21,i} \approx \frac{q_i}{2m_i} A_0 \frac{(\ell+1)k^\ell}{(2\ell+1)!!} \langle \psi_{\mathrm{L}} | (\nabla_i r_i^\ell Y_{\ell i}^{m*}) \cdot \Big(\frac{2}{\ell+1} \widehat{\vec{L}}_i + g_i \widehat{\vec{S}}_i \Big) | \psi_{\mathrm{H}} \rangle$$

The value of A_0 to use in addendum {A.25} is $-\varepsilon_k^{\mathrm{E}}/\sqrt{2}\mathrm{i}c$.

D.43.3 Weisskopf and Moszkowski estimates

This subsection explains where the radial, angular, and momentum factors in the Weisskopf and Moszkowski estimates come from. These factors represent the nondimensionalized matrix elements.

The electric matrix element is simplest. It is, written out in spherical coordinates using the assumed wave functions,

$$|h^{\mathrm{E}\ell}_{21}| \approx \int R_{\mathrm{L}}(r_i)^* (r_i/R)^\ell R_{\mathrm{H}}(r_i) r_i^2 \, \mathrm{d}r_i \;\; \sqrt{4\pi} \int \Theta^{m_{j\mathrm{L}}*}_{l_{\mathrm{L}} j_{\mathrm{L}} i} Y_{\ell i}^{m*} \Theta^{m_{j\mathrm{H}}}_{l_{\mathrm{H}} j_{\mathrm{H}} i} \sin^2\theta_i \mathrm{d}\theta_i \mathrm{d}\phi_i$$

The Weisskopf and Moszkowski estimates assume that the radial parts of wave functions equal a constant C until the nuclear edge R and are zero outside the nucleus. To perform the radial integral is then straightforward:

$$\int R_{\mathrm{L}}(r_i)^* (r_i/R)^\ell R_{\mathrm{H}}(r_i) r_i^2 \, \mathrm{d}r_i = \frac{\int_0^R C^2 (r_i/R)^\ell r_i^2 \, \mathrm{d}r_i}{\int_0^R C^2 r_i^2 \, \mathrm{d}r_i} = \frac{3}{\ell+3}$$

The first equality is true because the integral in the denominator is 1 on account of the normalization condition of wave functions. The second inequality follows from integrating.

The angular integral above is more tricky to ballpark. First of all, it will be assumed that the matrix element of interest is the lowest multipole order allowed by angular momentum conservation. That seems reasonable, given that normally higher multipole transitions will be very much slower. It follows that $\ell = |j_{\mathrm{H}} - j_{\mathrm{L}}|$. (The possibility that the initial and final angular momenta are equal will be ignored.)

The change in orbital angular momenta could in principle be up to one unit different from the change in net angular momenta because of the spins. But parity conservation allows only $|l_{\mathrm{H}} - l_{\mathrm{L}}| = \ell$.

To simplify even further, assume the following specific angular states:

$$\Theta^{m_{j\mathrm{L}}}_{l_{\mathrm{L}} j_{\mathrm{L}} i} = Y_{0i}^0 \uparrow \qquad \Theta^{m_{j\mathrm{H}}}_{l_{\mathrm{H}} j_{\mathrm{H}} i} = Y_{\ell i}^\ell \uparrow$$

which have

$$l_{\mathrm{L}} = 0 \quad j_{\mathrm{L}} = \tfrac{1}{2} \quad m_{j\mathrm{L}} = \tfrac{1}{2} \qquad l_{\mathrm{H}} = \ell \quad j_{\mathrm{H}} = \ell + \tfrac{1}{2} \quad m_{j\mathrm{H}} = \ell + \tfrac{1}{2}$$

If these states are substituted into the angular integral, the product of the spin states is 1 because spin states are orthonormal. What is left is

$$\sqrt{4\pi}\int Y_{0i}^{0*}Y_{\ell i}^{m*}Y_{\ell i}^{\ell}\sin^2\theta_i\mathrm{d}\theta_i\mathrm{d}\phi_i$$

Now $Y_0^0 = 1/\sqrt{4\pi}$ which is just a constant that can be taken out of the integral. There it cancels the corresponding square root in the definition of the matrix element. Then it is seen that the transition can only create a photon for which $m = \ell$. The reason is that spherical harmonics are orthonormal; the inner product is only nonzero if the two spherical harmonics are equal, and then it is 1. So the conclusion is that for the given states

$$\sqrt{4\pi}\int \Theta_{l_{\rm L}j_{\rm L}i}^{m_{j\rm L}*}Y_{\ell i}^{m*}\Theta_{l_{\rm H}j_{\rm H}i}^{m_{j\rm H}}\sin^2\theta_i\mathrm{d}\theta_i\mathrm{d}\phi_i = 1$$

The angular integral is 1. That makes the decay rate exactly 1 Weisskopf unit.

One glaring deficiency in the above analysis was the assumption that the initial proton state was a $Y_\ell^\ell\uparrow$ one. It would certainly be reasonable to have an initial nuclear state that has orbital angular momentum $l_{\rm H} = \ell$ and total angular momentum $j_{\rm H} = \ell + \frac{1}{2}$. But a bunch of these nuclei would surely each be oriented in its own random direction. So they would have different magnetic quantum numbers $m_{j\rm H}$. They would not all have $m_{j\rm H} = \ell + \frac{1}{2}$.

Fortunately, it turns out that this makes no difference. For example, by symmetry the state $Y_\ell^{-\ell}\downarrow$ decays just as happily to $Y_0^0\downarrow$ as $Y_\ell^\ell\uparrow$ does to $Y_0^0\uparrow$. For other values of $m_{j\rm H}$ it is a bit more nuanced. They produce an initial state of the form:

$$\Theta_{l_{\rm H}j_{\rm H}i}^{m_{j\rm H}} = \Theta_{\ell\,\ell+\frac{1}{2}\,i}^{m_j} = c_1 Y_\ell^{m_{j\rm H}-\frac{1}{2}}\uparrow + c_2 Y_\ell^{m_{j\rm H}+\frac{1}{2}}\downarrow$$

Now the first term produces decays to $Y_0^0\uparrow$ by the emission of a photon with $m_\ell = \ell - \frac{1}{2}$. However, because of the factor c_1 the number of such decays that occur per second is a factor c_1^2 less than the Weisskopf unit. But the second term produces decays to $Y_0^0\downarrow$ by the emission of a photon with $m_\ell = \ell + \frac{1}{2}$. This decay rate is a factor c_2^2 less than the Weisskopf unit. Since $c_1^2 + c_2^2 = 1$, (the normalization condition of the state), the total decay rate is still 1 Weisskopf unit.

So as long as the final state $\psi_{\rm L}$ has zero orbital angular momentum, the decay is at 1 Weisskopf unit. The orientation of the initial state makes no difference. That is reflected in table A.3. This table lists the angular factors to be applied to the Weisskopf unit to get the actual decay rate. The first row shows that, indeed, when the final angular momentum is ½, as occurs for zero angular momentum, and the initial angular momentum is $\ell + \frac{1}{2}$, then no correction is needed. The correction factor is 1.

More interesting is the possibility that the two states are swapped. Then the initial state is the one with zero orbital angular momentum. It might at first

seem that that will not make a difference either. After all, decay rates between *specific* states are exactly the same.

But there is in fact a difference. Previously, each initial nucleus had only two states to decay to: the spin-up and the spin-down version of the final state. Now however, each initial nucleus has $2j_{\rm L}+1$, i.e. $2\ell+2$ final states it can decay to, corresponding to the possible values of the final magnetic quantum number $m_{\rm L}$. That will increase the total decay rate correspondingly. In fact, suppose that the initial nuclei come in spin-up and spin-down pairs. Then each pair will decay at a rate of one Weisskopf unit to each possible final state. That is because this picture is the exact reverse of the decay of the final state. So the pairs would decay at a rate $2\ell+2$ faster than the Weisskopf unit. So by symmetry each nucleus of the pair decays $\ell+1$ times faster than the Weisskopf unit. That is reflected in the first column of table A.3. (Recall that ℓ is the difference in the j values.)

If neither the initial nor final state has zero orbital angular momentum, it gets more messy. Figuring out the correction factor in that case is something for those who love abstract mathematics.

Next consider magnetic multipole transitions. They are much messier to ballpark. It will again be assumed that the multipole order is the smallest possible. Unfortunately, now the final orbital angular momentum cannot be zero. Because of parity, that would require that the initial orbital angular momentum would be $\ell+1$. But that is too large because of the limitation (A.175) on the orbital angular momentum change in magnetic transitions. Therefore the simplest possible initial and final states have

$$l_{\rm L}=1 \quad j_{\rm L}=\tfrac{1}{2} \quad m_{j{\rm L}}=\tfrac{1}{2} \qquad l_{\rm H}=\ell \quad j_{\rm H}=\ell+\tfrac{1}{2} \quad m_{j{\rm H}}=\ell+\tfrac{1}{2}$$

For these quantum numbers, the initial and final states are

$$\psi_{\rm L}=R_{{\rm L},i}\Theta^{m_{j{\rm L}}}_{l_{\rm L}j_{\rm L}i}=R_{{\rm L},i}\left(\sqrt{\tfrac{2}{3}}Y^1_{1i}\downarrow-\sqrt{\tfrac{1}{3}}Y^0_{1i}\uparrow\right) \qquad \psi_{\rm H}=R_{{\rm H},i}\Theta^{m_{j{\rm H}}}_{l_{\rm H}j_{\rm H}i}=R_{{\rm H},i}Y^\ell_{\ell i}\uparrow$$

where the square roots come from figure 12.5 in the $j_a, j_b = 1, \frac{1}{2}$ tabulation.

Now consider the form of the magnetic matrix element (A.181). First note, {D.43.2}, that the angular momentum and gradient factors commute. That helps because then the angular momentum operators, being Hermitian, can be applied on the easier state $\psi_{\rm L}$.

The z-component part of the dot product in the matrix element is then the easiest. The z components of the angular momentum operators leave the state $\psi_{\rm L}$ essentially unchanged. They merely multiply the two terms by the eigenvalue $m_l\hbar$ respectively $m_s\hbar$.

Next, this gets multiplied by the z-component of the gradient. But multiplying by the gradient cannot change the spin. So the spin-down first term in $\psi_{\rm L}$ stays spin-down. That cannot match the spin-up of $\psi_{\rm H}$. So the first term does not produce a contribution.

The second term in $\psi_{\rm L}$ has the right spin. Since spin states are orthonormal, their inner product produces 1. But now there is a problem of matching the magnetic quantum number of $\psi_{\rm H}$. In particular, consider the harmonic polynomial $r^\ell Y_\ell^{m_l}$ in the gradient. The gradient reduces it to a combination of harmonic polynomials of one degree less, in other words, to $r^{\ell-1}Y_{\ell-1}^{m_l}$ polynomials. That limits m_l to a value no larger than $\ell-1$, and since the second term in $\psi_{\rm L}$ has magnetic quantum number 0, the value ℓ in $\psi_{\rm H}$ cannot be matched. The bottom line is that the z-component terms in the inner product of the matrix element do not produce a contribution.

However, the x- and y-component terms are another story. The angular momentum operators in these directions change the corresponding magnetic quantum numbers, chapter 12.11. In general, their application produces a mixture of $m+1$ and $m-1$ states. In particular, the x and y components of spin will produce a spin-up version of the first term in $\psi_{\rm L}$. That now matches the spin in $\psi_{\rm H}$ and a nonzero contribution results. Similarly, the orbital angular momentum operators will produce an $m_{\rm L}=1$ version of the second term in $\psi_{\rm L}$. Combined with the $\ell-1$ units from the gradient, that is enough to match the magnetic quantum number of $\psi_{\rm H}$. So there is a total of four nonzero contributions to the matrix element.

Now it is just a matter of working out the details to get the complete matrix element. The information in chapter 12.11 can be used to find the exact states produced from $\Theta^{m_{j\rm H}}_{l_{\rm H}j_{\rm H}i}$ by the x and y angular momentum operators. Each state is a multiple of the $Y_1^1\uparrow$ state. As far as the gradient term is concerned, the harmonic polynomials are of the general form

$$r^\ell Y_\ell^\ell = C(x+{\rm i}y)^\ell \qquad r^\ell Y_\ell^{\ell-1} = Dz(x+{\rm i}y)^{\ell-1} \qquad \ldots$$

as seen in table 4.3 or {D.64}. The constants $C, D, \ldots$ are of no importance here. The x and y derivatives of the first harmonic polynomial will give the needed $Y_{\ell-1}^{\ell-1}$ harmonic. (For values of ℓ greater than 1, the third harmonic could also make a contribution. However, it turns out that here the x and y contributions cancel each other.) The effect of the x-derivative on the first harmonic is simply to add a factor $\ell/(x+{\rm i}y)$ to it. Similarly, the y-derivative simply adds a factor ${\rm i}\ell/(x+{\rm i}y)$. Now if you look up Y_1^1 in table 4.3, you see it is a multiple of $x+{\rm i}y$. So the product with the gradient term produces a simple multiple of $Y_\ell^\ell\uparrow$. The inner product with $\psi_{\rm H}$ then produces that multiple (which still depends on r_i of course.) Identifying and adding the four multiples produces

$$h_{21}^{{\rm M}\ell} = -\left(g_i\ell - \frac{2\ell}{\ell+1}\right)\int R_{\rm L}(r_i)^*(r_i/R)^{\ell-1}R_{\rm H}(r_i)r_i^2\,{\rm d}r_i$$

The remaining radial integral may be ballparked exactly the same as for the electric case. The only difference is that the power of r_i is one unit smaller.

A similar analysis shows that the given initial state cannot decay to the version of the final state with negative magnetic quantum number $m_{j\rm L}=-\frac{1}{2}$.

And of course, if the initial and final states are swapped, there is again a factor $\ell + 1$ increase in decay rate.

More interestingly, the same expression turns out to hold if neither the initial nor the final angular momentum equals $\frac{1}{2}$, using the correction factor of table A.3. But the obtained magnetic multipole decay rate is more limited than the electric one. It does require that $|j_{\rm H} - j_{\rm L}| = \ell$ and that $|l_{\rm H} - l_{\rm L}| = \ell - 1$

The momentum factors (A.189) were identified using a computer program. This program crunched out the complete matrix elements using procedures exactly like the ones above. This program was also used to create table A.3 of angular factors. This guards against typos and provides an independent check on the Clebsch-Gordan values.

D.44 Derivation of group velocity

The objective of this note is to derive the wave function for a wave packet if time is large.

To shorten the writing, the Fourier integral (7.64) for Ψ will be abbreviated as:

$$\Psi = \int_{k_1}^{k_2} f(k) e^{{\rm i}\varphi t} \, {\rm d}k \qquad \varphi = k\frac{x}{t} - \omega \quad \varphi' = \frac{x}{t} - v_{\rm g} \quad \varphi'' = -v_{\rm g}'$$

where it will be assumed that φ is a well behaved functions of k and f at least twice continuously differentiable. Note that the wave number k_0 at which the group velocity equals x/t is a stationary point for φ. That is the key to the mathematical analysis.

The so-called "method of stationary phase" says that the integral is negligibly small as long as there are no stationary points $\varphi' = 0$ in the range of integration. Physically that means that the wave function is zero at large time positions that cannot be reached with any group velocity within the range of the packet. It therefore implies that the wave packet propagates with the group velocity, within the variation that it has.

To see why the integral is negligible if there are no stationary points, just integrate by parts:

$$\Psi = \frac{f(k)}{{\rm i}\varphi' t} e^{{\rm i}\varphi t}\Big|_{k_1}^{k_2} - \int_{k_1}^{k_2} \left(\frac{f(k)}{{\rm i}\varphi' t}\right)' e^{{\rm i}\varphi t} \, {\rm d}k$$

This is small of order $1/t$ for large times. And if $\overline{\Phi}_0(p)$ is chosen to smoothly become zero at the edges of the wave packet, rather than abruptly, you can keep integrating by parts to show that the wave function is much smaller still. That is important if you have to plot a wave packet for some book on quantum mechanics and want to have its surroundings free of visible perturbations.

For large time positions with x/t values within the range of packet group velocities, there will be a stationary point to φ. The wave number at the stationary point will be indicated by k_0, and the value of φ and its second derivative by φ_0 and φ_0''. (Note that the second derivative is minus the first derivative of the group velocity, and will be assumed to be nonzero in the analysis. If it would be zero, nontrivial modifications would be needed.)

Now split the exponential in the integral into two,

$$\Psi = e^{\mathrm{i}\varphi_0 t} \int_{k_1}^{k_2} f(k) e^{\mathrm{i}(\varphi-\varphi_0)t} \,\mathrm{d}k$$

It is convenient to write the difference in φ in terms of a new variable $\overline{k}$:

$$\varphi - \varphi_0 = \tfrac{1}{2}\varphi_0'' \overline{k}^2 \qquad \overline{k} \sim k - k_0 \quad \text{for} \quad k \to k_0$$

By Taylor series expansion it can be seen that $\overline{k}$ is a well behaved monotonous function of k. The integral becomes in terms $\overline{k}$:

$$\Psi = e^{\mathrm{i}\varphi_0 t} \int_{\overline{k}_1}^{\overline{k}_2} g(\overline{k}) e^{\mathrm{i}\frac{1}{2}\varphi_0''\overline{k}^2 t} \,\mathrm{d}\overline{k} \qquad g(\overline{k}) = f(k)\frac{\mathrm{d}k}{\mathrm{d}\overline{k}}$$

Now split function g apart as in

$$g(\overline{k}) = g(0) + [g(\overline{k}) - g(0)]$$

The part within brackets produces an integral

$$e^{\mathrm{i}\varphi_0 t} \int_{\overline{k}_1}^{\overline{k}_2} \frac{g(\overline{k}) - g(0)}{\mathrm{i}\varphi_0''\overline{k}t} \quad \mathrm{i}\varphi_0''\overline{k}t e^{\mathrm{i}\frac{1}{2}\varphi_0''\overline{k}^2 t} \,\mathrm{d}\overline{k}$$

and integration by parts shows that to be small of order $1/t$.

That leaves the first part, $g(0) = f(k_0)$, which produces

$$\Psi = e^{\mathrm{i}\varphi_0 t} f(k_0) \int_{\overline{k}_1}^{\overline{k}_2} e^{\mathrm{i}\frac{1}{2}\varphi_0''\overline{k}^2 t} \,\mathrm{d}\overline{k}$$

Change to a new integration variable

$$u \equiv \sqrt{\frac{|\varphi_0''|t}{2}}\overline{k}$$

Note that since time is large, the limits of integration will be approximately $u_1 = -\infty$ and $u_2 = \infty$ unless the stationary point is right at an edge of the wave packet. The integral becomes

$$\Psi = e^{\mathrm{i}\varphi_0 t} f(k_0) \sqrt{\frac{2}{|\varphi_0''|t}} \int_{u_1}^{u_2} e^{\pm \mathrm{i}u^2} \,\mathrm{d}u$$

where $\pm$ is the sign of φ_0''. The remaining integral is a "Fresnel integral" that can be looked up in a table book. Away from the edges of the wave packet, the integration range can be taken as all u, and then

$$\Psi = e^{\mathrm{i}\varphi_0 t} e^{\pm \mathrm{i}\pi/4} f(k_0) \sqrt{\frac{2\pi}{|\varphi_0''| t}}$$

Convert back to the original variables and there you have the claimed expression for the large time wave function.

Right at the edges of the wave packet, modified integration limits for u must be used, and the result above is not valid. In particular it can be seen that the wave packet spreads out a distance of order $\sqrt{t}$ beyond the stated wave packet range; however, for large times $\sqrt{t}$ is small compared to the size of the wave packet, which is proportional to t.

For the mathematically picky: the treatment above assumes that the wave packet momentum range is not small in an asymptotic sense, (i.e. it does not go to zero when t becomes infinite.) It is just small in the sense that the group velocity must be monotonous. However, Kaplun's extension theorem implies that the packet size can be allowed to become zero at least slowly. And the analysis is readily adjusted for faster convergence towards zero in any case.

D.45 Motion through crystals

This note derives the semi-classical motion of noninteracting electrons in crystals. The derivations will be one-dimensional, but the generalization to three dimensions is straightforward.

D.45.1 Propagation speed

The first question is the speed with which a more or less localized electron moves. An electron in free space moves with a speed found by dividing its linear momentum by its mass. However, in a solid, the energy eigenfunctions are Bloch waves and these do not have definite momentum.

Fortunately, the analysis for the wave packet of a free particle is virtually unchanged for a particle whose energy eigenfunctions are Bloch waves instead of simple exponentials. In the Fourier integral (7.64), simply add the periodic factor $\psi^{\mathrm{p}}_{\mathrm{p},k}(x)$. Since this factor is periodic, it is bounded, and it plays no part in limit process of infinite time. (You can restrict the times in the limit process to those at which x is always at the same position in the period.)

As a result the group velocity is again $\mathrm{d}\omega/\mathrm{d}k$. Since the energy is $E^{\mathrm{p}} = \hbar\omega$ and the crystal momentum $p_{\mathrm{cm}} = \hbar k$, the velocity of a localized electron can be written as

$$v = \frac{\mathrm{d}E^{\mathrm{p}}}{\mathrm{d}p_{\mathrm{cm}}}$$

In the absence of external forces, the electron will keep moving with the same velocity for all time. The large time wave function is

$$\Psi(x,t) \sim \frac{e^{\mp \mathrm{i}\pi/4}}{\sqrt{|v'_{\mathrm{g}0}|t}} \overline{\Phi}_0(k_0) \psi^{\mathrm{p}}_{\mathrm{p},\mathrm{k}_0}(x) e^{\mathrm{i}(k_0 x - \omega_0 t)} \qquad v_{\mathrm{g}0} = \frac{x}{t}$$

where k_0 is the wave number at which the group speed equals x/t. Note that the wave function looks locally just like a single Bloch wave for large time.

D.45.2 Motion under an external force

The acceleration due to an external force on an electrons is not that straightforward. First of all note that you cannot just add a constant external force. A constant force F_{ext} would produce an external potential of the form $V_{\mathrm{ext}} = -F_{\mathrm{ext}} x$ and that becomes infinite at infinite x. However, it can be assumed that the force is constant over the nonzero range of the wave packet.

Next there is a trick. Consider the expectation value $\langle \mathcal{T}_d \rangle$ of the translation operator $\mathcal{T}_d$ that translates the wave function over one atomic cell size d. If the wave packet consisted of just a single Bloch wave with wave number k_0, the expectation value of $\mathcal{T}_d$ would be $e^{\mathrm{i}k_0 d}$. A wave packet must however include a small range of k values. Then $\langle \mathcal{T}_d \rangle$ will be an *average* of $e^{\mathrm{i}kd}$ values over the k values of the wave packet. Still, if the range of k values is small enough, you can write

$$\langle \mathcal{T}_d \rangle = A e^{\mathrm{i}k_0 d}$$

where k_0 is a k value somewhere in the middle of the wave packet and A is a real number close to one. So $\langle \mathcal{T}_d \rangle$ still gives the typical k value in the wave packet.

Moreover, its magnitude $|\langle \mathcal{T}_d \rangle| = A$ is always less than one and the closer it is to one, the more compact the wave packet. That is because $\langle \mathcal{T}_d \rangle$ is an average of $e^{\mathrm{i}kd}$ values. These are all located on the unit circle in the complex plane, the plane with $\cos(kd)$ as the horizontal axis and $\sin(kd)$ as the vertical axis. If the wave packet would consist of just a single k value k_0, then the average of $e^{\mathrm{i}kd}$ would be exactly $e^{\mathrm{i}k_0 d}$, and be on the unit circle. If however the wave numbers spread out a bit around k_0, then the average moves inside the unit circle: if you average positions on a circle, the average is always inside the circle. In the extreme case that the k values get uniformly distributed over the entire circle, the average position is at the origin. That would make $|\langle \mathcal{T}_d \rangle|$ zero. Conversely, as long as $|\langle \mathcal{T}_d \rangle|$ stays very close to one, the wave packet must be very compact in terms of k.

The time evolution of $\langle \mathcal{T}_d \rangle$ can be found using chapter 7.2:

$$\frac{\mathrm{d}\langle \mathcal{T}_d \rangle}{\mathrm{d}t} = \frac{\mathrm{i}}{\hbar} \langle [H_0 + V_{\mathrm{ext}}, \mathcal{T}_d] \rangle \tag{D.28}$$

where H_0 is the Hamiltonian for the electron in the crystal, and $V_{\rm ext}$ the additional external potential. Now the commutator of H_0 and $\mathcal{T}_d$ is zero; the crystal Hamiltonian acts exactly the same on the wave function whether it is shifted one cell over or not. The remainder of the commutator gives, when applied on an arbitrary wave function,

$$[V_{\rm ext}, \mathcal{T}_d]\Psi \equiv V_{\rm ext}\mathcal{T}_d\Psi - \mathcal{T}_d V_{\rm ext}\Psi$$

Writing this out with the arguments of the functions explicitly shown gives:

$$V_{\rm ext}(x)\Psi(x+d) - V_{\rm ext}(x+d)\Psi(x+d) = (V_{\rm ext}(x) - V_{\rm ext}(x+d))\mathcal{T}_d\Psi(x)$$

Now assume that the external force $F_{\rm ext}$ is constant over the extent of the wave packet. In that case the difference in the potentials is just $F_{\rm ext}d$, and that is a constant that can be taken out of the expectation value of the commutator. So:

$$\frac{{\rm d}\langle\mathcal{T}_d\rangle}{{\rm d}t} = \frac{\rm i}{\hbar}F_{\rm ext}d\langle\mathcal{T}_d\rangle \tag{D.29}$$

The solution to this equation is:

$$\langle\mathcal{T}_d\rangle = \langle\mathcal{T}_d\rangle_0 e^{{\rm i}F_{\rm ext}dt/\hbar}$$

where $\langle\mathcal{T}_d\rangle_0$ is the value of $\langle\mathcal{T}_d\rangle$ at the starting time $t = 0$.

It follows that the magnitude of the $\langle\mathcal{T}_d\rangle$ does not change with time. In view of the earlier discussion, this means that the wave packet maintains its compactness in terms of k. (In physical space the wave packet will gradually spread out, as can be seen from the form of the large-time wave function given earlier.)

It further follows that the average wave number k_0 in the wave packet evolves as:

$$\frac{{\rm d}\hbar k_0}{{\rm d}t} = F_{\rm ext}$$

Since the packet remains compact, all wave numbers in the wave packet change the same way. This is Newton's second law in terms of crystal momentum.

D.45.3 Free-electron gas with constant electric field

This book discussed the effect of an applied electric field on free electrons in a periodic box in chapter 6.20. The effect was described as a change of the velocity of the electrons. Since the velocity is proportional to the wave number for free electrons, the velocity change corresponds to a change in the wave number. In this subsection the effect of the electric field will be examined in more detail. The solution will again be taken to be one-dimensional, but the extension to three dimensions is trivial.

Assume that a constant electric field is applied, so that the electrons experience a constant force $F_{\rm ext}$. The time-dependent Schrödinger equation is

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} - F_{\rm ext}x\Psi$$

Assume the initial condition to be

$$\Psi_0 = \sum_{k_0} c_{k_0} e^{\mathrm{i}k_0 x}$$

in which a subscript 0 indicates the initial time.

The exact solution to this problem is

$$\Psi = \sum_{k_0} c(k_0,t) e^{\mathrm{i}k(t)x} \qquad \frac{\mathrm{d}\hbar k}{\mathrm{d}t} = F_{\rm ext}$$

where the magnitude of the coefficients $|c(k_0,t)| = |c_{k_0}|$ is independent of time. This exact solution is in terms of states $e^{\mathrm{i}k(t)x}$ that change in time. The probability of the particle being in those states does not change.

Unfortunately, this solution is only periodic with period equal to the length of the box ℓ for times at which $F_{\rm ext}t/\hbar$ happens to be a whole multiple of the wave number spacing. At those times the Fermi sphere of occupied states has shifted the same whole multiple of wave number spacings to the right.

At intermediate times, the solution is not periodic, so it cannot be correctly described using the periodic box modes. The magnitude of the wave function is still periodic. However, the momentum has values inconsistent with the periodic box. The problem is that even though a constant force is periodic, the corresponding potential is not. Since quantum mechanics uses the potential instead of the force, the quantum solution is no longer periodic.

The problem goes away by letting the periodic box size become infinite. But that brings back the ugly normalization problems. For a periodic box, the periodic boundary conditions will need to be relaxed during the application of the electric field. In particular, a factor $e^{\mathrm{i}F_{\rm ext}\ell t/\hbar}$ difference in wave function and its x-derivative must be allowed between the ends of the box. Since the periodic boundary conditions are artificial anyway for modeling a piece of electrical wire, this may not be a big concern. In any case, for a big-enough periodic box, the times at which the solution returns to its original periodicity become spaced very close together.

D.46 Derivation of the WKB approximation

The purpose in this note is to derive an approximate solution to the Hamiltonian eigenvalue problem

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = -\frac{p_{\rm c}^2}{\hbar^2}\psi$$

where the classical momentum $p_c = \sqrt{2m(E-V)}$ is a known function for given energy. The approximation is to be valid when the values of $p_c/\hbar$ are large. In quantum terms, you can think of that as due to an energy that is macroscopically large. But to do the mathematics, it is easier to take a macroscopic point of view; in macroscopic terms, $p_c/\hbar$ is large because Planck's constant $\hbar$ is so small.

Since either way $p_c/\hbar$ is a large quantity, for the left hand side of the Hamiltonian eigenvalue problem above to balance the right hand side, the wave function must vary rapidly with position. Something that varies rapidly and nontrivially with position tends to be hard to analyze, so it turns out to be a good idea to write the wave function as an exponential,

$$\psi = e^{\mathrm{i}\tilde{\theta}}$$

and then approximate the argument $\tilde{\theta}$ of that exponential.

To do so, first the equation for $\tilde{\theta}$ will be needed. Taking derivatives of ψ using the chain rule gives in terms of $\tilde{\theta}$

$$\frac{\mathrm{d}\psi}{\mathrm{d}x} = e^{\mathrm{i}\tilde{\theta}}\,\mathrm{i}\frac{\mathrm{d}\tilde{\theta}}{\mathrm{d}x} \qquad \frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = -e^{\mathrm{i}\tilde{\theta}}\left(\frac{\mathrm{d}\tilde{\theta}}{\mathrm{d}x}\right)^2 + e^{\mathrm{i}\tilde{\theta}}\,\mathrm{i}\frac{\mathrm{d}^2\tilde{\theta}}{\mathrm{d}x^2}$$

Then plugging ψ and its second derivative above into the Hamiltonian eigenvalue problem and cleaning up gives:

$$\left(\frac{\mathrm{d}\tilde{\theta}}{\mathrm{d}x}\right)^2 = \frac{p_c^2}{\hbar^2} + \mathrm{i}\frac{\mathrm{d}^2\tilde{\theta}}{\mathrm{d}x^2} \tag{D.30}$$

For a given energy, $\tilde{\theta}$ will depend on both what x is and what $\hbar$ is. Now, since $\hbar$ is small, mathematically it simplifies things if you expand $\tilde{\theta}$ in a power series with respect to $\hbar$:

$$\tilde{\theta} = \frac{1}{\hbar}\left(f_0 + \hbar f_1 + \tfrac{1}{2}\hbar^2 f_2 + \ldots\right)$$

You can think of this as writing $\hbar\theta$ as a Taylor series in $\hbar$. The coefficients $f_0, f_1, f_2, \ldots$ will depend on x. Since $\hbar$ is small, the contribution of f_2 and further terms to ψ is small and can be ignored; only f_0 and f_1 will need to be figured out.

Plugging the power series into the equation for $\tilde{\theta}$ produces

$$\frac{1}{\hbar^2}f_0'^2 + \frac{1}{\hbar}2f_0'f_1' + \ldots = \frac{1}{\hbar^2}p_c^2 + \frac{1}{\hbar}\mathrm{i}f_0'' + \ldots$$

where primes denote x-derivatives and the dots stand for powers of $\hbar$ greater than $\hbar^{-1}$ that will not be needed. Now for two power series to be equal, the

coefficients of each individual power must be equal. In particular, the coefficients of $1/\hbar^2$ must be equal, $f_0'^2 = p_{\rm c}^2$, so there are two possible solutions

$$f_0' = \pm p_{\rm c}$$

For the coefficients of $1/\hbar$ to be equal, $2f_0'f_1' = {\rm i}f_0''$, or plugging in the solution for f_0',

$$f_1' = {\rm i}\frac{p_{\rm c}'}{2p_{\rm c}}$$

It follows that the x-derivative of $\tilde{\theta}$ is given by

$$\tilde{\theta}' = \frac{1}{\hbar}\left(\pm p_{\rm c} + \hbar{\rm i}\frac{p_{\rm c}'}{2p_{\rm c}} + \ldots\right)$$

and integrating gives $\tilde{\theta}$ as

$$\tilde{\theta} = \pm\frac{1}{\hbar}\int p_{\rm c}\,{\rm d}x + {\rm i}\tfrac{1}{2}\ln p_{\rm c} + \tilde{C}\ldots$$

where $\tilde{C}$ is an integration constant. Finally, $e^{{\rm i}\tilde{\theta}}$ now gives the two terms in the WKB solution, one for each possible sign, with $e^{{\rm i}\tilde{C}}$ equal to the constant $C_{\rm f}$ or $C_{\rm b}$.

D.47 Born differential cross section

This note derives the Born differential cross section of addendum {A.30}.

The general idea is to approximate (A.228) for large distances r. Then the asymptotic constant $C_{\rm f}$ in (A.216) can be identified, which gives the differential cross section according to (A.218). Note that the Born approximation took the asymptotic constant $C_{\rm f}^{\rm l}$ equal to one for simplicity.

The main difficulty in approximating (A.228) for large distances r is the argument of the exponential in the fraction. It is not accurate enough to just say that $|\vec{r}-\vec{r}\,'|$ is approximately equal to r. You need the more accurate approximation

$$|\vec{r}-\vec{r}\,'| = \sqrt{(\vec{r}-\vec{r}\,')\cdot(\vec{r}-\vec{r}\,')} = \sqrt{r^2 - 2\vec{r}\cdot\vec{r}\,' + \vec{r}\,'\cdot\vec{r}\,'} \sim r - \frac{\vec{r}}{r}\cdot\vec{r}\,'$$

The final approximation is from taking a factor r^2 out of the square root and then approximating the rest by a Taylor series. Note that the fraction in the final term is the unit vector $\hat{\imath}_r$ in the r-direction.

It follows that

$$\frac{e^{{\rm i}p_\infty|\vec{r}-\vec{r}\,'|/\hbar}}{|\vec{r}-\vec{r}\,'|} \sim \frac{e^{{\rm i}p_\infty r/\hbar}}{r}e^{-{\rm i}\vec{p}_\infty\cdot\vec{r}\,'/\hbar} \qquad \vec{p}_\infty = p_\infty\hat{\imath}_r$$

Also, in the second exponential, since $z' \equiv \hat{k} \cdot \vec{r}'$,

$$e^{\mathrm{i}p_\infty z'/\hbar} = e^{\mathrm{i}\vec{p}_\infty^{\,1}\cdot\vec{r}\,'/\hbar} \qquad \vec{p}_\infty^{\,1} = p_\infty \hat{k}$$

Writing out the complete expression (A.228) and comparing with (A.216) gives the constant $C_{\rm f}$ and hence the differential cross section.

D.48 About Lagrangian multipliers

This note will derive the Lagrangian multipliers for an example problem. Only calculus will be used. The example problem will be to find a stationary point of a function f of four variables if there are two constraints. Different numbers of variables and constraints would work out in similar ways as this example.

The four variables that example function f depends on will be denoted by x_1, x_2, x_3, and x_4. The two constraints will be taken to be equations of the form $g(x_1, x_2, x_3, x_4) = 0$ and $h(x_1, x_2, x_3, x_4) = 0$, for suitable functions g and h. Constraints can always be brought in such a form by taking everything in the constraint's equation to the left-hand side of the equals sign.

So the example problem is:

$$\begin{aligned} \text{stationarize:} \quad & f(x_1, x_2, x_3, x_4) \\ \text{subject to:} \quad & g(x_1, x_2, x_3, x_4) = 0,\ h(x_1, x_2, x_3, x_4) = 0 \end{aligned}$$

Stationarize means to find locations where the function has a minimum or a maximum, or any other point where it does not change under small changes of the variables x_1, x_2, x_3, x_4 as long as these satisfy the constraints.

The first thing to note is that rather than considering f to be a function of x_1, x_2, x_3, x_4, you can consider it instead to be to be a function of g and h and only two additional variables from x_1, x_2, x_3, x_4, say x_3 and x_4:

$$f(x_1, x_2, x_3, x_4) = \tilde{f}(g, h, x_3, x_4)$$

The reason you can do that is that you should in principle be able to reconstruct the two missing variables x_1 and x_2 given g, h, x_3, and x_4.

As a result, any small change in the function f, regardless of constraints, can be written using the expression for a total differential as:

$$\mathrm{d}f = \frac{\partial \tilde{f}}{\partial g}\mathrm{d}g + \frac{\partial \tilde{f}}{\partial h}\mathrm{d}h + \frac{\partial \tilde{f}}{\partial x_3}\mathrm{d}x_3 + \frac{\partial \tilde{f}}{\partial x_4}\mathrm{d}x_4.$$

At the desired stationary point, acceptable changes in variables are those that keep g and h constant at zero; they have $\mathrm{d}g = 0$ and $\mathrm{d}h = 0$. So for f to be stationary under all acceptable changes of variables, you must have that the final two terms are zero for any changes in variables. This means that the

partial derivatives in the final two terms must be zero since the changes $\mathrm{d}x_3$ and $\mathrm{d}x_4$ can be arbitrary.

For changes in variables that do go out of bounds, the change in f will *not* be zero; that change will be given by the first two terms in the right-hand side. So, the erroneous changes in f due to going out of bounds are these first two terms, and if we subtract them, we get zero net change for *any* arbitrary change in variables:

$$\mathrm{d}f - \frac{\partial \tilde{f}}{\partial g}\mathrm{d}g - \frac{\partial \tilde{f}}{\partial h}\mathrm{d}h = 0 \text{ always.}$$

In other words, if we "penalize" the change in f for going out of bounds by amounts $\mathrm{d}g$ and $\mathrm{d}h$ at the rate above, any change in variables will produce a penalized change of zero, whether it stays within bounds or not.

The two derivatives at the stationary point in the expression above are the Lagrangian multipliers or penalty factors, call them $\epsilon_1 = \partial \tilde{f}/\partial g$ and $\epsilon_2 = \partial \tilde{f}/\partial h$. In those terms

$$\mathrm{d}f - \epsilon_1 \mathrm{d}g - \epsilon_2 \mathrm{d}h = 0$$

for whatever is the change in the variables g, h, x_3, x_4, and that means for whatever is the change in the original variables x_1, x_2, x_3, x_4. Therefore, the change in the penalized function

$$f - \epsilon_1 g - \epsilon_2 h$$

is zero whatever is the change in the variables x_1, x_2, x_3, x_4.

In practical application, explicitly computing the Lagrangian multipliers ϵ_1 and ϵ_2 as the derivatives of function $\tilde{f}$ is not needed. You get four equations by putting the derivatives of the penalized f with respect to x_1 through x_4 equal to zero, and the two constraints provide two more equations. Six equations is enough to find the six unknowns x_1 through x_4, ϵ_1 and ϵ_2.

D.49 The generalized variational principle

The purpose of this note is to verify directly that the variation of the expectation energy is zero at any energy eigenstate, not just the ground state.

Suppose that you are trying to find some energy eigenstate ψ_n with eigenvalue E_n, and that you are close to it, but no cigar. Then the wave function can be written as

$$\psi = \varepsilon_1\psi_1 + \varepsilon_2\psi_2 + \ldots + \varepsilon_{n-1}\psi_{n-1} + (1+\varepsilon_n)\psi_n + \varepsilon_{n+1}\psi_{n+1} + \ldots$$

where ψ_n is the one you want and the remaining terms together are the small error in wave function, written in terms of the eigenfunctions. Their coefficients $\varepsilon_1, \varepsilon_2, \ldots$ are small.

The normalization condition $\langle\psi|\psi\rangle = 1$ is, using orthonormality:

$$1 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_{n-1}^2 + (1+\varepsilon_n)^2 + \varepsilon_{n+1}^2 + \ldots$$

The expectation energy is

$$\langle E \rangle = \varepsilon_1^2 E_1 + \varepsilon_2^2 E_2 + \ldots + \varepsilon_{n-1}^2 E_{n-1} + (1+\varepsilon_n)^2 E_n + \varepsilon_{n+1}^2 E_{n+1} + \ldots$$

or plugging in the normalization condition to eliminate $(1+\varepsilon_n)^2$

$$\begin{aligned} \langle E \rangle &= \varepsilon_1^2 (E_1 - E_n) + \varepsilon_2^2 (E_2 - E_n) + \ldots + \\ &\quad \varepsilon_{n-1}^2 (E_{n-1} - E_n) + E_n + \varepsilon_{n+1}^2 (E_{n+1} - E_n) + \ldots \end{aligned}$$

Assuming that the energy eigenvalues are arranged in increasing order, the terms before E_n in this sum are negative and the ones behind E_n positive. So E_n is neither a maximum nor a minimum; depending on conditions $\langle E \rangle$ can be greater or smaller than E_n.

Now, if you make small changes in the wave function, the values of $\varepsilon_1, \varepsilon_2, \ldots$ will slightly change, by small amounts that will be indicated by $\delta\varepsilon_1, \delta\varepsilon_2, \ldots$, and you get

$$\begin{aligned} \delta \langle E \rangle &= 2\varepsilon_1 (E_1 - E_n)\delta\varepsilon_1 + 2\varepsilon_2 (E_2 - E_n)\delta\varepsilon_2 + \ldots \\ &\quad + 2\varepsilon_{n-1}(E_{n-1} - E_n)\delta\varepsilon_{n-1} + 2\varepsilon_{n+1}(E_{n+1} - E_n)\delta\varepsilon_{n+1} + \ldots \end{aligned}$$

This is zero when $\varepsilon_1 = \varepsilon_2 = \ldots = 0$, so when ψ is the exact eigenfunction ψ_n. And it is nonzero as soon as any of $\varepsilon_1, \varepsilon_2, \ldots$ is nonzero; a change in that coefficient will produce a nonzero change in expectation energy. So the variational condition $\delta\langle E \rangle = 0$ is satisfied at the exact eigenfunction ψ_n, but not at any nearby different wave functions.

The bottom line is that if you locate the nearest wave function for which $\delta\langle E \rangle = 0$ for all acceptable small changes in that wave function, well, if you are in the vicinity of an energy eigenfunction, you are going to find that eigenfunction.

One final note. If you look at the expression above, it seems like none of the other eigenfunctions are eigenfunctions. For example, the ground state would be the case that ε_1 is one, and all the other coefficients zero. So a small change in ε_1 would seem to produce a change $\delta\langle E \rangle$ in expectation energy, and the expectation energy is supposed to be constant at eigenstates.

The problem is the normalization condition, whose differential form says that

$$0 = 2\varepsilon_1\delta\varepsilon_1 + 2\varepsilon_2\delta\varepsilon_2 + \ldots + 2\varepsilon_{n-1}\delta\varepsilon_{n-1} + 2(1+\varepsilon_n)\delta\varepsilon_n + 2\varepsilon_{n+1}\delta\varepsilon_{n+1} + \ldots$$

At $\varepsilon_1 = 1$ and $\varepsilon_2 = \ldots = \varepsilon_{n-1} = 1 + \varepsilon_n = \varepsilon_{n+1} = \ldots = 0$, this implies that the change $\delta\varepsilon_1$ must be zero. And that means that the change in expectation energy is in fact zero. You see that you really need to eliminate ε_1 from the list of coefficients near ψ_1, rather than ε_n as the analysis for ψ_n did, for the mathematics not to blow up. A coefficient that is not allowed to change at a point in the vicinity of interest is a confusing coefficient to work with.

D.50 Spin degeneracy

To see that generally speaking the basic form of the Hamiltonian produces energy degeneracy with respect to spin, but that it is not important for using the Born-Oppenheimer approximation, consider the example of three electrons.

Any three-electron energy eigenfunction $\psi^{\rm E}$ with $H\psi^{\rm E} = E^{\rm E}\psi^{\rm E}$ can be split into separate spatial functions for the distinct combinations of electron spin values as

$$\begin{aligned}\psi^{\rm E} \;=\; & \psi^{\rm E}_{+++}\uparrow\uparrow\uparrow + \psi^{\rm E}_{+--}\uparrow\downarrow\downarrow + \psi^{\rm E}_{-+-}\downarrow\uparrow\downarrow + \psi^{\rm E}_{--+}\downarrow\downarrow\uparrow + \\ & \psi^{\rm E}_{---}\downarrow\downarrow\downarrow + \psi^{\rm E}_{-++}\downarrow\uparrow\uparrow + \psi^{\rm E}_{+-+}\uparrow\downarrow\uparrow + \psi^{\rm E}_{++-}\uparrow\uparrow\downarrow.\end{aligned}$$

Since the assumed Hamiltonian H does not involve spin, each of the eight spatial functions $\psi_{\pm\pm\pm}$ above will separately have to be an eigenfunction of the Hamiltonian with eigenvalue $E^{\rm E}$ if nonzero. In addition, since the first four functions have an odd number of spin up states and the second four an even number, the antisymmetry requirements apply only within the two sets, not between them. The exchanges only affect the order of the spin states, not their number. So the two sets satisfy the antisymmetry requirements individually.

It is now seen that given a solution for the first four wave functions, there is an equally good solution for the second four wave functions that is obtained by inverting all the spins. Since the spins are not in the Hamiltonian, inverting the spins does not change the energy. They have the same energy, but are different because they have different spins.

However, they are orthogonal because their spins are, and the spatial operations in the derivation of the Born-Oppenheimer approximation in the next note do not change that fact. So they turn out to lead to nuclear wave functions that do not affect each other. More precisely, the inner products appearing in the coefficients $a_{n\underline{n}}$ are zero because the spins are orthogonal.

D.51 Born-Oppenheimer nuclear motion

This note gives a derivation of the Born-Oppenheimer Hamiltonian eigenvalue problems (9.14) for the wave functions of the nuclei.

First consider an exact eigenfunction ψ of the complete system, including both the electrons and the nuclei fully. Can it be related somehow to the simpler electron eigenfunctions $\psi^{\rm E}_1, \psi^{\rm E}_2, \ldots$ that ignored nuclear kinetic energy? Yes it can. *For any given set of nuclear coordinates*, the electron eigenfunctions are complete; they are the eigenfunctions of an Hermitian electron Hamiltonian. And that means that you can for any given set of nuclear coordinates write the exact wave function as

$$\psi = \sum_{\underline{n}} c_{\underline{n}} \psi^{\rm E}_{\underline{n}}$$

You can do this for any set of nuclear coordinates that you like, but the coefficients $c_{\underline{n}}$ will be different for different sets of nuclear coordinates. That is just another way of saying that the $c_{\underline{n}}$ are functions of the nuclear coordinates.

So, to be really precise, the wave function of I electrons and J nuclei can be written as:

$$\psi(\vec{r}_1, S_{z1}, \ldots, \vec{r}_I, S_{zI}, \vec{r}^{\,\mathrm{n}}_1, S^{\mathrm{n}}_{z1}, \ldots, \vec{r}^{\,\mathrm{n}}_J, S^{\mathrm{n}}_{zJ}) =$$
$$\sum_{\underline{n}} c_{\underline{n}}(\vec{r}^{\,\mathrm{n}}_1, S^{\mathrm{n}}_{z1}, \ldots, \vec{r}^{\,\mathrm{n}}_J, S^{\mathrm{n}}_{zJ}) \psi^{\mathrm{E}}_{\underline{n}}(\vec{r}_1, S_{z1}, \ldots, \vec{r}_I, S_{zI};\ \vec{r}^{\,\mathrm{n}}_1, S^{\mathrm{n}}_{z1}, \ldots, \vec{r}^{\,\mathrm{n}}_J, S^{\mathrm{n}}_{zJ})$$

where superscripts n indicate nuclear coordinates. (The nuclear spins are really irrelevant, but it cannot hurt to keep them in.)

Consider what this means physically. By construction, the square electron eigenfunctions $|\psi^{\mathrm{E}}_{\underline{n}}|^2$ give the probability of finding the electrons *assuming that they are in eigenstate $\underline{n}$ and that the nuclei are at the positions listed in the final arguments of the electron eigenfunction.* But then the probability that the nuclei are actually at those positions, and that the electrons are actually in eigenstate $\psi^{\mathrm{E}}_{\underline{n}}$, will have to be $|c_{\underline{n}}|^2$. After all, the full wave function ψ must describe the probability for the *entire* system to actually be in a specific state. That means that $c_{\underline{n}}$ must be the nuclear wave function $\psi^{\mathrm{N}}_{\underline{n}}$ for when the electrons are in energy eigenstate $\psi^{\mathrm{E}}_{\underline{n}}$. So from now on, just call it $\psi^{\mathrm{N}}_{\underline{n}}$ instead of $c_{\underline{n}}$. The full wave function is then

$$\boxed{\psi = \sum \psi^{\mathrm{N}}_{\underline{n}} \psi^{\mathrm{E}}_{\underline{n}}} \tag{D.31}$$

In the unsteady case, the $c_{\underline{n}}$, hence the $\psi^{\mathrm{N}}_{\underline{n}}$, will also be functions of time. The $\psi^{\mathrm{E}}_{\underline{n}}$ will remain time independent as long as no explicitly time-dependent terms are added. The derivation then goes exactly the same way as the time-independent Schrödinger equation (Hamiltonian eigenvalue problem) derived below, with $\mathrm{i}\hbar\partial/\partial t$ replacing E.

So far, no approximations have been made; the only thing that has been done is to define the nuclear wave functions $\psi^{\mathrm{N}}_{\underline{n}}$. But the objective is still to derive the claimed equation (9.14) for them. To do so plug the expression $\psi = \sum \psi^{\mathrm{N}}_{\underline{n}} \psi^{\mathrm{E}}_{\underline{n}}$ into the exact Hamiltonian eigenvalue problem:

$$\left[\widehat{T}^{\mathrm{N}} + \widehat{T}^{\mathrm{E}} + V^{\mathrm{NE}} + V^{\mathrm{EE}} + V^{\mathrm{NN}}\right] \sum_{\underline{n}} \psi^{\mathrm{N}}_{\underline{n}} \psi^{\mathrm{E}}_{\underline{n}} = E \sum_{\underline{n}} \psi^{\mathrm{N}}_{\underline{n}} \psi^{\mathrm{E}}_{\underline{n}}$$

Note first that the eigenfunctions can be taken to be real since the Hamiltonian is real. If the eigenfunctions were complex, then their real and imaginary parts separately would be eigenfunctions, and both of these are real. This argument applies to both the electron eigenfunctions separately as well as to the full eigenfunction. The trick is now to take an inner product of the equation

above with a chosen electron eigenfunction $\psi_n^{\rm E}$. More precisely, multiply the entire equation by $\psi_n^{\rm E}$, and integrate/sum over the electron coordinates and spins only, keeping the nuclear positions and spins at fixed values.

What do you get? Consider the terms in reverse order, from right to left. In the right hand side, the electron-coordinate inner product $\langle\psi_n^{\rm E}|\psi_{\underline{n}}^{\rm E}\rangle_e$ is zero unless $\underline{n}=n$, and then it is one, since the electron wave functions are orthonormal for given nuclear coordinates. So all we have left in the right-hand side is $E\psi_n^{\rm N}$, Check, $E\psi_n^{\rm N}$ is the correct right hand side in the nuclear-wave-function Hamiltonian eigenvalue problem (9.14).

Turning to the latter four terms in the left-hand side, remember that by definition the electron eigenfunctions satisfy

$$\left[\widehat{T}^{\rm E}+V^{\rm NE}+V^{\rm EE}+V^{\rm NN}\right]\psi_{\underline{n}}^{\rm E}=(E_{\underline{n}}^{\rm E}+V^{\rm NN})\psi_{\underline{n}}^{\rm E}$$

and if you then take an inner product of $\sum\psi_{\underline{n}}^{\rm N}(E_{\underline{n}}^{\rm E}+V^{\rm NN})\psi_{\underline{n}}^{\rm E}$ with $\psi_n^{\rm E}$, it is just like the earlier term, and you get $(E_n^{\rm E}+V^{\rm NN})\psi_n^{\rm N}$. Check, that are two of the terms in the left-hand side of (9.14) that you need.

That leaves only the nuclear kinetic term, and that one is a bit tricky. Recalling the definition (9.5) of the kinetic energy operator $\widehat{T}^{\rm N}$ in terms of the nuclear coordinate Laplacians, you have

$$-\sum_{j=1}^{J}\sum_{\alpha=1}^{3}\sum_{\underline{n}}\frac{\hbar^2}{2m_j^{\rm n}}\frac{\partial^2}{\partial {r_{\alpha j}^{\rm n}}^2}\psi_{\underline{n}}^{\rm N}\psi_{\underline{n}}^{\rm E}$$

Remember that not just the nuclear wave functions, but also the electron wave functions depend on the nuclear coordinates. So, if you differentiate out the product, you get

$$-\sum_{j=1}^{J}\sum_{\alpha=1}^{3}\sum_{\underline{n}}\left[\frac{\hbar^2}{2m_j^{\rm n}}\frac{\partial^2\psi_{\underline{n}}^{\rm N}}{\partial {r_{\alpha j}^{\rm n}}^2}\psi_{\underline{n}}^{\rm E}+\frac{\hbar^2}{m_j^{\rm n}}\frac{\partial\psi_{\underline{n}}^{\rm N}}{\partial r_{\alpha j}^{\rm n}}\frac{\partial\psi_{\underline{n}}^{\rm E}}{\partial r_{\alpha j}^{\rm n}}+\frac{\hbar^2}{2m_j^{\rm n}}\psi_{\underline{n}}^{\rm N}\frac{\partial^2\psi_{\underline{n}}^{\rm E}}{\partial {r_{\alpha j}^{\rm n}}^2}\right]$$

Now if you take the inner product with electron eigenfunction $\psi_n^{\rm E}$, the first term in the brackets gives you what you need, the expression for the kinetic energy of the nuclei. But you do not want the other two terms; these terms have the nuclear kinetic energy differentiations at least in part on the electron wave function instead of on the nuclear wave function.

Well, whether you like it or not, the exact equation is, collecting all terms and rearranging,

$$\boxed{\left[\widehat{T}^{\rm N}+V^{\rm NN}+E_n^{\rm E}\right]\psi_n^{\rm N}=E\psi_n^{\rm N}+\sum_{\underline{n}}a_{n\underline{n}}\psi_{\underline{n}}^{\rm N}} \tag{D.32}$$

where

$$\boxed{\widehat{T}^{\rm N} = -\sum_{j=1}^{J}\sum_{\alpha=1}^{3}\frac{\hbar^2}{2m_j^{\rm n}}\frac{\partial^2}{\partial {r_{\alpha j}^{\rm n}}^2}} \tag{D.33}$$

$$\boxed{a_{n\underline{n}} = \sum_{j=1}^{J}\sum_{\alpha=1}^{3}\frac{\hbar^2}{2m_j^{\rm n}}\left(2\left\langle \psi_n^{\rm E}\middle|\frac{\partial\psi_{\underline{n}}^{\rm E}}{\partial r_{\alpha j}^{\rm n}}\right\rangle\frac{\partial}{\partial r_{\alpha j}^{\rm n}} + \left\langle \psi_n^{\rm E}\middle|\frac{\partial^2\psi_{\underline{n}}^{\rm E}}{\partial {r_{\alpha j}^{\rm n}}^2}\right\rangle\right)} \tag{D.34}$$

The first thing to note is the final sum in (D.32). Unless you can talk away this sum as negligible, (9.14) is not valid. The "off-diagonal" coefficients, the $a_{n\underline{n}}$ for $\underline{n} \neq n$, are particularly bad news, because they produce interactions between the different potential energy surfaces, shifting energy from one value of n to another. These off-diagonal terms are called "vibronic coupling terms." (The word is a contraction of "vibration" and "electronic," if you are wondering.)

Let's have a closer look at (D.33) and (D.34) to see how big the various terms really are. At first appearance it might seem that both the nuclear kinetic energy $\widehat{T}^{\rm N}$ and the coefficients $a_{n\underline{n}}$ can be ignored, since both are inversely proportional to the nuclear masses, hence apparently thousands of times smaller than the electronic kinetic energy included in $E_n^{\rm E}$. But do not go too quick here. First ballpark the typical derivative, $\partial/\partial r_{\alpha j}^{\rm n}$ when applied to the nuclear wave function. You can estimate such a derivative as $1/\ell^{\rm N}$, where $\ell^{\rm N}$ is the typical length over which there are significant changes in a nuclear wave function $\psi_n^{\rm N}$. Well, there are significant changes in nuclear wave functions if you go from the middle of a nucleus to its outside, and that is a very small distance compared to the typical size of the electron blob $\ell^{\rm E}$. It means that the distance $\ell^{\rm N}$ is small. So the relative importance of the nuclear kinetic energy increases by a factor $(\ell^{\rm E}/\ell^{\rm N})^2$ relative to the electron kinetic energy, compensating quite a lot for the much higher nuclear mass. So keeping the nuclear kinetic energy is definitely a good idea.

How about the coefficients $a_{n\underline{n}}$? Well, *normally* the electron eigenfunctions only change appreciable when you vary the nuclear positions over a length comparable to the electron blob scale $\ell^{\rm E}$. Think back of the example of the hydrogen molecule. The ground state separation between the nuclei was found as 0.87Å. But you would not see a dramatic change in electron wave functions if you made it a few percent more or less. To see a dramatic change, you would have to make the nuclear distance 1.5Å, for example. So the derivatives $\partial/\partial r_{\alpha j}^{\rm n}$ applied to the electron wave functions are normally not by far as large as those applied to the nuclear wave functions, hence the $a_{n\underline{n}}$ terms are relatively small compared to the nuclear kinetic energy, and ignoring them is usually justified. So the final conclusion is that equation (9.14) is usually justified.

But there are exceptions. If different energy levels get close together, the electron wave functions become very sensitive to small effects, including small changes in the nuclear positions. When the wave functions have become sensitive

enough that they vary significantly under nuclear position changes comparable in size to the nuclear wave function blobs, you can no longer ignore the $a_{n\underline{n}}$ terms and (9.14) becomes invalid.

You can be a bit more precise about that claim with a few tricks. Consider the factors

$$\left\langle \psi^{\rm E}_n \middle| \frac{\partial \psi^{\rm E}_{\underline{n}}}{\partial r^{\rm n}_{\alpha j}} \right\rangle$$

appearing in the $a_{n\underline{n}}$, (D.34). First of all, these factors are zero when $\underline{n} = n$. The reason is that because of orthonormality, $\langle \psi^{\rm E}_n | \psi^{\rm E}_n \rangle = 1$, and taking the $\partial/\partial r^{\rm n}_{\alpha j}$ derivative of that, noting that the eigenfunctions are real, you see that the factor is zero.

For $\underline{n} \neq n$, the following trick works:

$$\begin{aligned}
\left\langle \psi^{\rm E}_n \middle| \frac{\partial}{\partial r^{\rm n}_{\alpha j}} H^{\rm E} - H^{\rm E} \frac{\partial}{\partial r^{\rm n}_{\alpha j}} \middle| \psi^{\rm E}_{\underline{n}} \right\rangle &= (E^{\rm E}_{\underline{n}} - E^{\rm E}_n) \left\langle \psi^{\rm E}_n \middle| \frac{\partial \psi^{\rm E}_{\underline{n}}}{\partial r^{\rm n}_{\alpha j}} \right\rangle \\
&= \frac{Z_j e^2}{4\pi\epsilon_0} \sum_{i=1}^{I} \left\langle \psi^{\rm E}_n \middle| \frac{r^{\rm n}_{\alpha j} - r_{\alpha i}}{r^3_{ij}} \middle| \psi^{\rm E}_{\underline{n}} \right\rangle
\end{aligned}$$

The first equality is just a matter of the definition of the electron eigenfunctions and taking the second $H^{\rm E}$ to the other side, which you can do since it is Hermitian. The second equality is a matter of looking up the Hamiltonian in chapter 9.2.1 and then working out the commutator in the leftmost inner product. ($V^{\rm NN}$ does not commute with the derivative, but you can use orthogonality on the cleaned up expression.) The bottom line is that the final inner product is finite, with no reason for it to become zero when energy levels approach. So, looking at the second equality, the first term in $a_{n\underline{n}}$, (D.34), blows up like $1/(E^{\rm E}_{\underline{n}} - E^{\rm E}_n)$ when those energy levels become equal.

As far as the final term in $a_{n\underline{n}}$ is concerned, like the second term, you would expect it to become important when the scale of nontrivial changes in electron wave functions with nuclear positions becomes comparable to the size of the nuclear wave functions. You can be a little bit more precise by taking one more derivative of the inner product expression derived above,

$$\left\langle \frac{\partial \psi^{\rm E}_n}{\partial r^{\rm n}_{\alpha j}} \middle| \frac{\partial \psi^{\rm E}_{\underline{n}}}{\partial r^{\rm n}_{\alpha j}} \right\rangle + \left\langle \psi^{\rm E}_n \middle| \frac{\partial^2 \psi^{\rm E}_{\underline{n}}}{\partial {r^{\rm n}_{\alpha j}}^2} \right\rangle = \frac{\partial}{\partial r^{\rm n}_{\alpha j}} \frac{1}{E^{\rm E}_{\underline{n}} - E^{\rm E}_n} \frac{Z_j e^2}{4\pi\epsilon_0} \sum_{i=1}^{I} \left\langle \psi^{\rm E}_n \middle| \frac{r^{\rm n}_{\alpha j} - r_{\alpha i}}{r_{ij}} \middle| \psi^{\rm E}_{\underline{n}} \right\rangle$$

The first term should not be large: while the left hand side of the inner product has a large component along $\psi^{\rm E}_{\underline{n}}$, the other side has zero component and vice-versa. The final term should be of order $1/(E^{\rm E}_{\underline{n}} - E^{\rm E}_n)^2$, as you can see if you first change the origin of the integration variable in the inner product to be at the nuclear position, to avoid having to differentiate the potential derivative. So you conclude that the second term of coefficient $a_{n\underline{n}}$ is of order $1/(E^{\rm E}_{\underline{n}} - E^{\rm E}_n)^2$.

In view of the fact that this term has one less derivative on the nuclear wave function, that is just enough to allow it to become significant at about the same time that the first term does.

The diagonal part of matrix $a_{n\underline{n}}$, i.e. the a_{nn} terms, is somewhat interesting since it produces a change in effective energy without involving interactions with the other potential energy surfaces, i.e. without interaction with the $\psi^{\rm N}_{\underline{n}}$ for $\underline{n} \neq n$. The diagonal part is called the "Born-Oppenheimer diagonal correction." Since as noted above, the first term in the expression (D.34) for the $a_{n\underline{n}}$ does not have a diagonal part, the diagonal correction is given by the second term.

Note that in a transient case that starts out as a single nuclear wave function $\psi^{\rm N}_n$, the diagonal term a_{nn} multiplies the predominant nuclear wave function $\psi^{\rm N}_n$, while the off-diagonal terms only multiply the small other nuclear wave functions. So despite not involving any derivative of the nuclear wave function, the diagonal term will initially be the main correction to the Born-Oppenheimer approximation. It will remain important at later times.

D.52 Simplification of the Hartree-Fock energy

This note derives the expectation energy for a wave function given by a single Slater determinant.

First note that if you multiply out a Slater determinant

$$\Psi = |\det \psi^{\rm s}_1\updownarrow_1, \psi^{\rm s}_2\updownarrow_2, \psi^{\rm s}_3\updownarrow_3, \ldots\rangle$$

you are going to get terms, or Hartree products if you want, of the form

$$\frac{\pm}{\sqrt{I!}}\, \psi^{\rm s}_{n_1}(\vec{r}_1)\updownarrow_{n_1}(S_{z1})\; \psi^{\rm s}_{n_2}(\vec{r}_2)\updownarrow_{n_2}(S_{z2})\; \psi^{\rm s}_{n_3}(\vec{r}_3)\updownarrow_{n_3}(S_{z3})\; \ldots$$

where the numbers $n_1, n_2, n_3, \ldots$ of the single-electron states can have values from 1 to I, but they must be *all different*. So there are $I!$ such terms: there are I possibilities among $1, 2, 3, \ldots, I$ for the number n_1 of the single-electron state for electron 1, which leaves $I-1$ remaining possibilities for the number n_2 of the single-electron state for electron 2, $I-2$ remaining possibilities for n_3, etcetera. That means a total of $I(I-1)(I-2)\ldots 2\,1 = I!$ terms. As far as the sign of the term is concerned, just don't worry about it. The only thing to remember is that whenever you exchange two n values, it changes the sign of the term. It has to be, because exchanging n values is equivalent to exchanging electrons, and the complete wave function must change sign under that.

To make the above more concrete, consider the example of a Slater determinant of three single-electron functions. It writes out to, taking $\sqrt{I!}$ to the other side for convenience,

$$|\det \psi^{\rm s}_1\updownarrow_1, \psi^{\rm s}_2\updownarrow_2, \psi^{\rm s}_3\updownarrow_3\rangle\sqrt{I!} =$$

$$
\begin{aligned}
&+\psi^{\rm s}_1(\vec{r}_1)\updownarrow_1(S_{z1})\psi^{\rm s}_2(\vec{r}_2)\updownarrow_2(S_{z2})\psi^{\rm s}_3(\vec{r}_3)\updownarrow_3(S_{z3})\\
&-\psi^{\rm s}_1(\vec{r}_1)\updownarrow_1(S_{z1})\psi^{\rm s}_3(\vec{r}_2)\updownarrow_3(S_{z2})\psi^{\rm s}_2(\vec{r}_3)\updownarrow_2(S_{z3})\\
&-\psi^{\rm s}_2(\vec{r}_1)\updownarrow_2(S_{z1})\psi^{\rm s}_1(\vec{r}_2)\updownarrow_1(S_{z2})\psi^{\rm s}_3(\vec{r}_3)\updownarrow_3(S_{z3})\\
&+\psi^{\rm s}_2(\vec{r}_1)\updownarrow_2(S_{z1})\psi^{\rm s}_3(\vec{r}_2)\updownarrow_3(S_{z2})\psi^{\rm s}_1(\vec{r}_3)\updownarrow_1(S_{z3})\\
&+\psi^{\rm s}_3(\vec{r}_1)\updownarrow_3(S_{z1})\psi^{\rm s}_1(\vec{r}_2)\updownarrow_1(S_{z2})\psi^{\rm s}_2(\vec{r}_3)\updownarrow_2(S_{z3})\\
&-\psi^{\rm s}_3(\vec{r}_1)\updownarrow_3(S_{z1})\psi^{\rm s}_2(\vec{r}_2)\updownarrow_2(S_{z2})\psi^{\rm s}_1(\vec{r}_3)\updownarrow_1(S_{z3})
\end{aligned}
$$

The first two rows in the expansion cover the possibility that $n_1 = 1$, with the first one the possibility that $n_2 = 2$ and the second one the possibility that $n_2 = 3$; note that then there are no choices left for n_3. Similarly the second two rows cover the two possibilities that $n_1 = 2$, and the third that $n_1 = 3$. You see that there are $3! = 6$ Hartree product terms total.

Next, recall that the Hamiltonian consists of single-electron Hamiltonians $h^{\rm e}_i$ and electron-pair repulsion potentials $v^{\rm ee}_{i\bar{\imath}}$. The expectation value of a single electron Hamiltonian $h^{\rm e}_i$ will be done first. In forming the inner product $\langle\Psi|h^{\rm e}_i|\Psi\rangle$, and taking Ψ apart into its Hartree product terms as above, you are going to end up with a large number of individual terms that all look like

$$
\begin{aligned}
\Big\langle &\frac{\pm}{\sqrt{I!}}\psi^{\rm s}_{n_1}(\vec{r}_1)\updownarrow_{n_1}(S_{z1})\psi^{\rm s}_{n_2}(\vec{r}_2)\updownarrow_{n_2}(S_{z2})\ldots\psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})\ldots\psi^{\rm s}_{n_I}(\vec{r}_I)\updownarrow_{n_I}(S_{zI})\Big|\\
&h^{\rm e}_i\Big|\frac{\pm}{\sqrt{I!}}\psi^{\rm s}_{\overline{n}_1}(\vec{r}_1)\updownarrow_{\overline{n}_1}(S_{z1})\psi^{\rm s}_{\overline{n}_2}(\vec{r}_2)\updownarrow_{\overline{n}_2}(S_{z2})\ldots\psi^{\rm s}_{\overline{n}_i}(\vec{r}_i)\updownarrow_{\overline{n}_i}(S_{zi})\ldots\psi^{\rm s}_{\overline{n}_I}(\vec{r}_I)\updownarrow_{\overline{n}_I}(S_{zI})\Big\rangle
\end{aligned}
$$

Note that overlines will be used to distinguish the wave function in the right hand side of the inner product from the one in the left hand side. Also note that to take this inner product, you have to integrate over $3I$ scalar position coordinates, and sum over I spin values.

But multiple integrals, and sums, can be factored into single integrals, and sums, as long as the integrands and limits only involve single variables. So you can factor out the inner product as

$$
\begin{aligned}
\frac{\pm}{\sqrt{I!}}\frac{\pm}{\sqrt{I!}}&\Big\langle\psi^{\rm s}_{n_1}(\vec{r}_1)\updownarrow_{n_1}(S_{z1})\Big|\psi^{\rm s}_{\overline{n}_1}(\vec{r}_1)\updownarrow_{\overline{n}_1}(S_{z1})\Big\rangle\\
&\times\Big\langle\psi^{\rm s}_{n_2}(\vec{r}_2)\updownarrow_{n_2}(S_{z2})\Big|\psi^{\rm s}_{\overline{n}_2}(\vec{r}_2)\updownarrow_{\overline{n}_2}(S_{z2})\Big\rangle\\
&\times\ldots\\
&\times\Big\langle\psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})\Big|h^{\rm e}_i\Big|\psi^{\rm s}_{\overline{n}_i}(\vec{r}_i)\updownarrow_{\overline{n}_i}(S_{zi})\Big\rangle\\
&\times\ldots\\
&\times\Big\langle\psi^{\rm s}_{n_I}(\vec{r}_I)\updownarrow_{n_I}(S_{zI})\Big|\psi^{\rm s}_{\overline{n}_I}(\vec{r}_I)\updownarrow_{\overline{n}_I}(S_{zI})\Big\rangle
\end{aligned}
$$

Now you can start the weeding-out process, because the single-electron functions are orthonormal. So factors in this product are zero unless all of the

following requirements are met:

$$n_1 = \overline{n}_1,\ n_2 = \overline{n}_2,\ \ldots,\ n_{i-1} = \overline{n}_{i-1},\ n_{i+1} = \overline{n}_{i+1},\ \ldots,\ n_I = \overline{n}_I$$

Note that $\langle \psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})|h^{\rm e}_i|\psi^{\rm s}_{\overline{n}_i}(\vec{r}_i)\updownarrow_{\overline{n}_i}(S_{zi})\rangle$ does not require $n_i = \overline{n}_i$ for a nonzero value, since the single-electron functions are most definitely not eigenfunctions of the single-electron Hamiltonians, (you would wish things were that easy!) But now remember that the numbers $n_1, n_2, n_3, \ldots$ in an individual term are all different. So the numbers $n_1, n_2, \ldots, n_{i-1}, n_{i+1}, \ldots$ include all the numbers that are *not* equal to n_i. Then so do $\overline{n}_1$, $\overline{n}_2$, $\ldots$, $\overline{n}_{i-1}$, $\overline{n}_{i+1}$,$\ldots$, because they are the same. And since $\overline{n}_i$ must be different from all of those, it can only be equal to n_i anyway.

So what is left? Well, with all the $\overline{n}$ values equal to the corresponding n values, all the plain inner products are one on account of orthonormality, and the only thing left is:

$$\frac{\pm}{\sqrt{I!}}\frac{\pm}{\sqrt{I!}}\left\langle \psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})\right|h^{\rm e}_i\left|\psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})\right\rangle$$

Also, the two signs are equal, because with all the $\overline{n}$ values equal to the corresponding n values, the wave function term in the right side of the inner product is the exact same one as in the left side. So the signs multiply to 1, and you can further factor out the spin inner product, which is one since the spin states are normalized:

$$\frac{1}{I!}\left\langle \psi^{\rm s}_{n_i}(\vec{r}_i)\right|h^{\rm e}_i\left|\psi^{\rm s}_{n_i}(\vec{r}_i)\right\rangle\left\langle \updownarrow_{n_i}(S_{zi})\right|\left.\updownarrow_{n_i}(S_{zi})\right\rangle = \frac{1}{I!}\left\langle \psi^{\rm s}_{n_i}(\vec{r}_i)\right|h^{\rm e}_i\left|\psi^{\rm s}_{n_i}(\vec{r}_i)\right\rangle \equiv \frac{1}{I!}E^{\rm e}_n$$

where for brevity the remaining inner product was called $E^{\rm e}_n$. Normally you would call it $E^{\rm e}_{n_i i}$, but an inner product integral does not care what the integration variable is called, so the thing has the same value regardless what the electron i is. Only the value of the single-electron function number $n_i = n$ makes a difference.

Next, how many such terms are there for a given electron i and single-electron function number n? Well, for a given n value for electron i, there are $I-1$ possible values left among $1, 2, 3, \ldots$ for the n value of the first of the other electrons, then $I-2$ left for the second of the other electrons, etcetera. So there are a total of $(I-1)(I-2)\ldots 1 = (I-1)!$ such terms. Since $(I-1)!/I! = 1/I$, if you sum them all together you get a total contribution from terms in which electron i is in state n equal to $E^{\rm e}_n/I$. Summing over the I electrons kills off the factor $1/I$ and so you finally get the total energy due to the single-electron Hamiltonians as

$$\sum_{n=1}^{I} E^{\rm e}_n \qquad E^{\rm e}_n = \left\langle \psi^{\rm s}_n(\vec{r})\right|h^{\rm e}\left|\psi^{\rm s}_n(\vec{r})\right\rangle$$

You might have guessed that answer from the start. Since the inner product integral is the same for all electrons, the subscripts i have been omitted.

The good news is that the reasoning to get the Coulomb and exchange contributions is pretty much the same. A single electron to electron repulsion term $v^{\rm ee}_{i\underline{i}}$ between an electron numbered i and another numbered $\underline{i}$ makes a contribution to the expectation energy equal to $\langle\Psi|v^{\rm ee}_{i\underline{i}}|\Psi\rangle$, and if you multiply out Ψ, you get terms of the general form:

$$\frac{1}{I!}\Big\langle \psi^{\rm s}_{n_1}(\vec{r}_1)\updownarrow_{n_1}(S_{z1})\psi^{\rm s}_{n_2}(\vec{r}_2)\updownarrow_{n_2}(S_{z2})\ldots\psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})\ldots\psi^{\rm s}_{n_{\underline{i}}}(\vec{r}_{\underline{i}})\updownarrow_{n_{\underline{i}}}(S_{z\underline{i}})\ldots\Big|$$
$$v^{\rm ee}_{i\underline{i}}\Big|\psi^{\rm s}_{\overline{n}_1}(\vec{r}_1)\updownarrow_{\overline{n}_1}(S_{z1})\psi^{\rm s}_{\overline{n}_2}(\vec{r}_2)\updownarrow_{\overline{n}_2}(S_{z2})\ldots\psi^{\rm s}_{\overline{n}_i}(\vec{r}_i)\updownarrow_{\overline{n}_i}(S_{zi})\ldots\psi^{\rm s}_{\overline{n}_{\underline{i}}}(\vec{r}_{\underline{i}})\updownarrow_{\overline{n}_{\underline{i}}}(S_{z\underline{i}})\ldots\Big\rangle$$

You can again split into a product of individual inner products, except that you cannot split between electrons i and $\underline{i}$ since $v^{\rm ee}_{i\underline{i}}$ involves both electrons in a nontrivial way. Still, you get again that all the other n values must be the same as the corresponding $\overline{n}$ values, eliminating those inner products from the expression:

$$\frac{1}{I!}\Big\langle \psi^{\rm s}_{n_i}(\vec{r}_i)\updownarrow_{n_i}(S_{zi})\psi^{\rm s}_{n_{\underline{i}}}(\vec{r}_{\underline{i}})\updownarrow_{n_{\underline{i}}}(S_{z\underline{i}})\Big|v^{\rm ee}_{i\underline{i}}\Big|\psi^{\rm s}_{\overline{n}_i}(\vec{r}_i)\updownarrow_{\overline{n}_i}(S_{zi})\psi^{\rm s}_{\overline{n}_{\underline{i}}}(\vec{r}_{\underline{i}})\updownarrow_{\overline{n}_{\underline{i}}}(S_{z\underline{i}})\Big\rangle$$

For given values of n_i and $n_{\underline{i}}$, there are $(I-2)!$ equivalent terms, since that is the number of possibilities left for the $n=\overline{n}$-values of the other $I-2$ electrons.

Next, $\overline{n}_i$ and $\overline{n}_{\underline{i}}$ must together be the same *pair* of numbers as n_i and $n_{\underline{i}}$, since they must be the two numbers left by the set of numbers not equal to n_i and $n_{\underline{i}}$. But that still leaves two possibilities, they can be in the same order or in reversed order:

$$\overline{n}_i = n_i,\ \overline{n}_{\underline{i}} = n_{\underline{i}} \qquad \text{or} \qquad \overline{n}_i = n_{\underline{i}},\ \overline{n}_{\underline{i}} = n_i.$$

The first possibility gives rise to the Coulomb terms, the second to the exchange ones. Note that the former case represents an inner product involving a Hartree product with itself, and the latter case an inner product of a Hartree product with the Hartree product that is the same save for the fact that it has n_i and $n_{\underline{i}}$ reversed, or equivalently, electrons i and $\underline{i}$ exchanged.

Consider the Coulomb terms first. For those the two Hartree products in the inner product are the same, so their signs multiply to one. Also, their spin states will be the same, so that inner product will be one too. And as noted there are $(I-2)!$ equivalent terms for given n_i and $n_{\underline{i}}$, so for each pair of electrons i and $\underline{i}\neq i$, and each pair of states $n=n_i$ and $\underline{n}=n_{\underline{i}}$, you get one term

$$\frac{1}{I(I-1)}J_{n\underline{n}}$$

with

$$J_{n\underline{n}} \equiv \Big\langle \psi^{\rm s}_n(\vec{r})\psi^{\rm s}_{\underline{n}}(\vec{\underline{r}})\Big|v^{\rm ee}\Big|\psi^{\rm s}_n(\vec{r})\psi^{\rm s}_{\underline{n}}(\vec{\underline{r}})\Big\rangle.$$

Again, the $J_{n\underline{n}}$ are the same regardless of what i and $\underline{i}$ are; they depend only on what $n = n_i$ and $\underline{n} = n_{\underline{i}}$ are. So the subscripts i and $\underline{i}$ were left out, after setting $\vec{r} = \vec{r}_i$ and $\underline{\vec{r}} = \vec{r}_{\underline{i}}$.

You now need to sum over all pairs of electrons with $i \neq \underline{i}$ and pairs of single-electron function numbers $n \neq \underline{n}$. Since there are a total of $I(I-1)$ electron pairs, it takes out the factor $1/I(I-1)$, and you get a contribution to the energy

$$\tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^{I} J_{n\underline{n}}$$

The factor $\frac{1}{2}$ was added since for every electron pair, you are summing both $v^{\rm ee}_{i\underline{i}}$ and $v^{\rm ee}_{\underline{i}i}$, and that counts the same energy twice.

The exchange integrals go exactly the same way; the only differences are that the Hartree product in the right hand side of the inner product has the values of $\overline{n}_i$ and $\overline{n}_{\underline{i}}$ reversed, producing a change of sign, and that the inner product of the spins is not trivial. Define

$$K_{n\underline{n}} \equiv \left\langle \psi^{\rm s}_n(\vec{r}) \psi^{\rm s}_{\underline{n}}(\underline{\vec{r}}) \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}}(\vec{r}) \psi^{\rm s}_n(\underline{\vec{r}}) \right\rangle.$$

and then the total contribution is

$$-\tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^{I} K_{n\underline{n}} \langle \uparrow_n | \uparrow_{\underline{n}} \rangle^2$$

Finally, you can leave the constraint $\underline{n} \neq n$ on the sums away since $K_{nn} = J_{nn}$, so they cancel each other.

D.53 Integral constraints

This note verifies the mentioned constraints on the Coulomb and exchange integrals.

To verify that $J_{nn} = K_{nn}$, just check their definitions.

The fact that

$$\begin{aligned} J_{n\underline{n}} &= \langle \psi^{\rm s}_n(\vec{r}_i) \psi^{\rm s}_{\underline{n}}(\vec{r}_{\underline{i}}) | v^{\rm ee}_{i\underline{i}} | \psi^{\rm s}_n(\vec{r}_i) \psi^{\rm s}_{\underline{n}}(\vec{r}_{\underline{i}}) \rangle \\ &= \int_{\text{all } \vec{r}_i} \int_{\text{all } \vec{r}_{\underline{i}}} |\psi^{\rm s}_n(\vec{r}_i) \psi^{\rm s}_{\underline{n}}(\vec{r}_{\underline{i}})|^2 \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{i\underline{i}}} \, {\rm d}^3\vec{r}_i \, {\rm d}^3\vec{r}_{\underline{i}}. \end{aligned}$$

is real and positive is self-evident, since it is an integral of a real and positive function.

The fact that

$$\begin{aligned} K_{n\underline{n}} &= \langle \psi_n^{\rm s}(\vec{r}_i)\psi_{\underline{n}}^{\rm s}(\vec{r}_{\underline{i}})|v_{i\underline{i}}^{\rm ee}|\psi_{\underline{n}}^{\rm s}(\vec{r}_i)\psi_n^{\rm s}(\vec{r}_{\underline{i}})\rangle \\ &= \int_{\text{all } \vec{r}_i}\int_{\text{all } \vec{r}_{\underline{i}}} \psi_n^{\rm s}(\vec{r}_i)^*\psi_{\underline{n}}^{\rm s}(\vec{r}_{\underline{i}})^* \frac{e^2}{4\pi\epsilon_0}\frac{1}{r_{i\underline{i}}}\psi_{\underline{n}}^{\rm s}(\vec{r}_i)\psi_n^{\rm s}(\vec{r}_{\underline{i}})\,{\rm d}^3\vec{r}_i\,{\rm d}^3\vec{r}_{\underline{i}} \end{aligned}$$

is real can be seen by taking complex conjugate, and then noting that the names of the integration variables do not make a difference, so you can swap them.

The same name swap shows that $J_{n\underline{n}}$ and $K_{n\underline{n}}$ are symmetric matrices; $J_{n\underline{n}} = J_{\underline{n}n}$ and $K_{n\underline{n}} = K_{\underline{n}n}$.

That $K_{n\underline{n}}$ is positive is a bit trickier; write it as

$$\int_{\text{all } \vec{r}_i} -ef^*(\vec{r}_i)\left(\int_{\text{all } \vec{r}_{\underline{i}}} \frac{-ef(\vec{r}_{\underline{i}})}{4\pi\epsilon_0}\frac{1}{r_{i\underline{i}}}\,{\rm d}^3\vec{r}_{\underline{i}}\right){\rm d}^3\vec{r}_i$$

with $f = {\psi_{\underline{n}}^{\rm s}}^*\psi_n^{\rm s}$. The part within parentheses is just the potential $V(\vec{r}_i)$ of a distribution of charges with density $-ef$. Sure, f may be complex but that merely means that the potential is too. The electric field is minus the gradient of the potential, $\vec{\mathcal{E}} = -\nabla V$, and according to Maxwell's equation, the divergence of the electric field is the charge density divided by ϵ_0: $\mathrm{div}\,\vec{\mathcal{E}} = -\nabla^2 V = -ef/\epsilon_0$. So $-ef^* = -\epsilon_0\nabla^2 V^*$ and the integral is

$$-\epsilon_0\int_{\text{all } \vec{r}_i} V\nabla^2 V^*\,{\rm d}^3\vec{r}_i$$

and integration by parts shows it is positive. Or zero, if $\psi_{\underline{n}}^{\rm s}$ is zero wherever $\psi_n^{\rm s}$ is not, and vice versa.

To show that $J_{n\underline{n}} \geqslant K_{n\underline{n}}$, note that

$$\langle \psi_n^{\rm s}(\vec{r}_i)\psi_{\underline{n}}^{\rm s}(\vec{r}_{\underline{i}}) - \psi_{\underline{n}}^{\rm s}(\vec{r}_i)\psi_n^{\rm s}(\vec{r}_{\underline{i}})|v^{\rm ee}|\psi_n^{\rm s}(\vec{r}_i)\psi_{\underline{n}}^{\rm s}(\vec{r}_{\underline{i}}) - \psi_{\underline{n}}^{\rm s}(\vec{r}_i)\psi_n^{\rm s}(\vec{r}_{\underline{i}})\rangle$$

is nonnegative, for the same reasons as $J_{n\underline{n}}$ but with $\psi_n^{\rm s}\psi_{\underline{n}}^{\rm s} - \psi_{\underline{n}}^{\rm s}\psi_n^{\rm s}$ replacing $\psi_n^{\rm s}\psi_{\underline{n}}^{\rm s}$. If you multiply out the inner product, you get that $2J_{n\underline{n}} - 2K_{n\underline{n}}$ is nonnegative, so $J_{n\underline{n}} \geqslant K_{n\underline{n}}$.

D.54 Derivation of the Hartree-Fock equations

This note derives the canonical Hartree-Fock equations. It will use some linear algebra; see the Notations section under "matrix" for some basic concepts. The derivation will be performed under the normally stated rules of engagement that the orbitals are of the form $\psi_n^{\rm s}\uparrow$ or $\psi_n^{\rm s}\downarrow$. So the spins are chosen, and only the spatial orbitals $\psi_n^{\rm s}$ are to be found.

The derivations must allow for the fact that in restricted Hartree-Fock, it is required that pairs of spin-up and spin-down orbitals have the same spatial

orbital. So there are three possible kinds of spatial orbitals. A spatial orbital may produce a single unpaired spin orbital that is spin-up, or a single unpaired spin orbital that is spin-down, or a pair of spin-up and spin-down orbitals with the same spatial orbital. These three types of spatial orbitals will be referred to as unpaired spin-up, unpaired spin-down, and restricted. Note that these names do not refer to properties of the spatial orbits themselves, of course, but to the properties of the spin orbits that these spatial orbitals produce.

Assume that there are N_u spin-up spatial orbitals, N_d spin-down ones, and N_r restricted ones. The total number of *spatial* orbitals, call it N, is then

$$N = N_\mathrm{u} + N_\mathrm{d} + N_\mathrm{r}$$

and that is the total number of unknown spatial orbitals to find. A corresponding number of N equations will be needed for them.

However, the total number of *spin* orbitals, I, is larger than N by an additional amount N_r, because the restricted spatial orbitals appear in both spin-up and spin-down versions. That makes the mathematics messy.

Things become a bit easier if the ordering of the orbitals is specified a priori. The ordering makes no difference physically. So it will be assumed that the spatial orbitals are ordered with the unpaired spin-up ones first, the unpaired spin-down ones second, and the restricted ones last. The ordering of the spin orbitals will be the same as that of the spatial orbitals, but with the restricted orbitals at the end appearing twice; first in the spin-up versions and then in the spin-down versions.

To find the spatial orbitals, the variational method as discussed in chapter 9.1.3 says that the expectation energy $\langle E \rangle$ must be unchanged under small changes in the orbitals, provided that the orbitals remain orthonormal. To easily enforce that orthonormality constraint requires that terms are added to the change in orbitals that penalize for any going out of bounds.

To do so, first note that $\langle E \rangle$ can be considered to be a real function from the real and imaginary parts of the spatial orbitals, and both these parts are real functions. The condition that any spatial orbital ψ^s_n must be normalized means that the inner product of the orbital with itself must be 1,

$$\langle \psi^\mathrm{s}_n | \psi^\mathrm{s}_n \rangle = 1$$

This condition is real too. However, the condition that any spatial mode ψ^s_n must be orthogonal to any other spatial mode $\psi^\mathrm{s}_{\underline{n}}$ means that the inner product of the two modes must be zero,

$$\langle \psi^\mathrm{s}_n | \psi^\mathrm{s}_{\underline{n}} \rangle = 0$$

In general this condition has both a real and an imaginary component. But it can be written as two real conditions;

$$\tfrac{1}{2}\Big(\langle \psi^\mathrm{s}_n | \psi^\mathrm{s}_{\underline{n}} \rangle + \langle \psi^\mathrm{s}_{\underline{n}} | \psi^\mathrm{s}_n \rangle \Big) = 0, \qquad \tfrac{1}{2}\mathrm{i}\Big(\langle \psi^\mathrm{s}_n | \psi^\mathrm{s}_{\underline{n}} \rangle - \langle \psi^\mathrm{s}_{\underline{n}} | \psi^\mathrm{s}_n \rangle \Big) = 0.$$

The reason is that if you swap the sides in an inner product, you get the complex conjugate; therefore the first equation above is the real part of the inner product and the second the imaginary part.

Since we now have a completely real problem in real independent variables, the penalty factors (the Lagrangian multipliers) in the problem will be real too. For reasons evident in a second, the penalty factor for the normalization condition above will be called ϵ_{nn}, while the ones for the two real orthogonality conditions will be called $2\epsilon_{n\underline{n},r}$ and $2\epsilon_{n\underline{n},i}$, respectively. To avoid enforcing the same orthogonality condition twice, it is here assumed that $\underline{n} > n$.

The reason for these notations is that in terms of them, the penalized variational condition that the spatial orbitals must satisfy, chapter 9.1.3, takes the simple form

$$\delta \langle E \rangle - \sum_{n=1}^{N} \sum_{\underline{n}=1}^{N} \epsilon_{n\underline{n}} \delta \langle \psi_n^{\rm s} | \psi_{\underline{n}}^{\rm s} \rangle = 0$$

where δ denotes a small change in the following quantity, $\underline{n}$ is now allowed to be both smaller or larger than n, and $\epsilon_{n\underline{n}}$ is a Hermitian matrix, meaning that $\epsilon_{\underline{n}n} = \epsilon_{n\underline{n}}^*$

Note however that two spatial orbitals do not have to be orthogonal if one is a unpaired spin-up one and the other an unpaired spin-down one. In that case the spins take care of orthogonality. This can be accomodated by stipulating that the penalty factors of the corresponding constraints are zero,

$$\epsilon_{n\underline{n}} = 0 \quad \text{if } \psi_n^{\rm s} \text{ is spin-up and } \psi_n^{\rm s} \text{ is spin-down, or vice versa}$$

Next the variational condition is to be evaluated for a small change $\delta\psi_m^{\rm s}$ in a sample spatial wave function $\psi_m^{\rm s}$ where m is no larger than N. This is straightforward for the inner products in the penalty terms. However, the expectation value of energy $\langle E \rangle$ was obtained in chapter 9.3.3 in terms of the spin, rather than spatial orbitals:

$$\begin{aligned} \langle E \rangle \;=\; & \sum_{n=1}^{I} \langle \psi_n^{\rm s} | h^{\rm e} | \psi_n^{\rm s} \rangle \\ & + \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\underline{n}=1}^{I} \langle \psi_n^{\rm s} \psi_{\underline{n}}^{\rm s} | v^{\rm ee} | \psi_n^{\rm s} \psi_{\underline{n}}^{\rm s} \rangle - \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\underline{n}=1}^{I} \langle \psi_n^{\rm s} \psi_{\underline{n}}^{\rm s} | v^{\rm ee} | \psi_{\underline{n}}^{\rm s} \psi_n^{\rm s} \rangle \langle \updownarrow_n | \updownarrow_{\underline{n}} \rangle^2 \end{aligned}$$

(From here on, the argument of the first orbital of a pair in either side of an inner product is taken to be the first inner product integration variable $\vec{r}$ and the argument of the second orbital is the second integration variable $\underline{\vec{r}}$)

Taking that into account, the variational condition for the $\delta\psi_m^{\rm s}$ takes the messy form

$$[2?] \Big(\langle \delta\psi_m^{\rm s} | h^{\rm e} | \psi_m^{\rm s} \rangle + \langle \psi_m^{\rm s} | h^{\rm e} | \delta\psi_m^{\rm s} \rangle \Big)$$

$$+[2?]\tfrac{1}{2}\sum_{\underline{n}=1}^{I}\Big(\langle\delta\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_{\underline{n}}|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_{\underline{n}}\rangle+\langle\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_{\underline{n}}|v^{\mathrm{ee}}|\delta\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_{\underline{n}}\rangle\Big)$$

$$+[2?]\tfrac{1}{2}\sum_{n=1}^{I}\Big(\langle\psi^{\mathrm{s}}_n\delta\psi^{\mathrm{s}}_m|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_n\psi^{\mathrm{s}}_m\rangle+\langle\psi^{\mathrm{s}}_n\psi^{\mathrm{s}}_m|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_n\delta\psi^{\mathrm{s}}_m\rangle\Big)$$

$$-\tfrac{1}{2}\sum_{\underline{n}=1}^{I}\Big(\langle\delta\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_{\underline{n}}|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_{\underline{n}}\psi^{\mathrm{s}}_m\rangle+\langle\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_{\underline{n}}|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_{\underline{n}}\delta\psi^{\mathrm{s}}_m\rangle\Big)[\langle\updownarrow_m|\updownarrow_{\underline{n}}\rangle^2?]$$

$$-\tfrac{1}{2}\sum_{n=1}^{I}\Big(\langle\psi^{\mathrm{s}}_n\delta\psi^{\mathrm{s}}_m|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_n\rangle+\langle\psi^{\mathrm{s}}_n\psi^{\mathrm{s}}_m|v^{\mathrm{ee}}|\delta\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_n\rangle\Big)[\langle\updownarrow_n|\updownarrow_m\rangle^2?]$$

$$-\sum_{\underline{n}=1}^{N}\epsilon_{m\underline{n}}\langle\delta\psi^{\mathrm{s}}_m|\psi^{\mathrm{s}}_{\underline{n}}\rangle-\sum_{n=1}^{N}\epsilon_{nm}\langle\psi^{\mathrm{s}}_n|\delta\psi^{\mathrm{s}}_m\rangle=0$$

Here [2?] means to *insert* a factor 2 there if m is one of the restricted spatial orbitals, because each of the two corresponding spin orbitals produces a term like that. And $[\langle\updownarrow_.|\updownarrow_.\rangle^2?]$ means *leave away* this inner product if m is one of the restricted spatial orbitals, because exactly one of the two corresponding spin orbitals has that inner product equal to one, and the other has it zero.

Note that the difference between $\underline{n}$ and n can from now on be ignored; the name of a summation variable makes no difference for the result, and there are no longer name conflicts in the individual terms. Note also that the sums over n (or $\underline{n}$) with upper limit I include the restricted spatial orbitals *twice*, once for each spin direction.

The second term in each row in the expression above is just the complex conjugate of the first. These second terms can be thrown out using the same trick as in chapter 9.1.3. (In other words, average with the same equation with $\delta\psi^{\mathrm{s}}_m$ replaced by $-\mathrm{i}\delta\psi^{\mathrm{s}}_m$ and divided by i.) And the integrals with the factors $\frac{1}{2}$ are pairwise the same; the difference is just a name swap of the inner product integration variables. So all there is really left is

$$[2?]\langle\delta\psi^{\mathrm{s}}_m|h^{\mathrm{e}}|\psi^{\mathrm{s}}_m\rangle+$$

$$+[2?]\sum_{n=1}^{I}\langle\delta\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_n|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_n\rangle$$

$$-\sum_{n=1}^{I}\langle\delta\psi^{\mathrm{s}}_m\psi^{\mathrm{s}}_n|v^{\mathrm{ee}}|\psi^{\mathrm{s}}_n\psi^{\mathrm{s}}_m\rangle[\langle\updownarrow_n|\updownarrow_m\rangle^2?]$$

$$-\sum_{n=1}^{N}\epsilon_{mn}\langle\delta\psi^{\mathrm{s}}_m|\psi^{\mathrm{s}}_n\rangle$$

Now write out the inner product over the first position coordinate $\vec{r}$, being the argument of $\delta\psi^{\rm s}_m$, for all terms:

$$\int_{\text{all } \vec{r}} \delta\psi^{\rm s}_m{}^* \Big(\\ [2?]h^{\rm e}\psi^{\rm s}_m \\ +[2?]\sum_{n=1}^{I}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_n\rangle\psi^{\rm s}_m \\ -\sum_{n=1}^{I}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_m\rangle\psi^{\rm s}_n[\langle\updownarrow_n|\updownarrow_m\rangle^2?] \\ -\sum_{n=1}^{N}\epsilon_{mn}\psi^{\rm s}_n \\ \Big)\,{\rm d}^3\vec{r} = 0$$

If this integral is to be zero for *whatever* is $\delta\psi^{\rm s}_m$, then the terms within the parentheses must be zero. (Otherwise just take $\delta\psi^{\rm s}_m$ proportional to the parenthetical expression; you would get the norm of the expression, and that is only zero if the expression is.)

Unavoidably then, the following equations, one for each value of m, must be satisfied:

$$[2?]h^{\rm e}\psi^{\rm s}_m + [2?]\sum_{n=1}^{I}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_n\rangle\psi^{\rm s}_m - \sum_{n=1}^{I}[\langle\updownarrow_n|\updownarrow_m\rangle^2?]\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_m\rangle\psi^{\rm s}_n = \sum_{n=1}^{N}\epsilon_{mn}\psi^{\rm s}_n$$

This can be cleaned up a bit by dividing by [2?]:

$$\boxed{h^{\rm e}\psi^{\rm s}_m + \sum_{n=1}^{I}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_n\rangle\psi^{\rm s}_m - \sum_{n=1}^{I}\left\{\begin{matrix}\langle\updownarrow_n|\updownarrow_m\rangle^2\\ \frac{1}{2}\end{matrix}\right\}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_m\rangle\psi^{\rm s}_n = \left\{\begin{matrix}1\\ \frac{1}{2}\end{matrix}\right\}\sum_{n=1}^{N}\epsilon_{mn}\psi^{\rm s}_n} \tag{D.35}$$

These are the general Hartree-Fock equations, one for each $m \leqslant N$. The upper value between braces applies if the spatial orbital $\psi^{\rm s}_m$ is not a restricted one; otherwise the lower value applies. Recall that the sums with upper limit I include the restricted spatial orbitals twice. And that ϵ_{mn} is zero if spatial orbital $\psi^{\rm s}_m$ is unpaired spin-up and $\psi^{\rm s}_n$ unpaired spin-down or vice-versa. For such index values, $\langle\updownarrow_m|\updownarrow_n\rangle$ is zero too.

Note that the general Hartree-Fock equation above includes N "eigenvalues" ϵ_{mn}. The canonical equations include just a single eigenvalue ϵ_m. So to get the

canonical Hartree-Fock equations, the sum in the right hand side must be further simplified to the form $\epsilon_m \psi^{\rm s}_m$.

The restricted closed-shell Hartree-Fock case will be done first, since it is the easiest one. Every spatial orbital is restricted, so the lower choice in the curly brackets always applies. The summation upper limits I, being the number of spin orbitals, can be reduced to the number of spatial orbitals N by adding a factor 2. We can also get rid of the factor $\frac{1}{2}$ in front of the ϵ_{mn} by simply redefining them by that factor. So for restricted closed-shell Hartree-Fock

$$h^{\rm e}\psi^{\rm s}_m + 2\sum_{n=1}^{N}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_n\rangle\psi^{\rm s}_m - \sum_{n=1}^{N}\langle\psi^{\rm s}_n|v^{\rm ee}|\psi^{\rm s}_m\rangle\psi^{\rm s}_n = \sum_{n=1}^{N}\epsilon_{mn}\psi^{\rm s}_n$$

Now the reason why all these ϵ_{mn} are there is because the set of N spatial orbitals that gives the lowest energy state are not unique. The equation above applies to a typical set. Only a special set will get rid of the ϵ_{mn} for n not equal to m, leaving only ϵ_{mm}, which can then be defined to be ϵ_m.

Each orbital in the special set will be some combination of the orbitals in the typical set above. In particular, any orbital in the special set, call it $\overline{\psi}^{\rm s}_\nu$, will be a linear combination of the orbitals $\psi^{\rm s}_n$ in the typical set as follows:

$$\overline{\psi}^{\rm s}_\nu \equiv \sum_{n=1}^{N} c_{n,\nu}\psi^{\rm s}_n \qquad \text{for any } \nu = 1, 2, \ldots, N$$

where the numbers $c_{1,\nu}, c_{2,\nu}, \ldots$ are the multiples of the typical orbitals $\psi^{\rm s}_1$, $\psi^{\rm s}_2$, ... The complete set of numbers $c_{n.\nu}$ for all possible values of both n and ν can be written as a "matrix," a table of numbers. This matrix will be indicated by C. The first index in $c_{n,\nu}$, n, says what row in C that coefficient is in, and the second index, ν, what column.

The multiples $c_{n,\nu}$ cannot be arbitrary, because the special orbitals must still be orthonormal. As noted earlier, they will be if they are normalized (so the inner product of any orbital with itself is 1), and mutually orthogonal (so the inner product of any orbital with any other one is zero). In short, the requirement is that

$$\langle\overline{\psi}^{\rm s}_\mu|\overline{\psi}^{\rm s}_\nu\rangle = \delta_{\mu\nu}$$

where $\delta_{\mu\nu}$ is one if $\mu = \nu$, and zero otherwise. The set of numbers $\delta_{\mu\nu}$ is called the "Kronecker delta" or "unit matrix" or "identity matrix." (The identity matrix is for matrices what the number 1 is for normal numbers; multiplying an arbitrary matrix or vector by the identity matrix does not change that matrix or vector.)

Substituting in the expression for the special orbitals above, making sure not to use the same name ν for two different indices, the requirement becomes

$$\sum_{m=1}^{N}\sum_{n=1}^{N}\langle c_{m,\mu}\psi^{\rm s}_m|c_{n,\nu}\psi^{\rm s}_n\rangle = \delta_{\mu\nu}$$

or noting that numbers come out of the left side of an inner product as complex conjugates,

$$\sum_{m=1}^{N}\sum_{n=1}^{N} c^*_{m,\mu} c_{n,\nu} \langle \psi^{\rm s}_m | \psi^{\rm s}_n \rangle = \delta_{\mu\nu}$$

Now since the set of typical orbitals $\psi^{\rm s}_n$ are already orthonormal, the inner product in the requirement above is only nonzero when m is n, and then it is one. So dropping the zero terms that have $m \neq n$, the requirement on the coefficients simplifies to

$$\sum_{n=1}^{N} c^*_{n,\mu} c_{n,\nu} = \delta_{\mu\nu}$$

What does that mean? Well, for given values of μ and ν, consider the coefficients $c_{n,\mu}$ to form a vector $\vec{u}_\mu$, where n indicates the component number of that vector. Similarly, consider the coefficients $c_{n,\nu}$ to form a vector $\vec{u}_\nu$. Then the left hand side in the requirement above is the inner (or dot, if real) product of these two vectors. So the set of vectors must be orthonormal, just like the special orbitals must be orthonormal. So the matrix of coefficients C must consist of orthonormal vectors. Mathematicians call such matrices "unitary," rather than orthonormal, since it is easily confused with "unit," and that keeps mathematicians in business explaining all the confusion.

The Hermitian adjoint matrix $C^\dagger$ of C is defined as the matrix you get by swapping the order of the indices of the elements of C and adding a complex conjugate. So by definition the factor $c^*_{n,\mu}$ in the requirement above equals the coefficient $c^\dagger_{\mu,n}$ of $C^\dagger$. And matrix multiplication is defined such that then the sum over n in the requirement gives exactly the coefficients of the product $C^\dagger C$. So the requirement above can be written as

$$C^\dagger C = I$$

where I is the unit matrix. That means $C^\dagger$ is the inverse matrix to C, $C^\dagger = C^{-1}$. Then you also have that C is the inverse of $C^\dagger$, $CC^\dagger = I$, which writes out to

$$\sum_{\nu=1}^{N} c_{n,\nu} c^*_{m,\nu} = \delta_{mn}.$$

This can be used to find the typical orbitals in terms of the special ones. To do so, premultiply the expression for the special orbitals as given earlier by $c^*_{m,\nu}$ and sum over ν:

$$\sum_{\nu=1}^{N} c^*_{m,\nu} \overline{\psi}^{\rm s}_\nu = \sum_{\nu=1}^{N} c^*_{m,\nu} \sum_{n=1}^{N} c_{n\nu} \psi^{\rm s}_n$$

As seen above, the sum over ν in the right hand side is just δ_{mn}, so in the sum over n, only the term with n equal to m is nonzero:

$$\sum_{\nu=1}^{N} c^*_{m,\nu} \overline{\psi}^{\rm s}_\nu = \psi^{\rm s}_m$$

That then gives any typical orbital $\psi^{\rm s}_m$ in terms of a sum of the special orbitals $\overline{\psi}^{\rm s}_\nu$.

Now plug that into the non canonical restricted closed-shell Hartree-Fock equations given earlier. Be careful not to use the same summation index name twice in the same term; this derivation will use

$$\psi^{\rm s}_m = \sum_{\nu=1}^{N} c^*_{m,\nu} \overline{\psi}^{\rm s}_\nu \qquad \psi^{\rm s}_n = \sum_{\lambda=1}^{N} c^*_{n,\lambda} \overline{\psi}^{\rm s}_\lambda \qquad \psi^{\rm s}_n = \sum_{\kappa=1}^{N} c^*_{n,\kappa} \overline{\psi}^{\rm s}_\kappa$$

for $\psi^{\rm s}_m$, the first occurrence of $\psi^{\rm s}_n$ in the terms, and the second occurrence, respectively. Premultiply it all by C, i.e. put $\sum_{m=1}^{N} c_{m,\mu}$ in front of each term. That cleans up to

$$h^{\rm e} \overline{\psi}^{\rm s}_\mu + 2 \sum_{\lambda=1}^{N} \langle \overline{\psi}^{\rm s}_\lambda | v^{\rm ee} | \overline{\psi}^{\rm s}_\lambda \rangle \overline{\psi}^{\rm s}_\mu - \sum_{\lambda=1}^{N} \langle \overline{\psi}^{\rm s}_\lambda | v^{\rm ee} | \overline{\psi}^{\rm s}_\mu \rangle \overline{\psi}^{\rm s}_\lambda = \sum_{m=1}^{N} \sum_{n=1}^{N} \sum_{\lambda=1}^{N} c_{m,\mu} \epsilon_{mn} c^*_{n,\lambda} \overline{\psi}^{\rm s}_\lambda$$

Note that the only thing that has changed more than just by symbol names is the matrix in the right hand side. Now for each separate value of λ, take $c^*_{n\lambda}$ as the λ-th orthonormal eigenvector of Hermitian matrix ϵ_{mn}, calling the eigenvalue ϵ_λ. Then by the definition of eigenvector,

$$\sum_{n=1}^{N} \epsilon_{mn} c^*_{n,\lambda} = \epsilon_\lambda c^*_{m,\lambda}$$

So the right hand side becomes

$$\sum_{m=1}^{I} \sum_{\lambda=1}^{N} c_{m,\mu} \epsilon_\lambda c^*_{m,\lambda} \overline{\psi}^{\rm s}_\lambda = \sum_{\lambda=1}^{N} \delta_{\mu\lambda} \epsilon_\lambda \overline{\psi}^{\rm s}_\lambda = \epsilon_\mu \overline{\psi}^{\rm s}_\mu$$

So, in terms of the special orbitals defined by the requirement that $c^*_{m,\mu}$ gives the μ-th eigenvector of ϵ_{mn}, the right hand side simplifies to the canonical one.

Since the old typical orbitals are no longer of interest, the overlines on the special orbitals can be dropped to save typing, and the Greek index names μ and λ can be renamed n and $\underline{n}$. That then finally produces the canonical closed-shell restricted Hartree-Fock equations:

$$\boxed{h^{\rm e} \psi^{\rm s}_n + 2 \sum_{\underline{n}=1}^{N} \langle \psi^{\rm s}_{\underline{n}} | v^{\rm ee} | \psi^{\rm s}_{\underline{n}} \rangle \psi^{\rm s}_n - \sum_{\underline{n}=1}^{N} \langle \psi^{\rm s}_{\underline{n}} | v^{\rm ee} | \psi^{\rm s}_n \rangle \psi^{\rm s}_{\underline{n}} = \epsilon_n \psi^{\rm s}_n} \tag{D.36}$$

Note that the left-hand side directly provides a Hermitian Fock operator if you identify it as $\mathcal{F}\psi^{\mathrm{s}}_n$; there is no need to involve spin in the closed-shell restricted case. This also provides a much simpler explanation than all the algebra above why all the earlier ϵ_{mn} with $m \neq n$ were not needed; existence of a set of orthonormal eigenfunctions of a Hermitian operator is automatic. So there is no *fundamental* need to enforce that separately through Lagrangian multipliers.

Turning now to the case of (fully) unrestricted Hartree-Fock (UHF), you might make the same simple argument as above and be done. But it is worthwhile to go through the full mathematics anyway, to better understand open-shell restricted Hartree-Fock later. In the unrestricted case, the non canonical equations are

$$h^{\mathrm{e}}\psi^{\mathrm{s}}_m + \sum_{n=1}^{N} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_n \rangle \psi^{\mathrm{s}}_m - \sum_{n=1}^{N} \langle \updownarrow_n | \updownarrow_m \rangle^2 \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_m \rangle \psi^{\mathrm{s}}_n = \sum_{n=1}^{N} \epsilon_{mn} \psi^{\mathrm{s}}_n$$

In this case, there are two different types of spatial orbitals; those appearing in spin-up spin orbitals, and those appearing in spin-down spin orbitals. You cannot just make arbitrary combinations of all these orbitals. If you combine spin-up and spin-down orbitals, they correspond to spin orbitals of uncertain spin. That would make the assumptions used to derive the Hartree-Fock equations invalid.

However, combinations of purely spin-up orbitals can still be made without problems, and so can combinations of purely spin down orbitals. To do the mathematics, the spatial orbitals can be separated into two sets. The set of orbital numbers n corresponding to spin-up spin orbitals will be indicated by U, and the set of numbers n corresponding to spin-down spin orbitals by D. So you can partition (separate) the non canonical equations above into equations for $m \in \mathrm{U}$ (meaning m is one of the values in set U),

$$h^{\mathrm{e}}\psi^{\mathrm{s}}_m + \sum_{n\in\mathrm{U}} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_n \rangle \psi^{\mathrm{s}}_m + \sum_{n\in\mathrm{D}} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_n \rangle \psi^{\mathrm{s}}_m - \sum_{n\in\mathrm{U}} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_m \rangle \psi^{\mathrm{s}}_n = \sum_{n\in\mathrm{U}} \epsilon_{mn} \psi^{\mathrm{s}}_n$$

and equations for $m \in \mathrm{D}$,

$$h^{\mathrm{e}}\psi^{\mathrm{s}}_m + \sum_{n\in\mathrm{U}} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_n \rangle \psi^{\mathrm{s}}_m + \sum_{n\in\mathrm{D}} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_n \rangle \psi^{\mathrm{s}}_m - \sum_{n\in\mathrm{D}} \langle \psi^{\mathrm{s}}_n | v^{\mathrm{ee}} | \psi^{\mathrm{s}}_m \rangle \psi^{\mathrm{s}}_n = \sum_{n\in\mathrm{D}} \epsilon_{mn} \psi^{\mathrm{s}}_n$$

In these two types of equations, the fact that the up and down spin states are orthogonal was used to get rid of one pair of sums, and another pair was eliminated by the fact that there are no Lagrangian variables ϵ_{mn} linking the sets, since the spatial orbitals in the two sets are allowed to be mutually non orthogonal.

Now separately replace the orbitals of the up and down states by a modified set just like for the restricted closed-shell case above, for each using the unitary

matrix of eigenvectors of the ϵ_{mn} coefficients appearing in the right hand side of the equations for that set. It leaves the equations intact except for changes in names, but gets rid of the ϵ_{mn} for $m \neq n$, leaving only ϵ_{mm} values, call them ϵ_m. Then combine the spin-up and spin-down equations again into a single expression. You get, in terms of revised symbol names,

$$\boxed{h^{\rm e}\psi^{\rm s}_n + \sum_{\underline{n}=1}^{N} \langle \psi^{\rm s}_{\underline{n}} | v^{\rm ee} | \psi^{\rm s}_{\underline{n}} \rangle \psi^{\rm s}_n - \sum_{\underline{n}=1}^{N} \langle \updownarrow_{\underline{n}} | \updownarrow_n \rangle^2 \langle \psi^{\rm s}_{\underline{n}} | v^{\rm ee} | \psi^{\rm s}_n \rangle \psi^{\rm s}_{\underline{n}} = \epsilon_n \psi^{\rm s}_n} \tag{D.37}$$

That leaves only the restricted open-shell Hartree-Fock method. Here, the partitioning also needs to include the set R of of restricted orbitals besides U and D. There is now a problem, because you cannot make combinations of restricted orbitals with spin-up or spin-down orbitals. That means that the ϵ_{mn} values where either m or n is restricted and the other is not, cannot be eliminated. Solutions range from just ignoring the whole thing to properly accounting for these ϵ_{mn} values by enforcing that restricted and non restricted orbitals must stay orthogonal as additional equations. This (even more) elaborate case will be left to the references that you can find in [46], in particular [28, pp. 242-253].

Woof.

D.55 Why the Fock operator is Hermitian

To verify that the Fock operator is Hermitian, first note that $h^{\rm e}$ is Hermitian since it is an Hamiltonian. Next if you form the inner product $\langle \overline{\psi^{\rm e}\updownarrow} | v^{\rm HF} \psi^{\rm s} \updownarrow \rangle$, the first term in $v^{\rm HF}$, the Coulomb term, can be taken to the other side since it is just a real function. The second term, the exchange one, produces the inner product,

$$-\sum_{\underline{n}=1}^{I} \left\langle \overline{\psi^{\rm e}(\vec{r})\updownarrow(S_z)} \middle| \langle \psi^{\rm s}_{\underline{n}}(\vec{\underline{r}})\updownarrow_{\underline{n}}(S_{z1}) | v^{\rm ee} | \psi^{\rm s}(\vec{\underline{r}})\updownarrow(S_{z1}) \rangle \psi^{\rm s}_{\underline{n}}(\vec{r})\updownarrow_{\underline{n}}(S_z) \right\rangle$$

and if you take the operator to the other side, you get

$$-\sum_{\underline{n}=1}^{I} \left\langle \langle \psi^{\rm s}_{\underline{n}}(\vec{r})\updownarrow_{\underline{n}}(S_z) | v^{\rm ee} | \overline{\psi^{\rm e}(\vec{r})\updownarrow(S_z)} \rangle \psi^{\rm s}_{\underline{n}}(\vec{\underline{r}})\updownarrow_{\underline{n}}(S_{z1}) \middle| \psi^{\rm s}(\vec{\underline{r}})\updownarrow(S_{z1}) \right\rangle$$

and writing out these inner products as six-dimensional spatial integrals and sums over spin, you see that they are the same.

D.56 Number of system eigenfunctions

This note derives the number of energy eigenfunctions $Q_{\vec{I}}$ for a given set $\vec{I} = (I_1, I_2, I_3, \ldots)$ of shelf occupation numbers, I_s being the number of particles on

shelf number s. The number of single-particle eigenfunctions on shelf number s is indicated by N_s.

Consider first the case of distinguishable particles, referring to figure 11.1 for an example. The question is how many different eigenfunctions can be created with the given shelf numbers. What are the ways to create different ones? Well, the first choice that can be made is what are the I_1 particles that go on shelf 1. If you pick out I_1 particles from the I total particles, you have I choices for particle 1, next there are $I-1$ choices left for particle 2, then $I-2$ for particle 3. The total number of possible ways of choosing the I_1 particles is then

$$I \times (I-1) \times (I-2) \times \ldots \times (I-I_1+1)$$

However, this overestimates the number of variations in eigenfunctions that you can create by selecting the I_1 particles: the only thing that makes a difference for the eigenfunctions is *what* particles you pick to go on shelf 1; the *order* in which you chose to pick them out of the total set of I makes no difference. If you chose a set of I_1 particles in an arbitrary order, you get no difference in eigenfunction compared to the case that you pick out the same particles sorted by number. To correct for this, the number of eigenfunction variations above must be divided by the number of different orderings in which a set of I_1 particles can come out of the total collection. That will give the number of different *sets* of particles, sorted by number, that can be selected. The number of ways that a set of I_1 particles can be ordered is

$$I_1! = I_1 \times (I_1-1) \times (I_1-2) \times \ldots \times 3 \times 2 \times 1;$$

there are I_1 possibilities for the particle that comes first in the sorted set, then I_1-1 possibilities left for the particle that comes second, etcetera. Dividing the earlier expression by $I_1!$, the number of different sets of I_1 particles that can be selected for shelf 1 becomes

$$\frac{I \times (I-1) \times (I-2) \times \ldots \times (I-I_1+1)}{I_1 \times (I_1-1) \times (I_1-2) \times \ldots \times 3 \times 2 \times 1}.$$

But further variations in eigenfunctions are still possible in the way these I_1 particles are distributed over the N_1 single-particle states on shelf 1. There are N_1 possible single-particle states for the first particle of the sorted set, times N_1 possible single-particle states for the second particle, etcetera, making a total of $N_1^{I_1}$ variations. That number of variations exists for each of the individual sorted sets of particles, so the total number of variations in eigenfunctions is the product:

$$N_1^{I_1} \frac{I \times (I-1) \times (I-2) \times \ldots \times (I-I_1+1)}{I_1 \times (I_1-1) \times (I_1-2) \times \ldots \times 3 \times 2 \times 1}.$$

This can be written more concisely by noting that the bottom of the fraction is per definition $I_1!$ while the top equals $I!/(I-I_1)!$: note that the terms missing

from $I!$ in the top are exactly $(I - I_1)!$. (In the special case that $I = I_1$, all particles on shelf 1, this still works since mathematics defines $0! = 1$.) So, the number of variations in eigenfunctions so far is:

$$N_1^{I_1} \frac{I!}{I_1!(I - I_1)!}.$$

The fraction is known in mathematics as "I choose I_1."

Further variations in eigenfunctions are possible in the way that the I_2 particles on shelf 2 are chosen and distributed over the single-particle states on that shelf. The analysis is just like the one for shelf 1, except that shelf 1 has left only $I - I_1$ particles for shelf 2 to chose from. So the number of additional variations related to shelf 2 becomes

$$N_2^{I_2} \frac{(I - I_1)!}{I_2!(I - I_1 - I_2)!}.$$

The same way the number of eigenfunction variations for shelves 3, 4, ... can be found, and the grand total of different eigenfunctions is

$$N_1^{I_1} \frac{I!}{I_1!(I - I_1)!} \times N_2^{I_2} \frac{(I - I_1)!}{I_2!(I - I_1 - I_2)!} \times N_3^{I_3} \frac{(I - I_1 - I_2)!}{I_3!(I - I_1 - I_2 - I_3)!} \times \ldots$$

This terminates at the shelf number S beyond which there are no more particles left, when $I - I_1 - I_2 - I_3 - \ldots - I_B = 0$. All further shelves will be empty. Empty shelves might just as well not exist, they do not change the eigenfunction count. Fortunately, there is no need to exclude empty shelves from the mathematical expression above, it can be used either way. For example, if shelf 2 would be empty, e.g. $I_2 = 0$, then $N_2^{I_2} = 1$ and $I_2! = 1$, and the factors $(I - I_1)!$ and $(I - I_1 - I_2)!$ cancel each other. So the factor due to empty shelf 2 becomes multiplying by one, it does not change the eigenfunction count.

Note that various factors cancel in the eigenfunction count above, it simplifies to the final expression

$$Q^{\mathrm{d}}_{\vec{I}} = I! \frac{N_1^{I_1}}{I_1!} \times \frac{N_2^{I_2}}{I_2!} \times \frac{N_3^{I_3}}{I_3!} \times \ldots$$

Mathematicians like to symbolically write a product of indexed factors like this using the product symbol Π:

$$Q^{\mathrm{d}}_{\vec{I}} = I! \prod_{\text{all } s} \frac{N_s^{I_s}}{I_s!}.$$

It means exactly the same as the written-out product.

Next the eigenfunction count for fermions. Refer now to figure 11.3. For any shelf s, it is given that there are I_s particles on that shelf, and the only

variations in eigenfunctions that can be achieved are in the way that these particles are distributed over the N_s single-particle eigenfunctions on that shelf. The fermions are identical, but to simplify the reasoning, for now assume that you stamp numbers on them from 1 to I_s. Then fermion 1 can go into N_s single-particle states, leaving $N_s - 1$ states that fermion 2 can go into, then $N_s - 2$ states that fermion 3 can go into, etcetera. That produces a total of

$$N_s \times (N_s - 1) \times (N_s - 2) \times \ldots \times (N_s - I_s + 1) = \frac{N_s!}{(N_s - I_s)!}$$

variations. But most of these differ only in the order of the numbers stamped on the fermions; differences in the numbers stamped on the electrons do not constitute a difference in eigenfunction. The only difference is in whether a state is occupied by a fermion or not, not what number is stamped on it. Since, as explained under distinguishable particles, the number of ways I_s particles can be ordered is $I_s!$, it follows that the formula above over-counts the number of variations in eigenfunctions by that factor. To correct, divide by $I_s!$, giving the number of variations as $N_s!/(N_s - I_s)!I_s!$, or "N_s choose I_s." The combined number of variations in eigenfunctions for all shelves then becomes

$$Q^{\rm f}_{\vec{I}} = \frac{N_1!}{(N_1 - I_1)!I_1!} \times \frac{N_2!}{(N_2 - I_2)!I_2!} \times \frac{N_3!}{(N_3 - I_3)!I_3!} \times \ldots = \prod_{\text{all } s} \frac{N_s!}{(N_s - I_s)!I_s!}.$$

If a shelf is empty, it makes again no difference; the corresponding factor is again one. But another restriction applies for fermions: there should not be any eigenfunctions if any shelf number I_s is greater than the number of states N_s on that shelf. There can be at most one particle in each state. Fortunately, mathematics defines factorials of negative integer numbers to be infinite, and the infinite factor $(N_s - I_s)!$ in the bottom will turn the eigenfunction count into zero as it should. The formula can be used whatever the shelf numbers are.



Figure D.3: Schematic of an example boson distribution on a shelf.

Last but not least, the eigenfunction count for bosons. Refer now to figure 11.2. This one is tricky, but a trick solves it. To illustrate the idea, take shelf 2 in figure 11.2 as an example. It is reproduced in condensed form in figure D.3. The figure merely shows the particles and the lines separating the single-particle states. Like for the fermions, the question is, how many ways can the I_s bosons be arranged inside the N_s single-particle states? In other words, how many variations are there on a schematic like the one shown in figure D.3? To figure it out, stamp identifying numbers on all the elements, particles and single-state separating lines alike, ranging from 1 to $I_s + N_s - 1$. Following

the same reasoning as before, there are $(I_s + N_s - 1)!$ different ways to order these numbered objects. As before, now back off. All the different orderings of the numbers stamped on the bosons, $I_s!$ of them, produce no difference in eigenfunction, so divide by $I_s!$ to fix it up. Similarly, all the different orderings of the single-particle state boundaries produce no difference in eigenfunction, so divide by $(N_s - 1)!$. The number of variations in eigenfunctions possible by rearranging the particles on a single shelf s is then $(I_s + N_s - 1)!/I_s!(N_s - 1)!$. The total for all shelves is

$$\begin{aligned} Q^{\rm b}_{\vec{I}} &= \frac{(I_1 + N_1 - 1)!}{I_1!(N_1 - 1)!} \times \frac{(I_2 + N_2 - 1)!}{I_2!(N_2 - 1)!} \times \frac{(I_3 + N_3 - 1)!}{I_3!(N_3 - 1)!} \times \ldots \\ &= \prod_{\text{all } s} \frac{(I_s + N_s - 1)!}{I_s!(N_s - 1)!}. \end{aligned}$$

D.57 The particle energy distributions

This note derives the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein energy distributions of weakly interacting particles for a system for which the net energy is precisely known.

The objective is to find the shelf numbers $\vec{I} = (I_1, I_2, I_3, \ldots)$ for which the number of eigenfunctions $Q_{\vec{I}}$ is maximal. Actually, it is mathematically easier to find the maximum of $\ln(Q_{\vec{I}})$, and that is the same thing: if $Q_{\vec{I}}$ is as big as it can be, then so is $\ln(Q_{\vec{I}})$. The advantage of working with $\ln(Q_{\vec{I}})$ is that it simplifies all the products in the expressions for the $Q_{\vec{I}}$ derived in derivation {D.56} into sums: mathematics says that $\ln(ab)$ equals $\ln(a)$ plus $\ln(b)$ for any (positive) a and b.

It will be assumed, following derivation {N.24}, that if the maximum value is found among *all* shelf occupation numbers, whole numbers or not, it suffices. More daringly, errors less than a particle are not going to be taken seriously.

In finding the maximum of $\ln(Q_{\vec{I}})$, the shelf numbers cannot be completely arbitrary; they are constrained by the conditions that the sum of the shelf numbers must equal the total number of particles I, and that the particle energies must sum together to the given total energy E:

$$\sum_s I_s = I \qquad \sum_s I_s E^{\rm p}_s = E.$$

Mathematicians call this a constrained maximization problem.

According to calculus, without the constraints, you can just put the derivatives of $\ln(Q_{\vec{I}})$ with respect to all the shelf numbers I_s to zero to find the maximum. With the constraints, you have to add "penalty terms" that correct for any going out of bounds, {D.48}, and the correct function whose derivatives

must be zero is

$$F = \ln(Q_{\vec{I}}) - \epsilon_1 \Big(\sum_s I_s - I \Big) - \epsilon_2 \Big(\sum_s I_s E_s^{\mathrm{p}} - E \Big)$$

where the constants ϵ_1 and ϵ_2 are unknown penalty factors called the Lagrangian multipliers.

At the shelf numbers for which the number of eigenfunctions is largest, the derivatives $\partial F / \partial I_s$ must be zero. However, that condition is difficult to apply exactly, because the expressions for $Q_{\vec{I}}$ as given in the text involve the factorial function, or rather, the gamma function. The gamma function does not have a simple derivative. Here typical textbooks will flip out the Stirling approximation of the factorial, but this approximation is simply incorrect in parts of the range of interest, and where it applies, the error is unknown.

It is a much better idea to approximate the differential quotient by a difference quotient, as in

$$0 = \frac{\partial F}{\partial I_s} \approx \frac{\Delta F}{\Delta I_s} \equiv$$
$$\frac{F(I_1, I_2, \ldots, I_{s-1}, I_s + 1, I_{s+1}, \ldots) - F(I_1, I_2, \ldots, I_{s-1}, I_s, I_{s+1}, \ldots)}{I_s + 1 - I_s}.$$

This approximation is very minor, since according to the so-called mean value theorem of mathematics, the location where $\Delta F / \Delta I_s$ is zero is at most one particle away from the desired location where $\partial F / \partial I_s$ is zero. Better still, $I_s + \frac{1}{2} \equiv I_{s,\mathrm{best}}$ will be no more that half a particle off, and the analysis already had to commit itself to ignoring fractional parts of particles anyway. The difference quotient leads to simple formulae because the gamma function satisfies the condition $(n+1)! = (n+1)\, n!$ for any value of n, compare the notations section under "!".

Now consider first distinguishable particles. The function F to differentiate is defined above, and plugging in the expression for $Q_{\vec{I}}^{\mathrm{d}}$ as found in derivation {D.56} produces

$$F = \ln(I!) + \sum_s \left[I_s \ln(N_s) - \ln(I_s!) \right] - \epsilon_1 \Big(\sum_s I_s - I \Big) - \epsilon_2 \Big(\sum_s I_s E_s^{\mathrm{p}} - E \Big)$$

For any value of the shelf number s, in the limit $I_s \downarrow -1$, F tends to negative infinity because $I_s!$ tends to positive infinity in that limit and its logarithm appears with a minus sign. In the limit $I_s \uparrow +\infty$, F tends once more to negative infinity, since $\ln(I_s!)$ for large values of I_s is according to the so-called Stirling formula approximately equal to $I_s \ln(I_s) - I_s$, so the $-\ln(I_s!)$ term in F goes to minus infinity more strongly than the terms proportional to I_s might go to plus infinity. If F tends to minus infinity at both ends of the range $-1 < I_s < \infty$, there must be a maximum value of F somewhere within that range where

the derivative with respect to I_s is zero. More specifically, working out the difference quotient:

$$\frac{\Delta F}{\Delta I_s} = \ln(N_s) - \ln(I_s + 1) - \epsilon_1 - \epsilon_2 E_s^{\mathrm{p}} = 0$$

and $-\ln(I_s+1)$ is infinity at $I_s = -1$ and minus infinity at $I_s = \infty$. Somewhere in between, $\Delta F/\Delta I_s$ will cross zero. In particular, combining the logarithms and then taking an exponential, the best estimate for the shelf occupation number is

$$I_{s,\mathrm{best}} = I_s + \tfrac{1}{2} = \frac{N_s}{e^{\epsilon_2 E_s^{\mathrm{p}} + \epsilon_1}} - \tfrac{1}{2}$$

The correctness of the final half particle is clearly doubtful within the made approximations. In fact, it is best ignored since it only makes a difference at high energies where the number of particles per shelf becomes small, and surely, the correct probability of finding a particle must go to zero at infinite energies, not to minus half a particle! Therefore, the best estimate $\iota^{\mathrm{d}} \equiv I_{s,\mathrm{best}}/N_s$ for the number of particles per single-particle energy state becomes the Maxwell-Boltzmann distribution. Note that the derivation might be off by a particle for the lower energy shelves. But there are a lot of particles in a macroscopic system, so it is no big deal.

The case of identical fermions is next. The function to differentiate is now

$$\begin{aligned} F &= \sum_s \left[\ln(N_s!) - \ln(I_s!) - \ln((N_s - I_s)!)\right] \\ &\quad - \epsilon_1\Big(\sum_s I_s - I\Big) - \epsilon_2\Big(\sum_s I_s E_s^{\mathrm{p}} - E\Big) \end{aligned}$$

This time F is minus infinity when a shelf number reaches $I_s = -1$ or $I_s = N_s + 1$. So there must be a maximum to F when I_s varies between those limits. The difference quotient approximation produces

$$\frac{\Delta F}{\Delta I_s} = -\ln(I_s + 1) + \ln(N_s - I_s) - \epsilon_1 - \epsilon_2 E_s^{\mathrm{p}} = 0$$

which can be solved to give

$$I_{s,\mathrm{best}} = I_s + \tfrac{1}{2} = \frac{N_s}{e^{\epsilon_2 E_s^{\mathrm{p}} + \epsilon_1} + 1} + \tfrac{1}{2}\frac{1 - e^{\epsilon_2 E_s^{\mathrm{p}} + \epsilon_1}}{1 + e^{\epsilon_2 E_s^{\mathrm{p}} + \epsilon_1}}$$

The final term, less than half a particle, is again best left away, to ensure that $0 \leqslant I_{s,\mathrm{best}} \leqslant N_s$ as it should. That gives the Fermi-Dirac distribution.

Finally, the case of identical bosons, is, once more, the tricky one. The function to differentiate is now

$$\begin{aligned} F &= \sum_s \left[\ln((I_s + N_s - 1)!) - \ln(I_s!) - \ln((N_s - 1)!)\right] \\ &\quad -\epsilon_1\Big(\sum_s I_s - I\Big) - \epsilon_2\Big(\sum_s I_s E_s^{\mathrm{p}} - E\Big) \end{aligned}$$

For now, assume that $N_s > 1$ for all shelves. Then F is again minus infinity for $I_s = -1$. For $I_s \uparrow \infty$, however, F will behave like $-(\epsilon_1 + \epsilon_2 E_s^{\rm p})I_s$. This tends to minus infinity if $\epsilon_1 + \epsilon_2 E_s^{\rm p}$ is positive, so for now assume it is. Then the difference quotient approximation produces

$$\frac{\Delta F}{\Delta I_s} = \ln(I_s + N_s) - \ln(I_s + 1) - \epsilon_1 - \epsilon_2 E_s^{\rm p} = 0$$

which can be solved to give

$$I_{s,\rm best} = I_s + \tfrac{1}{2} = \frac{N_s - 1}{e^{\epsilon_2 E_s^{\rm p} + \epsilon_1} - 1} - \tfrac{1}{2}.$$

The final half particle is again best ignored to get the number of particles to become zero at large energies. Then, if it is assumed that the number N_s of single-particle states on the shelves is large, the Bose-Einstein distribution is obtained. If N_s is not large, the number of particles could be less than the predicted one by up to a factor 2, and if N_s is one, the entire story comes part. And so it does if $\epsilon_1 + \epsilon_2 E_s^{\rm p}$ is not positive.

Before addressing these nasty problems, first the physical meaning of the Lagrangian multiplier ϵ_2 needs to be established. It can be inferred from examining the case that two different systems, call them A and B, are in thermal contact. Since the interactions are assumed weak, the eigenfunctions of the combined system are the products of those of the separate systems. That means that the number of eigenfunctions of the combined system $Q_{\vec{I}_A \vec{I}_B}$ is the product of those of the individual systems. Therefore the function to differentiate becomes

$$\begin{aligned} F \;=\; & \ln(Q_{\vec{I}_A} Q_{\vec{I}_B}) \\ & -\epsilon_{1,A}\Big(\sum_{s_A} I_{s_A} - I_A\Big) - \epsilon_{1,B}\Big(\sum_{s_B} I_{s_B} - I_B\Big) \\ & -\epsilon_2\Big(\sum_{s_A} I_{s_A} E_{s_A}^{\rm p} + \sum_{s_B} I_{s_B} E_{s_B}^{\rm p} - E\Big) \end{aligned}$$

Note the constraints: the number of particles in system A must be the correct number I_A of particles in that system, and similar for system B. However, since the systems are in thermal contact, they can exchange energy through the weak interactions and there is no longer a constraint on the energy of the individual systems. Only the combined energy must equal the given total. That means the two systems share the same Lagrangian variable ϵ_2. For the rest, the equations for the two systems are just like if they were not in thermal contact, because the logarithm in F separates, and then the differentiations with respect to the shelf numbers I_{s_A} and I_{s_B} give the same results as before.

It follows that two systems that have the same value of ϵ_2 can be brought into thermal contact and nothing happens, macroscopically. However, if two

systems with different values of ϵ_2 are brought into contact, the systems will adjust, and energy will transfer between them, until the two ϵ_2 values have become equal. That means that ϵ_2 is a temperature variable. From here on, the temperature will be *defined* as $T = 1/\epsilon_2 k_{\rm B}$, so that $\epsilon_2 = 1/k_{\rm B}T$, with $k_{\rm B}$ the Boltzmann constant. The same way, for now the chemical potential μ will simply be defined to be the constant $-\epsilon_1/\epsilon_2$. Chapter 11.14.4 will eventually establish that the temperature defined here is the ideal gas temperature, while derivation {D.61} will establish that μ is the Gibbs free energy per atom that is normally defined as the chemical potential.

Returning now to the nasty problems of the distribution for bosons, first assume that every shelf has at least two states, and that $(E_s^{\rm p} - \mu)/k_{\rm B}T$ is positive even for the ground state. In that case there is no problem with the derived solution. However, Bose-Einstein condensation will occur when either the number density is increased by putting more particles in the system, or the temperature is decreased. Increasing particle density is associated with increasing chemical potential μ because

$$I_s = \frac{N_s - 1}{e^{(E_s^{\rm p}-\mu)/k_{\rm B}T} - 1}$$

implies that every shelf particle number increases when μ increases. Decreasing temperature by itself decreases the number of particles, and to compensate and keep the number of particles the same, μ must then once again increase. When μ gets very close to the ground state energy, the exponential in the expression for the number of particles on the ground state shelf $s = 1$ becomes very close to one, making the total denominator very close to zero, so the number of particles I_1 in the ground state blows up. When it becomes a finite fraction of the total number of particles I even when I is macroscopically large, Bose-Einstein condensation is said to have occurred.

Note that under reasonable assumptions, it will only be the ground state shelf that ever acquires a finite fraction of the particles. For, assume the contrary, that shelf 2 also holds a finite fraction of the particles. Using Taylor series expansion of the exponential for small values of its argument, the shelf occupation numbers are

$$I_1 = \frac{(N_1 - 1)k_{\rm B}T}{E_1^{\rm p} - \mu}$$

$$I_2 = \frac{(N_2 - 1)k_{\rm B}T}{E_1^{\rm p} - \mu + (E_2^{\rm p} - E_1^{\rm p})}$$

$$I_3 = \frac{(N_3 - 1)k_{\rm B}T}{E_1^{\rm p} - \mu + (E_2^{\rm p} - E_1^{\rm p}) + (E_3^{\rm p} - E_2^{\rm p})}$$

$$\vdots$$

For I_2 to also be a finite fraction of the total number of particles, $E^{\rm p}_2 - E^{\rm p}_1$ must be similarly small as $E^{\rm p}_1 - \mu$. But then, reasonably assuming that the energy levels are at least roughly equally spaced, and that the number of states will not decrease with energy, so must I_3 be a finite fraction of the total, and so on. You cannot have a large number of shelves each having a finite fraction of the particles, because there are not so many particles. More precisely, a sum roughly like $\sum_{s=2}^{\infty} \mathrm{const}/s\Delta E$, (or worse), sums to an amount that is much larger than the term for $s = 2$ alone. So if I_2 would be a finite fraction of I, then the sum would be much larger than I.

What happens during condensation is that μ becomes much closer to $E^{\rm p}_1$ than $E^{\rm p}_1$ is to the next energy level $E^{\rm p}_2$, and only the ground state shelf ends up with a finite fraction of the particles. The remainder is spread out so much that the shelf numbers immediately above the ground state only contain a negligible fraction of the particles. It also follows that for all shelves except the ground state one, μ may be approximated as being $E^{\rm p}_1$. (Specific data for particles in a box is given in chapter 11.14.1. The entire story may of course need to be modified in the presence of confinement, compare chapter 6.12.)

The other problem with the analysis of the occupation numbers for bosons is that the number of single-particle states on the shelves had to be at least two. There is no reason why a system of weakly-interacting spinless bosons could not have a unique single-particle ground state. And combining the ground state with the next one on a single shelf is surely not an acceptable approximation in the presence of potential Bose-Einstein condensation. Fortunately, the mathematics still partly works:

$$\frac{\Delta F}{\Delta I_1} = \ln(I_1 + 1) - \ln(I_1 + 1) - \epsilon_1 - \epsilon_2 E^{\rm p}_1 = 0$$

implies that $\epsilon_1 - \epsilon_2 E^{\rm p}_1 = 0$. In other words, μ is equal to the ground state energy $E^{\rm p}_1$ exactly, rather than just extremely closely as above.

That then is the condensed state. Without a chemical potential that can be adjusted, for any given temperature the states above the ground state contain a number of particles that is completely unrelated to the actual number of particles that is present. Whatever is left can be dumped into the ground state, since there is no constraint on I_1.

Condensation stops when the number of particles in the states above the ground state wants to become larger than the actual number of particles present. Now the mathematics changes, because nature says "Wait a minute, there is no such thing as a negative number of particles in the ground state!" Nature now adds the constraint that $I_1 = 0$ rather than negative. That adds another penalty term, $\epsilon_3 I_1$ to F and ϵ_3 takes care of satisfying the equation for the ground state shelf number. It is a sad story, really: below the condensation temperature, the ground state was awash in particles, above it, it has zero. None.

A system of weakly interacting helium atoms, spinless bosons, would have a unique single-particle ground state like this. Since below the condensation temperature, the elevated energy states have no clue about an impending lack of particles actually present, physical properties such as the specific heat stay analytical until condensation ends.

It may be noted that above the condensation temperature it is only the most probable set of the occupation numbers that have exactly zero particles in the unique ground state. The expectation value of the number in the ground state will include neighboring sets of occupation numbers to the most probable one, and the number has nowhere to go but up, compare {D.61}.

D.58 The canonical probability distribution

This note deduces the canonical probability distribution. Since the derivations in typical textbooks seem crazily convoluted and the made assumptions not at all as self-evident as the authors suggest, a more mathematical approach will be followed here.

Consider a big system consisting of many smaller subsystems $A, B, \ldots$ with a given total energy E. Call the combined system the collective. Following the same reasoning as in derivation {D.57} for two systems, the thermodynamically stable equilibrium state has shelf occupation numbers of the subsystems satisfying

$$\frac{\partial \ln Q_{\vec{I}_A}}{\partial I_{s_A}} - \epsilon_{1,A} - \epsilon_2 E^{\rm p}_{s_A} = 0$$

$$\frac{\partial \ln Q_{\vec{I}_B}}{\partial I_{s_B}} - \epsilon_{1,B} - \epsilon_2 E^{\rm p}_{s_B} = 0$$

$$\ldots$$

where ϵ_2 is a shorthand for $1/k_{\rm B}T$.

An individual system, take A as the example, no longer has an individual energy that is for certain. Only the collective has that. That means that when A is taken out of the collective, its shelf occupation numbers will have to be described in terms of probabilities. There will still be an expectation value for the energy of the system, but system energy eigenfunctions $\psi^{\rm S}_{q_A}$ with somewhat different energy $E^{\rm S}_{q_A}$ can no longer be excluded with certainty. However, still assume, following the fundamental assumption of quantum statistics, {N.23}, that the physical differences between the system energy eigenfunctions do not make (enough of) a difference to affect which ones are likely or not. So, the probability P_{q_A} of a system eigenfunction $\psi^{\rm S}_{q_A}$ will be assumed to depend only on its energy $E^{\rm S}_{q_A}$:

$$P_{q_A} = P(E^{\rm S}_{q_A}).$$

where P is some as yet unknown function.

For the isolated example system A, the question is now no longer "What shelf numbers have the most eigenfunctions?" but "What shelf numbers have the highest probability?" Note that all system eigenfunctions $\psi^{\rm S}_{q_A}$ for a given set of shelf numbers $\vec{I}_A$ have the same system energy $E^{\rm S}_{\vec{I}_A} = \sum_{s_A} I_{s_A} E^{\rm p}_{s_A}$. Therefore, the probability of a given set of shelf numbers $P_{\vec{I}_A}$ will be the number of eigenfunctions with those shelf numbers times the probability of each individual eigenfunction:

$$P_{\vec{I}_A} = Q_{\vec{I}_A} P(E^{\rm S}_{\vec{I}_A}).$$

Mathematically, the function whose partial derivatives must be zero to find the most probable shelf numbers is

$$F = \ln\left(P_{\vec{I}_A}\right) - \epsilon_{1,A}\Big(\sum_{s_A} I_{s_A} - I_A\Big).$$

The maximum is now to be found for the shelf number probabilities, not their eigenfunction counts, and there is no longer a constraint on energy.

Substituting $P_{\vec{I}_A} = Q_{\vec{I}_A} P(E^{\rm S}_{\vec{I}_A})$, taking apart the logarithm, and differentiating, produces

$$\frac{\partial \ln Q_{\vec{I}_A}}{\partial I_{s_A}} + \frac{{\rm d}\ln(P)}{{\rm d}E^{\rm S}_{\vec{I}_A}} E^{\rm p}_{s_A} - \epsilon_{1,A} = 0$$

That is exactly like the equation for the shelf numbers of system A when it was part of the collective, except that the derivative of the as yet unknown function $\ln(P_A)$ takes the place of $-\epsilon_2$, i.e. $-1/k_{\rm B}T$. It follows that the two must be the same, because the shelf numbers cannot change when the system A is taken out of the collective it is in thermal equilibrium with. For one, the net energy would change if that happened, and energy is conserved.

It follows that ${\rm d}\ln P/{\rm d}E^{\rm S}_{\vec{I}_A} = -1/k_{\rm B}T$ at least in the vicinity of the most probable energy $E^{\rm S}_{\vec{I}_A}$. Hence in the vicinity of that energy

$$P(E^{\rm S}_A) = \frac{1}{Z_A} e^{-E^{\rm S}_A/k_{\rm B}T}$$

which is the canonical probability. Note that the given derivation only ensures it to be true in the vicinity of the most probable energy. Nothing says it gives the correct probability for, say, the ground state energy. But then the question becomes "What difference does it make?" Suppose the ground state has a probability of 0. followed by only 100 zeros instead of the predicted 200 zeros? What would change in the price of eggs?

Note that the canonical probability is self-consistent: if two systems at the same temperature are combined, the probabilities of the combined eigenfunctions multiply, as in

$$P_{AB} = \frac{1}{Z_A Z_B} e^{-(E^{\rm S}_A + E^{\rm S}_B)/k_{\rm B}T}.$$

That is still the correct expression for the combined system, since its energy is the sum of those of the two separate systems. Also for the partition functions

$$Z_A Z_B = \sum_{q_A} \sum_{q_B} e^{-(E^{\rm S}_{q_A} + E^{\rm S}_{q_B})/k_{\rm B}T} = Z_{AB}.$$

D.59 Analysis of the ideal gas Carnot cycle

Refer to figure D.4 for the physical device to be analyzed. The refrigerant circulating through the device is an ideal gas with constant specific heats, like a thin gas of helium atoms. Chapter 11.14 will examine ideal gases in detail, but for now some reminders from introductory classical physics classes about ideal gasses must do. The internal energy of the gas is $E = mIC_V T$ where mI is its mass and C_v is a constant for a gas like helium whose atoms only have translational kinetic energy. Also, the ideal gas law says that $PV = mIRT$, where P is the pressure, V the volume, and the constant R is the gas constant, equal to the universal gas constant divided by the molar mass.

The differential version of the first law, energy conservation, (11.11), says that

$$\mathrm{d}E = \delta Q - P\,\mathrm{d}V$$

or getting rid of internal energy and pressure using the given expressions,

$$mIC_v\,\mathrm{d}T = \delta Q - mIRT\frac{\mathrm{d}V}{V}.$$

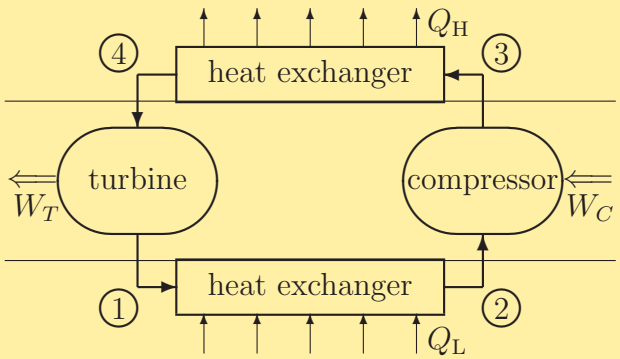


Figure D.4: Schematic of the Carnot refrigeration cycle.

Now for the transitions through the heat exchangers, from 1 to 2 or from 3 to 4 in figure D.4, the temperature is approximated to be constant. The first law above can then be integrated to give the heat added to the substance as:

$$Q_{\rm L} = mIRT_{\rm L}\left(\ln V_2 - \ln V_1\right) \qquad Q_{\rm H} = -mIRT_{\rm H}\left(\ln V_4 - \ln V_3\right).$$

Remember that unlike $Q_{\rm L}$, $Q_{\rm H}$ is taken positive if it comes out of the substance.

On the other hand, for the transitions through the adiabatic turbine and compressor, the heat δQ added is zero. Then the first law can be divided through by T and integrated to give

$$mIC_v\,(\ln T_{\rm H} - \ln T_{\rm L}) = -mIR\,(\ln V_3 - \ln V_2)$$

$$mIC_v\,(\ln T_{\rm L} - \ln T_{\rm H}) = -mIR\,(\ln V_1 - \ln V_4)$$

Adding these two expressions shows that

$$\ln V_3 - \ln V_2 + \ln V_1 - \ln V_4 = 0 \quad \Longrightarrow \quad \ln V_3 - \ln V_4 = \ln V_2 - \ln V_1$$

and plugging that into the expressions for the exchanged heats shows that $Q_{\rm H}/T_{\rm H} = Q_{\rm L}/T_{\rm L}$.

D.60 Checks on the expression for entropy

According to the microscopic definition, the differential of the entropy S should be

$$\mathrm{d}S = -k_{\rm B}\mathrm{d}\left[\sum_q P_q \ln P_q\right]$$

where the sum is over all system energy eigenfunctions $\psi^{\rm S}_q$ and P_q is their probability. The differential can be simplified to

$$\mathrm{d}S = -k_{\rm B}\sum_q \left[\ln P_q + 1\right]\,\mathrm{d}P_q = -k_{\rm B}\sum_q \ln P_q\,\mathrm{d}P_q,$$

the latter equality since the sum of the probabilities is always one, so $\sum_q \mathrm{d}P_q = 0$.

This is to be compared with the macroscopic differential for the entropy. Since the macroscopic expression requires thermal equilibrium, P_q in the microscopic expression above can be equated to the canonical value $e^{-E^{\rm S}_q/k_{\rm B}T}/Z$ where $E^{\rm S}_q$ is the energy of system eigenfunction $\psi^{\rm S}_q$. It simplifies the microscopic differential of the entropy to

$$\mathrm{d}S = -k_{\rm B}\sum_q \left[-\frac{E^{\rm S}_q}{k_{\rm B}T} - \ln Z\right]\,\mathrm{d}P_q = -k_{\rm B}\sum_q \left[-\frac{E^{\rm S}_q}{k_{\rm B}T}\right]\,\mathrm{d}P_q = \frac{1}{T}\sum_q E^{\rm S}_q\,\mathrm{d}P_q, \tag{D.38}$$

the second inequality since Z is a constant in the summation and $\sum_q \mathrm{d}P_q = 0$.

The macroscopic expression for the differential of entropy is given by (11.18),

$$\mathrm{d}S = \frac{\delta Q}{T}.$$

Substituting in the differential first law (11.11),

$$\mathrm{d}S = \frac{1}{T}\,\mathrm{d}E + \frac{1}{T}P\,\mathrm{d}V$$

and plugging into that the definitions of E and P,

$$\mathrm{d}S = \frac{1}{T}\,\mathrm{d}\left[\sum_q P_q E_q^{\mathrm{S}}\right] - \frac{1}{T}\left[\sum_q P_q \frac{\mathrm{d}E_q^{\mathrm{S}}}{\mathrm{d}V}\right]\,\mathrm{d}V$$

and differentiating out the product in the first term, one part drops out versus the second term and what is left is the differential for S according to the microscopic definition (D.38). So, the macroscopic and microscopic definitions agree to within a constant on the entropy. That means that they agree completely, because the macroscopic definition has no clue about the constant.

Now consider the case of a system with zero indeterminacy in energy. According to the fundamental assumption, all the eigenfunctions with the correct energy should have the same probability in thermal equilibrium. From the entropy's point of view, thermal equilibrium should be the stable most messy state, having the maximum entropy. For the two views to agree, the maximum of the microscopic expression for the entropy should occur when all eigenfunctions of the given energy have the same probability. Restricting attention to only the energy eigenfunctions ψ_q^{S} with the correct energy, the maximum entropy occurs when the derivatives of

$$F = -k_{\mathrm{B}} \sum_q P_q \ln P_q - \epsilon \left(\sum_q P_q - 1\right)$$

with respect to the P_q are zero. Note that the constraint that the sum of the probabilities must be one has been added as a penalty term with a Lagrangian multiplier, {D.48}. Taking derivatives produces

$$-k_{\mathrm{B}} \ln(P_q) - k_{\mathrm{B}} - \epsilon = 0$$

showing that, yes, all the P_q have the same value at the maximum entropy. (Note that the minima in entropy, all P_q zero except one, do not show up in the derivation; $P_q \ln P_q$ is zero when $P_q = 0$, but its derivative does not exist there. In fact, the infinite derivative can be used to verify that no maxima exist with any of the P_q equal to zero if you are worried about that.)

If the energy is uncertain, and only the expectation energy is known, the penalized function becomes

$$F = -k_{\mathrm{B}} \sum_q P_q \ln P_q - \epsilon_1 \left(\sum P_q - 1\right) - \epsilon_2 \left(\sum E_q^{\mathrm{S}} P_q - E\right)$$

and the derivatives become

$$-k_{\mathrm{B}}\ln(P_q) - k_{\mathrm{B}} - \epsilon_1 - \epsilon_2 E_q^{\mathrm{S}} = 0$$

which can be solved to show that

$$P_q = C_1 e^{-E_q^{\mathrm{S}}/C_2}$$

with C_1 and C_2 constants. The requirement to conform with the given definition of temperature identifies C_2 as $k_{\mathrm{B}}T$ and the fact that the probabilities must sum to one identifies C_1 as $1/Z$.

For two systems A and B in thermal contact, the probabilities of the combined system energy eigenfunctions are found as the products of the probabilities of those of the individual systems. The maximum of the combined entropy, constrained by the given total energy E, is then found by differentiating

$$\begin{aligned} F &= -k_{\mathrm{B}} \sum_{q_A} \sum_{q_B} P_{q_A} P_{q_B} \ln(P_{q_A} P_{q_B}) \\ &\quad -\epsilon_{1,A}(\sum_{q_A} P_{q_A} - 1) - \epsilon_{1,B}(\sum_{q_B} P_{q_B} - 1) \\ &\quad -\epsilon_2(\sum_{q_A} P_{q_A} E_{q_A}^{\mathrm{S}} + \sum_{q_B} P_{q_B} E_{q_B}^{\mathrm{S}} - E) \end{aligned}$$

F can be simplified by taking apart the logarithm and noting that the probabilities P_{q_A} and P_{q_B} sum to one to give

$$\begin{aligned} F &= -k_{\mathrm{B}} \sum_{q_A} P_{q_A} \ln(P_{q_A}) - k_{\mathrm{B}} \sum_{q_B} P_{q_B} \ln(P_{q_B}) \\ &\quad -\epsilon_{1,A}(\sum_{q_A} P_{q_A} - 1) - \epsilon_{1,B}(\sum_{q_B} P_{q_B} - 1) \\ &\quad -\epsilon_2(\sum_{q_A} P_{q_A} E_{q_A}^{\mathrm{S}} + \sum_{q_B} P_{q_B} E_{q_B}^{\mathrm{S}} - E) \end{aligned}$$

Differentiation now produces

$$-k_{\mathrm{B}}\ln(P_{q_A}) - k_{\mathrm{B}} - \epsilon_{1,A} - \epsilon_2 E_{q_A}^{\mathrm{S}} = 0$$

$$-k_{\mathrm{B}}\ln(P_{q_B}) - k_{\mathrm{B}} - \epsilon_{1,B} - \epsilon_2 E_{q_B}^{\mathrm{S}} = 0$$

which produces $P_{q_A} = C_{1,A} e^{-E_{q_A}^{\mathrm{S}}/C_2}$ and $P_{q_B} = C_{1,B} e^{-E_{q_B}^{\mathrm{S}}/C_2}$ and the common constant C_2 then implies that the two systems have the same temperature.

D.61 Chemical potential in the distributions

The following convoluted derivation of the distribution functions comes fairly straightly from Baierlein [4, pp. 170-]. Let it not deter you from reading the rest of this otherwise very clearly written and engaging little book. Even a nonengineering author should be allowed one mistake.

The derivations of the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein distributions given previously, {D.57} and {D.58}, were based on finding the most numerous or most probable distribution. That implicitly assumes that significant deviations from the most numerous/probable distributions will be so rare that they can be ignored. This note will bypass the need for such an assumption since it will directly derive the actual expectation values of the single-particle state occupation numbers ι. In particular for fermions, the derivation will be solid as a rock.

The mission is to derive the expectation number ι_n of particles in an arbitrary single-particle state $\psi^{\rm p}_n$. This expectation value, as any expectation value, is given by the possible values times their probability:

$$\iota_n = \sum_q i_n P_q$$

where i_n is the number of particles that system energy eigenfunction $\psi^{\rm S}_q$ has in single-particle state $\psi^{\rm p}_n$, and P_q the probability of the eigenfunction. Since thermal equilibrium is assumed, the canonical probability value $e^{-E^{\rm S}_q/k_{\rm B}T}/Z$ can be substituted for P_q. Then, if the energy $E^{\rm S}_q$ is written as the sum of the ones of the single particle states times the number of particles in that state, it gives:

$$\iota_n = \frac{1}{Z} \sum_q i_n e^{-(i_1 E^{\rm p}_1 + i_2 E^{\rm p}_2 + \ldots + i_{n-1} E^{\rm p}_{n-1} + i_n E^{\rm p}_n + i_{n+1} E^{\rm p}_{n+1} + \ldots)/k_{\rm B}T}.$$

Note that i_n is the occupation number of single-particle state $\psi^{\rm p}_n$, just like I_s was the occupation number of shelf s. Dealing with single-particle state occupation numbers has an advantage over dealing with shelf numbers: you do not have to figure out how many system eigenfunctions there are. For a given set of single-particle state occupation numbers $\vec{\imath} = |i_1, i_2, \ldots\rangle$, there is exactly *one* system energy eigenfunction. Compare figures 11.2 and 11.3: if you know how many particles there are in each single-particle state, you know everything there is to know about the eigenfunction depicted. (This does not apply to distinguishable particles, figure 11.1, because for them the numbers on the particles can still vary for given occupation numbers, but as noted in chapter 11.11, there is no such thing as identical distinguishable particles anyway.)

It has the big consequence that the sum over the eigenfunctions can be

replaced by sums over all sets of occupation numbers:

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1} \sum_{i_2} \cdots \sum_{i_{n-1}} \sum_{i_n} \sum_{i_{n+1}} \cdots}_{i_1+i_2+\ldots+i_{n-1}+i_n+i_{n+1}+\ldots=I} i_n e^{-(i_1 E_1^{\mathrm{p}} + i_2 E_2^{\mathrm{p}} + \ldots + i_{n-1} E_{n-1}^{\mathrm{p}} + i_n E_n^{\mathrm{p}} + i_{n+1} E_{n+1}^{\mathrm{p}} + \ldots)/k_{\mathrm{B}}T}$$

Each set of single-particle state occupation numbers corresponds to exactly one eigenfunction, so each eigenfunction is still counted exactly once. Of course, the occupation numbers do have to add up to the correct number of particles in the system.

Now consider first the case of I identical bosons. For them the occupation numbers may have values up to a maximum of I:

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1=0}^{I} \sum_{i_2=0}^{I} \cdots \sum_{i_{n-1}=0}^{I} \sum_{i_n=0}^{I} \sum_{i_{n+1}=0}^{I} \cdots}_{i_1+i_2+\ldots+i_{n-1}+i_n+i_{n+1}+\ldots=I} i_n e^{-(i_1 E_1^{\mathrm{p}} + i_2 E_2^{\mathrm{p}} + \ldots + i_{n-1} E_{n-1}^{\mathrm{p}} + i_n E_n^{\mathrm{p}} + i_{n+1} E_{n+1}^{\mathrm{p}} + \ldots)/k_{\mathrm{B}}T}$$

One simplification that is immediately evident is that all the terms that have $i_n = 0$ are zero and can be ignored. Now apply a trick that only a mathematician would think of: define a new summation index i'_n by setting $i_n = 1 + i'_n$. Then the summation over i'_n can start at 0 and will run up to $I - 1$. Plugging $i_n = 1 + i'_n$ into the sum above gives

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1=0}^{I} \cdots \sum_{i_{n-1}=0}^{I} \sum_{i'_n=0}^{I-1} \sum_{i_{n+1}=0}^{I} \cdots}_{i_1+\ldots+i_{n-1}+i'_n+i_{n+1}+\ldots=I-1} (1 + i'_n) e^{-(i_1 E_1^{\mathrm{p}} + \ldots + i_{n-1} E_{n-1}^{\mathrm{p}} + E_n^{\mathrm{p}} + i'_n E_n^{\mathrm{p}} + i_{n+1} E_{n+1}^{\mathrm{p}} + \ldots)/k_{\mathrm{B}}T}$$

This can be simplified by taking the constant part of the exponential out of the summation. Also, the constraint in the bottom shows that the occupation numbers can no longer be any larger than $I - 1$ (since the original i_n is at least one), so the upper limits can be reduced to $I - 1$. Finally, the prime on i'_n may as well be dropped, since it is just a summation index and it does not make a difference what name you give it. So, altogether,

$$\iota_n = \frac{1}{Z} e^{-E_n^{\mathrm{p}}/k_{\mathrm{B}}T} \underbrace{\sum_{i_1=0}^{I-1} \cdots \sum_{i_{n-1}=0}^{I-1} \sum_{i_n=0}^{I-1} \sum_{i_{n+1}=0}^{I-1} \cdots}_{i_1+\ldots+i_{n-1}+i_n+i_{n+1}+\ldots=I-1} (1 + i_n) e^{-(i_1 E_1^{\mathrm{p}} + \ldots + i_{n-1} E_{n-1}^{\mathrm{p}} + i_n E_n^{\mathrm{p}} + i_{n+1} E_{n+1}^{\mathrm{p}} + \ldots)/k_{\mathrm{B}}T}$$

The right hand side falls apart into two sums: one for the 1 in $1 + i_n$ and one for the i_n in $1 + i_n$. The first sum is essentially the partition function Z^- for a system with $I - 1$ particles instead of I. The second sum is essentially Z^- times the expectation value ι_n^- for such a system. To be precise

$$\iota_n = \frac{1}{Z} e^{-E_n^{\mathrm{p}}/k_{\mathrm{B}}T} Z^- \left[1 + \iota_n^-\right]$$

This equation is exact, no approximations have been made yet.

The system with $I - 1$ particles is the same in all respects to the one for I particles, except that it has one less particle. In particular, the single-particle energy eigenfunctions are the same, which means the volume of the box is the same, and the expression for the canonical probability is the same, meaning that the temperature is the same.

But when the system is macroscopic, the occupation counts for $I-1$ particles must be virtually identical to those for I particles. Clearly the physics should not change noticeably depending on whether 10^{20} or $10^{20} + 1$ particles are present. If $\iota_n^- = \iota_n$, then the above equation can be solved to give:

$$\iota_n = 1 \Big/ \left[\frac{Z}{Z^-} e^{E_n^{\mathrm{p}}/k_{\mathrm{B}}T} - 1\right]$$

The final formula is the Bose-Einstein distribution with

$$e^{-\mu/k_{\mathrm{B}}T} = \frac{Z}{Z^-}$$

Solve for μ:

$$\mu = -k_{\mathrm{B}}T \ln\left(\frac{Z}{Z^-}\right) = \frac{-k_{\mathrm{B}}T \ln(Z) + k_{\mathrm{B}}T \ln(Z^-)}{I - (I-1)}$$

The final fraction is a difference quotient approximation for the derivative of the Helmholtz free energy with respect to the number of particles. Now a single particle change is an extremely small change in the number of particles, so the difference quotient will be to very great accuracy be equal to the derivative of the Helmholtz free energy with respect to the number of particles. And as noted earlier, in the obtained expressions, volume and temperature are held constant. So, $\mu = (\partial F/\partial I)_{T,V}$, and (11.39) identified that as the chemical potential. Do note that μ is on a single-particle basis, while $\bar{\mu}$ was taken to be on a molar basis. The Avogadro number $I_{\mathrm{A}} = 6.022\,1\ 10^{26}$ particles per kmol converts between the two.

Now consider the case of I identical fermions. Then, according to the exclusion principle, there are only two allowed possibilities for the occupation numbers: they can be zero or one:

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1=0}^{1} \cdots \sum_{i_{n-1}=0}^{1} \sum_{i_n=0}^{1} \sum_{i_{n+1}=0}^{1} \cdots}_{i_1+\ldots+i_{n-1}+i_n+i_{n+1}+\ldots=I} i_n e^{-(i_1 E_1^{\mathrm{p}} + \ldots + i_{n-1}E_{n-1}^{\mathrm{p}} + i_n E_n^{\mathrm{p}} + i_{n+1}E_{n+1}^{\mathrm{p}} + \ldots)/k_{\mathrm{B}}T}$$

Again, all terms with $i_n = 0$ are zero, so you can set $i_n = 1 + i'_n$ and get

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1=0}^{1} \cdots \sum_{i_{n-1}=0}^{1} \sum_{i'_n=0}^{0} \sum_{i_{n+1}=0}^{1} \cdots}_{i_1+\ldots+i_{n-1}+i'_n+i_{n+1}+\ldots=I-1}$$
$$(1+i'_n) e^{-(i_1 E_1^{\mathrm{p}} + \ldots + i_{n-1} E_{n-1}^{\mathrm{p}} + E_n^{\mathrm{p}} + i'_n E_n^{\mathrm{p}} + i_{n+1} E_{n+1}^{\mathrm{p}} + \ldots)/k_{\mathrm{B}}T}$$

But now there is a difference: even for a system with $I-1$ particles i'_n can still have the value 1 but the upper limit is zero. Fortunately, since the above sum only sums over the single value $i'_n = 0$, the factor $(1 + i'_n)$ can be replaced by $(1 - i'_n)$ without changing the answer. And then the summation can include $i'_n = 1$ again, because $(1 - i'_n)$ is zero when $i'_n = 1$. This sign change produces the sign change in the Fermi-Dirac distribution compared to the Bose-Einstein one; the rest of the analysis is the same.

Here are some additional remarks about the only approximation made, that the systems with I and $I-1$ particles have the same expectation occupation numbers. For fermions, this approximation is justified to the gills, because it can be easily be seen that the obtained value for the occupation number is *in between* those of the systems with $I-1$ and I particles. Since nobody is going to count whether a macroscopic system has 10^{20} particles or $10^{20} + 1$, this is truly as good as any theoretical prediction can possibly get.

But for bosons, it is a bit trickier because of the possibility of condensation. Assume, reasonably, that when a particle is added, the occupation numbers will not go down. Then the derived expression overestimates both expectation occupation numbers ι_n and ι_n^-. However, it could at most be wrong, (i.e. have a finite relative error) for a finite number of states, and the number of single-particle states will be large. (In the earlier derivation using shelf numbers, the actual ι_n was found to be lower than the Bose-Einstein value by a factor $(N_s - 1)/N_s$ with N_s the number of states on the shelf.)

If the factor $Z e^{E_1^{\mathrm{p}}/k_{\mathrm{B}}T}/Z^-$ is one exactly, which definitely means Bose-Einstein condensation, then $i_1 = 1 + i_1^-$. In that case, the additional particle that the system with I particles has goes with certainty into the ground state. So the ground state better be unique then; the particle cannot go into two ground states.

D.62 Fermi-Dirac integrals at low temperature

This note finds the basic Fermi-Dirac integrals for the free-electron gas at low temperature. To summarize the main text, the number of particles and total energy per unit volume are to be found from

$$\frac{I}{\mathcal{V}} = \int_0^\infty \iota^{\mathrm{f}} \mathcal{D} \, \mathrm{d}E^{\mathrm{p}} \qquad \frac{E}{\mathcal{V}} = \int_0^\infty E^{\mathrm{p}} \iota^{\mathrm{f}} \mathcal{D} \, \mathrm{d}E^{\mathrm{p}}$$

where the Fermi-Dirac distribution and the density of states are:

$$\iota^{\mathrm{f}} = \frac{1}{e^{(E^{\mathrm{p}}-\mu)/k_{\mathrm{B}}T}+1} \qquad \mathcal{D} = \frac{n_s}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\sqrt{E^{\mathrm{p}}}$$

and the number of spin states $n_s = 2s+1 = 2$ for systems of electrons. This may be rewritten in terms of the scaled energies

$$u = \frac{E^{\mathrm{p}}}{k_{\mathrm{B}}T} \qquad u_0 = \frac{\mu}{k_{\mathrm{B}}T}$$

to give

$$\frac{I}{\mathcal{V}} = \frac{n_s}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\mu^{3/2}\int_{u=0}^{\infty}\frac{(u/u_0)^{1/2}}{e^{u-u_0}+1}\,\mathrm{d}(u/u_0)$$

$$\frac{E}{\mathcal{V}} = \frac{n_s}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}\mu^{5/2}\int_{u=0}^{\infty}\frac{(u/u_0)^{3/2}}{e^{u-u_0}+1}\,\mathrm{d}(u/u_0)$$

To find the number of particles per unit volume for small but nonzero temperature, in the final integral change integration variable to $v = (u/u_0) - 1$, then take the integral apart as

$$\int_{-1}^{0}\sqrt{1+v}\,\mathrm{d}v - \int_{-1}^{0}\frac{\sqrt{1+v}e^{u_0 v}\,\mathrm{d}v}{e^{u_0 v}+1} + \int_{0}^{\infty}\frac{\sqrt{1+v}\,\mathrm{d}v}{e^{u_0 v}+1}$$

and clean it up, by dividing top and bottom of the center integral by the exponential and then inverting the sign of v in the integral, to give

$$\int_{-1}^{0}\sqrt{1+v}\,\mathrm{d}v + \int_{0}^{1}\frac{(\sqrt{1+v}-\sqrt{1-v})\,\mathrm{d}v}{e^{u_0 v}+1} + \int_{1}^{\infty}\frac{\sqrt{1+v}\,\mathrm{d}v}{e^{u_0 v}+1}$$

In the second integral, the range that is not killed off by the exponential in the bottom is very small for large u_0 and you can therefore approximate $\sqrt{1+v}-\sqrt{1-v}$ as v, or using a Taylor series if still higher precision is required. (Note that the Taylor series only includes odd terms. That makes the final expansions proceed in powers of $1/u_0^2$.) The range of integration can be extended to infinity, since the exponential in the bottom is exponentially large beyond $v = 1$. For the same reason, the third integral can be ignored completely. Note that $\int_0^\infty x\,\mathrm{d}x/(e^x+1) = \pi^2/12$, see [41, 18.81-82, p. 132] for this and additional integrals.

Finding the number of particles per unit volume $I/\mathcal{V}$ this way and then solving the expression for the Fermi level μ gives

$$\mu = E_{\mathrm{F}}^{\mathrm{p}} - \frac{\pi^2}{12}\left(\frac{k_{\mathrm{B}}T}{E_{\mathrm{F}}^{\mathrm{p}}}\right)^2 E_{\mathrm{F}}^{\mathrm{p}} + \ldots \qquad E_{\mathrm{F}}^{\mathrm{p}} = \left(\frac{6\pi^2}{n_s}\right)^{2/3}\frac{\hbar^2}{2m}\left(\frac{I}{V}\right)^{2/3} \tag{D.39}$$

This used the approximations that $\mu \approx E_{\rm F}^{\rm p}$ and u_0^{-2} is small, so

$$u_0^{-2} = \left(\frac{k_{\rm B}T}{\mu}\right)^2 \approx \left(\frac{k_{\rm B}T}{E_{\rm F}^{\rm p}}\right)^2 \qquad \left(1+\frac{\pi^2}{8}u_0^{-2}\right)^{-2/3} \approx 1 - \frac{2}{3}\frac{\pi^2}{8}u_0^{-2}$$

The integral in the expression for the total energy per unit volume goes exactly the same way. That gives the average energy per particle as

$$\frac{E}{I} = E_{\rm ave}^{\rm p} = \frac{3}{5}E_{\rm F}^{\rm p} + \frac{\pi^2}{4}\left(\frac{k_{\rm B}T}{E_{\rm F}^{\rm p}}\right)^2 E_{\rm F}^{\rm p} + \ldots \tag{D.40}$$

To get the specific heat at constant volume, divide by m and differentiate with respect to temperature:

$$C_v = \frac{\pi^2}{2}\frac{k_{\rm B}T}{E_{\rm F}^{\rm p}}\frac{k_{\rm B}}{m} + \ldots$$

D.63 Angular momentum uncertainty

Suppose that an eigenstate, call it $|m\rangle$, of $\widehat{J}_z$ is also an eigenstate of $\widehat{J}_x$. Then $[\widehat{J}_z, \widehat{J}_x]|m\rangle$ must be zero, and the commutator relations say that this is equivalent to $\widehat{J}_y|m\rangle = 0$, which makes $|m\rangle$ also an eigenvector of $\widehat{J}_y$, and with the eigenvalue zero to boot. So the angular momentum in the y-direction must be zero. Repeating the same argument using the $[\widehat{J}_x, \widehat{J}_y]$ and $[\widehat{J}_y, \widehat{J}_z]$ commutator pairs shows that the angular momentum in the other two directions is zero too. So there is no angular momentum at all, $|m\rangle$ is an $|0\,0\rangle$ state.

D.64 Spherical harmonics by ladder operators

One application of ladder operators is to find the spherical harmonics, which as noted in chapter 4.2.3 is not an easy problem. To do it with ladder operators, show that

$$\boxed{\widehat{L}_x = \frac{\hbar}{\mathrm{i}}\left(-\sin\phi\frac{\partial}{\partial\theta} - \frac{\cos\theta\cos\phi}{\sin\theta}\frac{\partial}{\partial\phi}\right) \quad \widehat{L}_y = \frac{\hbar}{\mathrm{i}}\left(\cos\phi\frac{\partial}{\partial\theta} - \frac{\cos\theta\sin\phi}{\sin\theta}\frac{\partial}{\partial\phi}\right)} \tag{D.41}$$

then that

$$\boxed{L^+ = \hbar e^{\mathrm{i}\phi}\left(\frac{\partial}{\partial\theta} + \mathrm{i}\frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\phi}\right) \qquad L^- = \hbar e^{-\mathrm{i}\phi}\left(-\frac{\partial}{\partial\theta} + \mathrm{i}\frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\phi}\right)} \tag{D.42}$$

Note that the spherical harmonics are of the form $Y_l^m = e^{\mathrm{i}m\phi}\Theta_l^m(\theta)$, so

$$L^+ Y_l^m = \hbar e^{\mathrm{i}(m+1)\phi}\sin^m\theta\frac{\mathrm{d}(\Theta_l^m/\sin^m\theta)}{\mathrm{d}\theta}$$

$$L^- Y_l^m = -\hbar e^{\mathrm{i}(m-1)\phi} \frac{1}{\sin^m\theta} \frac{\mathrm{d}(\Theta_l^m \sin^m\theta)}{\mathrm{d}\theta}$$

Find the Y_l^l harmonic from $\widehat{L}^+ Y_l^l = 0$. That gives

$$\boxed{Y_l^l = \sqrt{\frac{1}{4\pi}\frac{(2l+1)!}{(2^l\, l!)^2}}\, e^{\mathrm{i}l\phi} \sin^l\theta = \sqrt{\frac{1}{4\pi}\frac{1\,3\,5\,\ldots\,(2l+1)}{2\,4\,6\,\ldots\,2l}}(x+\mathrm{i}y)^l} \tag{D.43}$$

Now apply $\widehat{L}^-$ to find the rest of the ladder.

Interestingly enough, the solution of the one-dimensional harmonic oscillator problem can also be found using ladder operators. It turns out that, in the notation of that problem,

$$H^+ = -\mathrm{i}\widehat{p} + m\omega\widehat{x} \quad H^- = \mathrm{i}\widehat{p} + m\omega\widehat{x}$$

are commutator eigenoperators of the harmonic oscillator Hamiltonian, with eigenvalues $\pm\hbar\omega$. So, you can play the same games of constructing ladders. Easier, really, since there is no equivalent to square angular momentum to worry about in that problem: there is only one ladder. See [25, pp. 42-47] for details. An equivalent derivation is given in addendum {A.15.5} based on quantum field theory.

D.65 How to make Clebsch-Gordan tables

The procedure of finding the Clebsch-Gordan coefficients for the combination of any two spin ladders is exactly the same as for electron ones, so it is simple enough to program.

To further simplify things, it turns out that the coefficients are all square roots of rational numbers (i.e. ratios of integers such as 102/38.) The step-up and step-down operators by themselves produce square roots of rational numbers, so at first glance it would appear that the individual Clebsch-Gordan coefficients would be sums of square roots. But the square roots of a given coefficient are all compatible and can be summed into one. To see why, consider the coefficients that result from applying the combined step down ladder $\widehat{J}_{ab}^-$ a few times on the top of the ladder $|j\ j\rangle_a |j\ j\rangle_b$. Every contribution to the coefficient of a state $|j\ m\rangle_a |j\ m\rangle_b$ comes from applying $\widehat{J}_a^-$ for $j_a - m_a$ times and $\widehat{J}_b^-$ for $j_b - m_b$ times, so all contributions have compatible square roots. $\widehat{J}_{ab}^-$ merely adds an m_{ab} dependent normalization factor.

You might think this pattern would be broken when you start defining the tops of lower ladders, since that process uses the step up operators. But because $\widehat{J}^+\widehat{J}^-$ and $\widehat{J}^-\widehat{J}^+$ are rational numbers (not square roots), applying the up operators is within a rational number the same as applying the down ones, and the pattern turns out to remain.

Additional note: There is also a direct expression for the Clebsch-Gordan coefficients:

$$\langle j\,m|j_1\,m_1\rangle|j_2\,m_2\rangle =$$

$$\delta_{m_1+m_2\,m}$$

$$\sqrt{(2j+1)(j_1+j_2-j)!(j_1-j_2+j)!(-j_1+j_2+j)!/(j_1+j_2+j+1)!}$$

$$\sqrt{(j_1+m_1)!(j_1-m_1)!(j_2+m_2)!(j_2-m_2)!(j+m)!(j-m)!}$$

$$\sum_{z=z_{\rm l}}^{z_{\rm h}} \frac{(-1)^z}{\prod_{i=1}^{3}(z-z_{{\rm l}i})!(z_{{\rm h}i}-z)!}$$

where δ is the Kronecker delta and

$$z_{\rm l1}=0 \quad z_{\rm l2}=j_1+m_2-j \quad z_{\rm l3}=j_2-m_1-j \quad z_{\rm l}=\min(z_{\rm l1},z_{\rm l2},z_{\rm l3})$$

$$z_{\rm h1}=j_1+j_2-j \quad z_{\rm h2}=j_1-m_1 \quad z_{\rm h2}=j_2+m_2 \quad z_{\rm h}=\min(z_{\rm h1},z_{\rm h2},z_{\rm h3})$$

Carefully coded, this one seems to be numerically superior at larger angular momenta. Either way, these coefficients will overflow pretty quickly.

There are also resources on the web to compute these coefficients. See {N.13} for additional information.

D.66 The triangle inequality

The normal triangle inequality continues to apply for expectation values in quantum mechanics.

The way to show that is, like other triangle inequality proofs, rather curious: examine the combination of $\widehat{\vec{J}}_a$, not with $\widehat{\vec{J}}_b$, but with an arbitrary multiple λ of $\widehat{\vec{J}}_b$:

$$\left\langle \left(\vec{J}_a+\lambda\vec{J}_b\right)^2 \right\rangle = \left\langle (J_{x,a}+\lambda J_{x,b})^2 \right\rangle + \left\langle (J_{y,a}+\lambda J_{y,b})^2 \right\rangle + \left\langle (J_{z,a}+\lambda J_{z,b})^2 \right\rangle$$

For $\lambda=1$ this produces the expectation value of $\left(\vec{J}_a+\vec{J}_b\right)^2$, for $\lambda=-1$, the one for $\left(\vec{J}_a-\vec{J}_b\right)^2$. In addition, it is positive for all values of λ, since it consists of expectation values of square Hermitian operators. (Just examine each term in terms of its own eigenstates.)

If you multiply out, you get

$$\left\langle \left(\vec{J}_a+\lambda\vec{J}_b\right)^2 \right\rangle = J_a^2+2M\lambda+J_b^2\lambda^2$$

where $J_a \equiv \sqrt{\left\langle J_{xa}^2 + J_{ya}^2 + J_{za}^2 \right\rangle}$, $J_b \equiv \sqrt{\left\langle J_{xb}^2 + J_{yb}^2 + J_{zb}^2 \right\rangle}$, and M represents mixed terms that do not need to be written out. In order for this quadratic form in λ to always be positive, the discriminant must be negative:

$$M^2 - J_a^2 J_b^2 \leqslant 0$$

which means, taking square roots,

$$-J_a J_b \leqslant M \leqslant J_a J_b$$

and so

$$J_a^2 - 2J_a J_b + J_b^2 \leqslant \left\langle \left(\vec{J}_a + \vec{J}_b \right)^2 \right\rangle \leqslant J_a^2 + 2J_a J_b + J_b^2$$

or

$$|J_a - J_b|^2 \leqslant \left\langle \left(\vec{J}_a + \vec{J}_b \right) \right\rangle^2 \leqslant |J_a + J_b|^2$$

and taking square roots gives the triangle inequality.

Note that this derivation does not use any properties specific to angular momentum and does not require the simultaneous existence of the components. With a bit of messing around, the azimuthal quantum number relation $|j_a - j_b| \leqslant j_{ab} \leqslant j_a + j_b$ can be derived from it if a unique value for j_{ab} exists; the key is to recognize that $J = j + \delta$ where δ is an increasing function of j that stays below $\frac{1}{2}$, and the j values must be half integers. This derivation is not as elegant as using the ladder operators, but the result is the same.

D.67 Momentum of shells

Table 12.1 was originally taken from [36], who in turn took it from the book of Mayer and Jensen. However, the final table contains three typos, as can be seen from the fact that in three cases the numbers of states do not add up to the correct total. (The errors are: for 3 particles with spin 9/2, the 13/2 combined state is omitted, for 4 particles with spin 9/2, the spin 8 state should be double, and for 4 particles with spin 11/2, a spin 7 (double) state is missing. Similarly, [5, p. 140] has the same missing 13/2 combined state, and in addition for 3 particles with spin 7/2, there is a 1/2 state that should not be there.)

So table 12.1 was instead computer-generated, and should therefore be free of typos. Since the program had to be written anyway, some more values were generated and are in table D.1.

Deducing the table using Clebsch-Gordan coefficients would be a messy exercise indeed. A simpler procedure, [31], will here be illustrated for the example that the number of fermions is $I = 3$ and the angular momentum of the single-particle states is $j^{\mathrm{p}} = \frac{5}{2}$. Then the possibilities for the single-particle angular momentum in the z-direction are $m^{\mathrm{p}} = \frac{5}{2}$, $\frac{3}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$, $-\frac{3}{2}$, and $-\frac{5}{2}$. So there

j^p	I	possible combined angular momentum j																	
		$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$	$\frac{9}{2}$	$\frac{11}{2}$	$\frac{13}{2}$	$\frac{15}{2}$	$\frac{17}{2}$	$\frac{19}{2}$	$\frac{21}{2}$	$\frac{23}{2}$	$\frac{25}{2}$	$\frac{27}{2}$	$\frac{29}{2}$	$\frac{31}{2}$	$\frac{33}{2}$	$\frac{35}{2}$
		$\frac{37}{2}$	$\frac{39}{2}$	$\frac{41}{2}$	$\frac{43}{2}$	$\frac{45}{2}$	$\frac{47}{2}$	$\frac{49}{2}$	$\frac{51}{2}$	$\frac{53}{2}$	$\frac{55}{2}$	$\frac{57}{2}$	$\frac{59}{2}$	$\frac{61}{2}$	$\frac{63}{2}$				
$\frac{13}{2}$	1							1											
	3		1	1	1	2	2	2	2	2	1	2	1	1	1	1		1	
	5	1	3	5	5	7	7	8	8	8	7	8	6	6	5	5	3	3	2
		2	1	1		1													
	7	3	4	7	9	10	11	13	12	13	12	12	10	11	8	8	6	5	4
		4	2	2	1	1		1											
$\frac{15}{2}$	1								1										
	3		1	1	1	2	2	2	3	2	2	2	2	1	2	1	1	1	1
			1																
	5	2	4	6	8	9	11	11	13	12	13	12	12	11	11	9	9	7	7
		5	5	3	3	2	2	1	1		1								
	7	4	10	13	17	21	24	25	29	28	29	29	29	26	27	23	22	19	18
		14	14	10	9	7	6	4	4	2	2	1	1		1				

j^p	I	possible combined angular momentum j																	
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
		19	20	21	22	23	24	25	26	27	28	29	30	31	32					
$\frac{13}{2}$	2	1		1		1		1		1		1		1						
	4	2		4	1	5	3	5	3	6	3	5	3	4	2	3	1	2	1	1
			1																	
	6	4	1	7	5	11	7	13	9	13	10	12	8	11	7	8	5	6	3	4
		2	2	1	1		1													
$\frac{15}{2}$	2	1		1		1		1		1		1		1		1				
	4	3		4	2	6	3	7	4	7	5	7	4	7	4	5	3	4	2	3
		1	2	1	1		1													
	6	6	2	11	9	17	13	22	17	23	19	24	18	23	17	19	15	16	11	13
		8	9	6	6	3	4	2	2	1	1		1							
	8	7	4	16	13	25	21	31	26	35	29	35	29	34	27	30	23	25	19	20
		14	15	10	10	6	7	4	4	2	2	1	1		1					

Table D.1: Additional combined angular momentum values.

are 6 different one particle states, and these will give rise to 6!/3!(6-3)! = 20 different antisymmetric states for 3 particles, chapter 5.7.

The combination states can be chosen to have definite values of the combined angular momentum j and momentum in the z-direction m. In the absence of any antisymmetrization requirements, that can be seen from the way that states combine using Clebsch-Gordan coefficients. And these states of definite combined angular momentum must either be antisymmetric and allowable, or symmetric and not allowed. The reason is that exchanging fermions does not do anything physically, since the fermions are identical. So the angular momentum and particle exchange operators commute. Therefore, the eigenstates of the angular momentum operators can also be taken to be eigenstates of the particle exchange operators, which means either symmetric (eigenvalue 1) or antisymmetric (eigenvalue -1).

Let m be the total magnetic quantum number of the 3 fermions in any combination of $j^{\rm p} = \frac{5}{2}$ single-particle states. First note that m is the sum of the three $m^{\rm p}$ values of the individual particles. Next, the highest that $m^{\rm p}$ can be is $\frac{5}{2}$, but the fermions cannot all three be in the same $m^{\rm p} = \frac{5}{2}$ state, only one can. Three fermions need three different states, so the highest the combined m can be is $\frac{5}{2} + \frac{3}{2} + \frac{1}{2}$. This triplet of values of $m^{\rm p}$ gives exactly one antisymmetric combination of states with $m = \frac{9}{2}$. (There is only one Slater determinant for three different given states, chapter 5.7). Since the combined angular momentum of this state in any arbitrary direction can never be observed to be more than $\frac{9}{2}$, because that would violate the above argument in a rotated coordinate system, it must be a $j = \frac{9}{2}$ state. The first conclusion is therefore that the angular momenta cannot combine into a total greater than $j = \frac{9}{2}$. And since j cannot be less than m, there must be states with $j = \frac{9}{2}$.

But note that if $j = m = \frac{9}{2}$ is a valid combination of single-particle states, then so should be the states with $j = \frac{9}{2}$ for the other values of m; these can be thought of as fully equivalent states simply oriented under a different angle. That means that there are a total of 10 combination states with $j = \frac{9}{2}$, in which m is any one of $\frac{9}{2}, \frac{7}{2}, \ldots, -\frac{9}{2}$.

Next consider what combinations have $m = \frac{7}{2}$. The only combination of three different $m^{\rm p}$ values that adds up to $\frac{7}{2}$ is $\frac{5}{2} + \frac{3}{2} - \frac{1}{2}$. So there is only one combined state with $m = \frac{7}{2}$. Since it was already inferred above that there must be one such state with $j = \frac{9}{2}$, that must be the only one. So apparently there is no state with $j = \frac{7}{2}$: such a state would show up as a second $m = \frac{7}{2}$ state under the right orientation.

There are two independent possibilities to create a triplet of different states with $m = \frac{5}{2}$: $\frac{5}{2} + \frac{3}{2} - \frac{3}{2}$ or $\frac{5}{2} + \frac{1}{2} - \frac{1}{2}$. One combination of such a type is already identified as being a $j = \frac{9}{2}$ state, so the second must correspond to a $j = \frac{5}{2}$ state. Since the orientation should again not make a difference, there must be a total of 6 such states, one for each of the different values of m in the range from $\frac{5}{2}$ to $-\frac{5}{2}$.

There are three ways to create a triplet of states with $m = \frac{3}{2}$: $\frac{5}{2} + \frac{3}{2} - \frac{5}{2}$, $\frac{5}{2} + \frac{1}{2} - \frac{3}{2}$, and $\frac{3}{2} + \frac{1}{2} - \frac{1}{2}$. Two of these are already identified as being $j = \frac{9}{2}$ and $j = \frac{5}{2}$, so there must be one set of 4 states with $j = \frac{3}{2}$.

That makes a total of 20 states, so there must not be any states with $j = \frac{1}{2}$. Indeed, there are only three ways to produce $m = \frac{1}{2}$: $\frac{5}{2} + \frac{1}{2} - \frac{5}{2}$, $\frac{5}{2} - \frac{1}{2} - \frac{3}{2}$, and $\frac{3}{2} + \frac{1}{2} - \frac{3}{2}$, and each of these three states is already assigned to a value of j.

It is tricky, but it works. And it is easily put on a computer.

For bosons, the idea is the same, except that states with equal values of m^{p} can no longer be excluded.

D.68 Awkward questions about spin

Now of course you ask: how do you know how the mathematical expressions for spin states change when the coordinate system is rotated around some axis? Darn.

If you did a basic course in linear algebra, they will have told you how the components of normal vectors change when the coordinate system is rotated, but not spin vectors, or spinors, which are two-dimensional vectors in three-dimensional space.

You need to go back to the fundamental meaning of angular momentum. The effect of rotations of the coordinate system around the z-axis was discussed in addendum {A.19}. The expressions given there can be straightforwardly generalized to rotations around a line in the direction of an arbitrary unit vector (n_x, n_y, n_z). Rotation by an angle φ multiplies the n-direction angular momentum eigenstates by $e^{\mathrm{i}m\varphi}$ if $m\hbar$ is the angular momentum in the n-direction. For electron spin, the values for m are $\pm\frac{1}{2}$, so, using the Euler formula (2.5) for the exponential, the eigenstates change by a factor

$$\cos\left(\tfrac{1}{2}\varphi\right) \pm \mathrm{i}\sin\left(\tfrac{1}{2}\varphi\right)$$

For arbitrary combinations of the eigenstates, the first of the two terms above still represents multiplication by the number $\cos\left(\frac{1}{2}\varphi\right)$.

The second term may be compared to the effect of the n-direction angular momentum operator $\widehat{J}_n$, which multiplies the angular momentum eigenstates by $\pm\frac{1}{2}\hbar$; it is seen to be $2\mathrm{i}\sin\left(\frac{1}{2}\varphi\right)\widehat{J}_n/\hbar$. So the operator that describes rotation of the coordinate system over an angle φ around the n-axis is

$$\boxed{\mathcal{R}_{n,\varphi} = \cos\left(\tfrac{1}{2}\varphi\right) + \mathrm{i}\sin\left(\tfrac{1}{2}\varphi\right)\frac{2}{\hbar}\widehat{J}_n} \tag{D.44}$$

Further, in terms of the x, y, and z angular momentum operators, the angular momentum in the n-direction is

$$\widehat{J}_n = n_x\widehat{J}_x + n_y\widehat{J}_y + n_z\widehat{J}_z$$

If you put it in terms of the Pauli spin matrices, $\hbar$ drops out:

$$\mathcal{R}_{n,\varphi} = \cos\left(\tfrac{1}{2}\varphi\right) + \mathrm{i}\sin\left(\tfrac{1}{2}\varphi\right)(n_x\sigma_x + n_y\sigma_y + n_z\sigma_z)$$

Using this operator, you can find out how the spin-up and spin-down states are described in terms of correspondingly defined basis states along the x- or y-axis, and then deduce these correspondingly defined basis states in terms of the z ones.

Note however that the very idea of defining the positive x and y angular momentum states from the z ones by rotating the coordinate system over 90° is somewhat specious. If you rotate the coordinate system over 450° instead, you get a different answer! Off by a factor -1, to be precise. But that is as bad as the indeterminacy gets; whatever way you rotate the axis system to the new position, the basis vectors you get will either be the same or only a factor -1 different {D.69}.

More awkwardly, the negative momentum states obtained by rotation do not lead to real positive numerical factors for the corresponding ladder operators. Presumably, this reflects the fact that at the wave function level, nature does not have the rotational symmetry that it has for observable quantities. Anyway, if nature does not bother to obey such symmetry, then there seems no point in pretending it does. Especially since the nonpositive ladder factors would mess up various formulae. The negative spin states found by rotation go out of the window. Bye, bye.

D.69 More awkwardness about spin

How about that? A note on a note.

The previous note brought up the question: why can you only change the spin states you find in a given direction by a factor -1 by rotating your point of view? Why not by i, say?

With a bit of knowledge of linear algebra and some thought, you can see that this question is really: how can you change the spin states if you perform an arbitrary number of coordinate system rotations that end up in the same orientation as they started?

One way to answer this is to show that the effect of any two rotations of the coordinate system can be achieved by a single rotation over a suitably chosen net angle around a suitably chosen net axis. (Mathematicians call this showing the "group" nature of the rotations.) Applied repeatedly, any set of rotations of the starting axis system back to where it was becomes a single rotation around a single axis, and then it is easy to check that at most a change of sign is possible.

(To show that any two rotations are equivalent to one, just crunch out the multiplication of two rotations, which shows that it takes the algebraic form of a single rotation, though with a unit vector $\vec{n}$ not immediately evident to be of

length one. By noting that the determinant of the rotation matrix must be one, it follows that the length is in fact one.)

D.70 Emergence of spin from relativity

This note will give a (relatively) simple derivation of the Dirac equation to show how relativity naturally gives rise to spin. The equation will be derived without ever mentioning the word spin while doing it, just to prove it can be done. Only Dirac's assumption that Einstein's square root disappears,

$$\sqrt{(mc^2)^2 + \sum_{i=1}^{3} (\widehat{p}_i c)^2} = \alpha_0 mc^2 + \sum_{i=1}^{3} \alpha_i \widehat{p}_i c,$$

will be used and a few other assumptions that have nothing to do with spin.

The conditions on the coefficient matrices α_i for the linear combination to equal the square root can be found by squaring both sides in the equation above and then comparing sides. They turn out to be:

$$\alpha_i^2 = 1 \text{ for every } i \qquad \alpha_i\alpha_j + \alpha_j\alpha_i = 0 \text{ for } i \neq j \tag{D.45}$$

Now assume that the matrices α_i are Hermitian, as appropriate for measurable energies, and choose to describe the wave function vector in terms of the eigenvectors of matrix α_0. Under those conditions α_0 will be a diagonal matrix, and its diagonal elements must be ± 1 for its square to be the unit matrix. So, choosing the order of the eigenvectors suitably,

$$\alpha_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

where the sizes of the positive and negative unit matrices in α_0 are still undecided; one of the two could in principle be of zero size.

However, since $\alpha_0\alpha_i + \alpha_i\alpha_0$ must be zero for the three other Hermitian α_i matrices, it is seen from multiplying that out that they must be of the form

$$\alpha_1 = \begin{pmatrix} 0 & \sigma_1^\dagger \\ \sigma_1 & 0 \end{pmatrix} \quad \alpha_2 = \begin{pmatrix} 0 & \sigma_2^\dagger \\ \sigma_2 & 0 \end{pmatrix} \quad \alpha_3 = \begin{pmatrix} 0 & \sigma_3^\dagger \\ \sigma_3 & 0 \end{pmatrix}.$$

The σ_i matrices, whatever they are, must be square in size or the α_i matrices would be singular and could not square to one. This then implies that the positive and negative unit matrices in α_0 must be the same size.

Now try to satisfy the remaining conditions on α_1, α_2, and α_3 using just complex numbers, rather than matrices, for the σ_i. By multiplying out the conditions (D.45), you see that

$$\alpha_i\alpha_i = 1 \Longrightarrow \sigma_i^\dagger\sigma_i = \sigma_i\sigma_i^\dagger = 1$$

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 0 \Longrightarrow \sigma_i^\dagger\sigma_j + \sigma_j^\dagger\sigma_i = \sigma_i\sigma_j^\dagger + \sigma_j\sigma_i^\dagger = 0.$$

The first condition above would require each σ_i to be a number of magnitude one, in other words, a number that can be written as $e^{\mathrm{i}\phi_i}$ for some real angle ϕ_i. The second condition is then according to the Euler formula (2.5) equivalent to the requirement that

$$\cos(\phi_i - \phi_j) = 0 \text{ for } i \neq j;$$

this implies that all three angles would have to be 90 degrees apart. That is impossible: if ϕ_2 and ϕ_3 are each 90 degrees apart from ϕ_1, then ϕ_2 and ϕ_3 are either the same or apart by 180 degrees; not by 90 degrees.

It follows that the components σ_i cannot be numbers, and must be matrices too. Assume, reasonably, that they correspond to some measurable quantity and are Hermitian. In that case the conditions above on the σ_i are the same as those on the α_i, with one critical difference: there are only three σ_i matrices, not four. And so the analysis repeats.

Choose to describe the wave function in terms of the eigenvectors of the σ_3 matrix; this does not conflict with the earlier choice since all half wave function vectors are eigenvectors of the positive and negative unit matrices in α_0. So you have

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and the other two matrices must then be of the form

$$\sigma_1 = \begin{pmatrix} 0 & \tau_1^\dagger \\ \tau_1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & \tau_2^\dagger \\ \tau_2 & 0 \end{pmatrix}$$

But now the components τ_1 and τ_2 can indeed be just complex numbers, since there are only two, and two angles can be apart by 90 degrees. You can take $\tau_1 = e^{\mathrm{i}\phi_1}$ and then $\tau_2 = e^{\mathrm{i}(\phi_1+\pi/2)}$ or $e^{\mathrm{i}(\phi_1-\pi/2)}$. The existence of two possibilities for τ_2 implies that on the wave function level, nature is not mirror symmetric; momentum in the positive y-direction interacts differently with the x and z momenta than in the opposite direction. Since the observable effects are mirror symmetric, do not worry about it and just take the first possibility.

So, the goal of finding a formulation in which Einstein's square root falls apart has been achieved. However, you can clean up some more, by redefining the value of τ_1 away. If the four-dimensional wave function vector takes the form (a_1, a_2, a_3, a_4), define $\bar{a}_1 = e^{\mathrm{i}\phi_1/2}a_1$, $\bar{a}_2 = e^{-\mathrm{i}\phi_1/2}a_2$ and similar for $\bar{a}_3$ and $\bar{a}_4$.

In that case, the final cleaned-up σ matrices are

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \tag{D.46}$$

The "s" word has not been mentioned even once in this derivation. So, now please express audible surprise that the σ_i matrices turn out to be the Pauli (it can now be said) spin matrices of chapter 12.10.

But there is more. Suppose you define a new coordinate system rotated 90 degrees around the z-axis. This turns the old y-axis into a new x-axis. Since τ_2 has an additional factor $e^{\mathrm{i}\pi/2}$, to get the normalized coefficients, you must include an additional factor $e^{\mathrm{i}\pi/4}$ in $\bar{a}_1$, which by the fundamental definition of angular momentum discussed in addendum {A.19} means that it describes a state with angular momentum $\frac{1}{2}\hbar$. Similarly a_3 corresponds to a state with angular momentum $\frac{1}{2}\hbar$ and a_2 and a_4 to ones with $-\frac{1}{2}\hbar$.

For nonzero momentum, the relativistic evolution of spin and momentum becomes coupled. But still, if you look at the eigenstates of positive energy, they take the form:

$$\begin{pmatrix} \vec{a} \\ \varepsilon(\vec{p}\cdot\vec{\sigma})\vec{a} \end{pmatrix}$$

where ε is a small number in the nonrelativistic limit and $\vec{a}$ is the two-component vector (a_1, a_2). The operator corresponding to rotation of the coordinate system around the momentum vector commutes with $\vec{p}\cdot\vec{\sigma}$, hence the entire four-dimensional vector transforms as a combination of a spin $\frac{1}{2}\hbar$ state and a spin $-\frac{1}{2}\hbar$ state for rotation around the momentum vector.

D.71 Electromagnetic commutators

The purpose of this note is to identify the two commutators of chapter 13.1; the one that produces the velocity (or rather, the rate of change in expectation position), and the one that produces the force (or rather the rate of change in expectation linear momentum). All basic properties of commutators used in the derivations below are described in chapter 4.5.4.

The Hamiltonian is

$$H = \frac{1}{2m}\left(\widehat{\vec{p}} - q\vec{A}\right)\cdot\left(\widehat{\vec{p}} - q\vec{A}\right) + q\varphi = \frac{1}{2m}\sum_{j=1}^{3}(\widehat{p}_j - qA_j)^2 + q\varphi$$

when the dot product is written out in index notation.

The rate of change in the expectation value of a position vector component r_i is according to chapter 7.2 given by

$$\frac{\mathrm{d}\langle r_i\rangle}{\mathrm{d}t} = \left\langle \frac{\mathrm{i}}{\hbar}[H, r_i] \right\rangle$$

so you need the commutator

$$[H, r_i] = \left[\frac{1}{2m}\sum_{j=1}^{3}(\widehat{p}_j - qA_j)^2 + q\varphi, r_i\right]$$

Now the term $q\varphi$ can be dropped, since functions of position commute with each other. On the remainder, use the fact that each of the two factors $\widehat{p}_j - qA_j$ comes out at its own side of the commutator, to give

$$[H, r_i] = \frac{1}{2m} \sum_{j=1}^{3} \Big\{ (\widehat{p}_j - qA_j)[\widehat{p}_j - qA_j, r_i] + [\widehat{p}_j - qA_j, r_i](\widehat{p}_j - qA_j) \Big\}$$

and then again, since the vector potential is just a function of position too, the qA_j can be dropped from the commutators. What is left is zero unless j is the same as i, since different components of position and momentum commute, and when $j = i$, it is minus the canonical commutator, (minus since the order of r_i and $\widehat{p}_i$ is inverted), and the canonical commutator has value $\mathrm{i}\hbar$, so

$$[H, r_i] = -\frac{1}{m} \mathrm{i}\hbar (\widehat{p}_i - qA_i)$$

Plugging this in the time derivative of the expectation value of position, you get

$$\frac{\mathrm{d}\langle r_i \rangle}{\mathrm{d}t} = \frac{1}{m} \langle \widehat{p}_i - qA_i \rangle$$

so the normal momentum mv_i is indeed given by the operator $\widehat{p}_i - qA_i$.

On to the other commutator! The i-th component of Newton's second law in expectation form,

$$m \frac{\mathrm{d}\langle v_i \rangle}{\mathrm{d}t} = \left\langle \frac{\mathrm{i}}{\hbar} [H, \widehat{p}_i - qA_i] \right\rangle - q \left\langle \frac{\partial A_i}{\partial t} \right\rangle$$

requires the commutator

$$[H, \widehat{p}_i - qA_i] = \left[\frac{1}{2m} \sum_{j=1}^{3} (\widehat{p}_j - qA_j)^2 + q\varphi, p_i - qA_i \right]$$

The easiest is the term $q\varphi$, since both φ and A_i are functions of position and commute. And the commutator with $\widehat{p}_i$ is the generalized fundamental operator of chapter 4.5.4,

$$[q\varphi, p_i] = \mathrm{i}\hbar q \frac{\partial \varphi}{\partial r_i}$$

and plugging that into Newton's equation, you can verify that the electric field term of the Lorentz law has already been obtained.

In what is left of the desired commutator, again take each factor $\widehat{p}_j - qA_j$ to its own side of the commutator:

$$\frac{1}{2m} \sum_{j=1}^{3} \Big\{ (\widehat{p}_j - qA_j)[\widehat{p}_j - qA_j, p_i - qA_i] + [\widehat{p}_j - qA_j, p_i - qA_i](\widehat{p}_j - qA_j) \Big\}$$

Work out the simpler commutator appearing here first:

$$[\widehat{p}_j - qA_j, p_i - qA_i] = -q[p_j, A_i] - q[A_j, p_i] = \mathrm{i}\hbar q \frac{\partial A_i}{\partial r_j} - \mathrm{i}\hbar q \frac{\partial A_j}{\partial r_i}$$

the first equality because momentum operators and functions commute, and the second equality is again the generalized fundamental commutator.

Note that by assumption the derivatives of $\vec{A}$ are constants, so the side of $\widehat{p}_j - qA_j$ that this result appears is not relevant and what is left of the Hamiltonian becomes

$$\frac{q\mathrm{i}\hbar}{m} \sum_{j=1}^{3} \left\{ \frac{\partial A_i}{\partial r_j} - \frac{\partial A_j}{\partial r_i} \right\} (\widehat{p}_j - qA_j)$$

Now let $\bar{\imath}$ be the index following i in the sequence 123123... and $\bar{\bar{\imath}}$ the one preceding it (or the second following). Then the sum above will have a term where $j = i$, but that term is seen to be zero, a term where $j = \bar{\imath}$, and a term where $j = \bar{\bar{\imath}}$. The total is then:

$$\frac{q\mathrm{i}\hbar}{m} \left\{ (\widehat{p}_{\bar{\imath}} - qA_{\bar{\imath}}) \left(\frac{\partial A_i}{\partial r_{\bar{\imath}}} - \frac{\partial A_{\bar{\imath}}}{\partial r_i} \right) - (\widehat{p}_{\bar{\bar{\imath}}} - qA_{\bar{\bar{\imath}}}) \left(\frac{\partial A_{\bar{\bar{\imath}}}}{\partial r_i} - \frac{\partial A_i}{\partial r_{\bar{\bar{\imath}}}} \right) \right\}$$

and that is

$$-\frac{q\mathrm{i}\hbar}{m} \left\{ (\widehat{p}_{\bar{\imath}} - qA_{\bar{\imath}}) \left(\nabla \times \vec{A} \right)_{\bar{\bar{\imath}}} - (\widehat{p}_{\bar{\bar{\imath}}} - qA_{\bar{\bar{\imath}}}) \left(\nabla \times \vec{A} \right)_{\bar{\imath}} \right\}$$

and the expression in brackets is the i-th component of $(\widehat{\vec{p}} - q\vec{A}) \times \left(\nabla \times \vec{A} \right)$ and produces the $q\vec{v} \times \vec{\mathcal{B}}$ term in Newton's equation provided that $\vec{\mathcal{B}} = \nabla \times \vec{A}$.

D.72 Various electrostatic derivations.

This section gives various derivations for the electromagnetostatic solutions of chapter 13.3.

D.72.1 Existence of a potential

This subsection shows that if the curl of the electric field $\vec{\mathcal{E}}$ (or of any other vector field, like the magnetic one or a force field), is zero, it is minus the gradient of some potential.

That potential can be defined to be

$$\varphi(\vec{r}) = -\int_{\vec{r}_0}^{\vec{r}} \vec{\mathcal{E}}(\underline{\vec{r}})\, \mathrm{d}\underline{\vec{r}} \tag{D.47}$$

where $\vec{r}_0$ is some arbitrarily chosen reference point. You might think that the value of $\varphi(\vec{r})$ would depend on what integration path you took from the reference point to $\vec{r}$, but the Stokes' theorem of calculus says that the difference between integrals leading to the same path must be zero since $\nabla \times \vec{\mathcal{E}}$ is zero.

Now if you evaluate φ at a neighboring point $\vec{r} + \hat{\imath}\partial x$ by a path first going to $\vec{r}$ and from there straight to $\vec{r} + \hat{\imath}\partial x$, the difference in integrals is just the integral over the final segment:

$$\varphi(\vec{r} + \hat{\imath}\partial x) - \varphi(\vec{r}) = -\int_{\vec{r}}^{\vec{r}+\hat{\imath}\partial x} \vec{\mathcal{E}}(\underline{\vec{r}})\,\mathrm{d}\underline{\vec{r}} \tag{D.48}$$

Dividing by ∂x and then taking the limit $\partial x \to 0$ shows that minus the x-derivative of φ gives the x-component of the electric field. The same of course for the other components, since the x-direction is arbitrary.

Note that if regions are multiply connected, the potential may not be quite unique. The most important example of that is the magnetic potential of an infinite straight electric wire. Since the curl of the magnetic field is nonzero inside the wire, the path of integration must stay clear of the wire. It then turns out that the value of the potential depends on how many times the chosen integration path wraps around the wire. Indeed, the magnetic potential is $\varphi_m = -I\theta/2\pi\epsilon_0 c^2$. and as you know, an angle like θ is indeterminate by any integer multiple of 2π.

D.72.2 The Laplace equation

The homogeneous Poisson equation,

$$\nabla^2 \varphi = 0 \tag{D.49}$$

for some unknown function φ is called the Laplace equation. It is very important in many areas of physics and engineering. This note derives some of its generic properties.

The so-called mean-value property says that the average of φ over the surface of any sphere in which the Laplace equation holds is the value of φ at the center of the sphere. To see why, for convenience take the center of the sphere as the origin of a spherical coordinate system. Now

$$\begin{aligned}
0 &= \int_{\text{sphere}} \nabla^2 \varphi \,\mathrm{d}^3\vec{r} \\
&= \iiint \frac{\partial \varphi}{\partial r} r^2 \sin\theta \,\mathrm{d}\theta \mathrm{d}\phi \\
&= \frac{1}{4\pi} \iint \frac{\partial \varphi}{\partial r} \sin\theta \,\mathrm{d}\theta \mathrm{d}\phi \\
&= \frac{\partial}{\partial r} \frac{1}{4\pi} \iint \varphi \sin\theta \,\mathrm{d}\theta \mathrm{d}\phi
\end{aligned}$$

the first equality since φ satisfies the Laplace equation, the second because of the divergence theorem, the third because the integral is zero, so a constant factor does not make a difference, and the fourth by changing the order of integration and differentiation. It follows that the average of φ is the same on all spherical surfaces centered around the origin. Since this includes as a limiting case the origin and the average of φ over the single point at the origin is just φ at the origin, the mean value property follows.

The so called maximum-minimum principle says that either φ is constant everywhere or its maximum and minimum are on a boundary or at infinity. The reason is the mean-value property above. Suppose there is an absolute maximum in the interior of the region in which the Laplace equation applies. Enclose the maximum by a small sphere. Since the values of φ would be less than the maximum on the surface of the sphere, the average value on the surface must be less than the maximum too. But the mean value theorem says it must be the same. The only way around that is if φ is completely constant in the sphere, but then the "maximum" is not a true maximum. And then you can start "sphere-hopping" to show that φ is constant everywhere. Minima go the same way.

The only solution of the Laplace equation in all of space that is zero at infinity is zero everywhere. In more general regions, as long as the solution is zero on all boundaries, including infinity where relevant, then the solution is zero everywhere. The reason is the maximum-minimum principle: if there was a point where the solution was positive/negative, then there would have to be an interior maximum/minimum somewhere.

The solution of the Laplace equation for given boundary values is unique. The reason is that the difference between any two solutions must satisfy the Laplace equation with zero boundary values, hence must be zero.

D.72.3 Egg-shaped dipole field lines

The egg shape of the ideal dipole field lines can be found by assuming that the dipole is directed along the z-axis. Then the field lines in the xz-plane satisfy

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathcal{E}_z}{\mathcal{E}_x} = \frac{2z^2 - x^2}{3zx}$$

Change to a new variable u by replacing z by xu to get:

$$x\frac{\mathrm{d}u}{\mathrm{d}x} = -\frac{1+u^2}{3u} \quad \Longrightarrow \quad \int \frac{3u\,\mathrm{d}u}{1+u^2} = -\int \frac{\mathrm{d}x}{x}$$

Integrating and replacing u again by z/x gives

$$(x^2 + z^2)^{3/2} = Cx^2$$

where C represents the integration constant from the integration. Near the origin, $x \sim z^{3/2}/C$; therefore the field line has infinite curvature at the origin, explaining the pronounced egg shape. Rewritten in spherical coordinates, the field lines are given by $r = C \sin^2\theta$ and ϕ constant, and that is also valid outside the xz-plane.

D.72.4 Ideal charge dipole delta function

Next is the delta function in the electric field generated by a charge distribution that is contracted to an ideal dipole. To find the precise delta function, the electric field can be integrated over a small sphere, but still large enough that on its surface the ideal dipole potential is valid. The integral will give the strength of the delta function. Since the electric field is minus the gradient of the potential, an arbitrary component $\mathcal{E}_i$ integrates to

$$\int_{\text{sphere}} \mathcal{E}_i \,\mathrm{d}^3\vec{r} = -\int_{\text{sphere}} \nabla\cdot(\varphi\hat{\imath}_i)\,\mathrm{d}^3\vec{r} = -\int_{\text{sphere surface}} \varphi n_i \,\mathrm{d}A$$

where $\hat{\imath}_i$ is the unit vector in the i-direction and the divergence theorem of calculus was used to convert the integral to an integral over the surface area A of the sphere. Noting that the vector $\vec{n}$ normal to the surface of the sphere equals $\vec{r}/r$, and that the potential is the ideal dipole one, you get

$$\int_{\text{sphere}} \mathcal{E}_i \,\mathrm{d}^3\vec{r} = -\frac{1}{4\pi\epsilon_0}\int_{\text{sphere surface}} \frac{\vec{\wp}\cdot\vec{r}}{r^3}\frac{r_i}{r}\,\mathrm{d}A$$

For simplicity, take the z-axis along the dipole moment; then $\vec{\wp}\cdot\vec{r} = \wp z$. For the x-component $\mathcal{E}_x$, $r_i = x$ so that the integrand is proportional to xz, and that integrates to zero over the surface of the sphere because the negative x-values cancel the positive ones at the same z. The same for the y-component of the field, so only the z-component, or more generally, the component in the same direction as $\vec{\wp}$, has a delta function. For $\mathcal{E}_z$, you are integrating z^2, and by symmetry that is the same as integrating x^2 or y^2, so it is the same as integrating $\frac{1}{3}r^2$. Since the surface of the sphere equals $4\pi r^2$, the delta function included in the expression for the field of a dipole as listed in table 13.2 is obtained.

D.72.5 Integrals of the current density

In subsequent derivations, various integrals of the current density $\vec{\jmath}$ are needed. In all cases it is assumed that the current density vanishes strongly outside some region. Of course, normally an electric motor or electromagnet has electrical leads going towards and away from of it; it is assumed that these are stranded so tightly together that their net effect can be ignored.

Consider an integral like $\int {r_i}^m r_{\bar{\imath}}^n j_i \,\mathrm{d}^3\vec{r}$ where j_i is any component j_1, j_2, or j_3 of the current density, $\bar{\imath}$ is the index following i in the sequence $\ldots 123123\ldots$,

m and n are nonnegative integers, and the integration is over all of space. By integration by parts in the i-direction, and using the fact that the current densities vanish at infinity,

$$\int r_i^m r_{\bar{\imath}}^n j_i \,\mathrm{d}^3\vec{r} = -\int \frac{r_i^{m+1}}{m+1} r_{\bar{\imath}}^n \frac{\partial j_i}{\partial r_i}\,\mathrm{d}^3\vec{r}$$

Now use the fact that the divergence of the current density is zero since the charge density is constant for electromagnetostatic solutions:

$$\int r_i^m r_{\bar{\imath}}^n j_i \,\mathrm{d}^3\vec{r} = \int \frac{r_i^{m+1}}{m+1} r_{\bar{\imath}}^n \frac{\partial j_{\bar{\imath}}}{\partial r_{\bar{\imath}}}\,\mathrm{d}^3\vec{r} + \int \frac{r_i^{m+1}}{m+1} r_{\bar{\imath}}^n \frac{\partial j_{\bar{\bar{\imath}}}}{\partial r_{\bar{\bar{\imath}}}}\,\mathrm{d}^3\vec{r}$$

where $\bar{\bar{\imath}}$ is the index preceding i in the sequence $\ldots 123123\ldots$. The final integral can be integrated in the $\bar{\bar{\imath}}$-direction and is then seen to be zero because $\vec{\jmath}$ vanishes at infinity.

The first integral in the right hand side can be integrated by parts in the $\bar{\imath}$-direction to give the final result:

$$\int r_i^m r_{\bar{\imath}}^n j_i \,\mathrm{d}^3\vec{r} = -\int \frac{r_i^{m+1}}{m+1} n r_{\bar{\imath}}^{n-1} j_{\bar{\imath}}\,\mathrm{d}^3\vec{r} \tag{D.50}$$

It follows from this equation with $m = 0$, $n = 1$ that

$$\int r_i j_{\bar{\imath}}\,\mathrm{d}^3\vec{r} = -\int r_{\bar{\imath}} j_i \,\mathrm{d}^3\vec{r} = \mu_{\bar{\bar{\imath}}} \qquad \vec{\mu} \equiv \tfrac{1}{2}\int \vec{r}\times\vec{\jmath}\,\mathrm{d}^3\vec{r} \tag{D.51}$$

with $\vec{\mu}$ the current distribution's dipole moment. In these expressions, you can swap indices as

$$(i,\bar{\imath},\bar{\bar{\imath}}) \to (\bar{\imath},\bar{\bar{\imath}},i) \qquad \text{or} \qquad (i,\bar{\imath},\bar{\bar{\imath}}) \to (\bar{\bar{\imath}},i,\bar{\imath})$$

because only the relative ordering of the indices in the sequence $\ldots 123123\ldots$ is relevant.

In quantum applications, it is often necessary to relate the dipole moment to the angular momentum of the current carriers. Since the current density is the charge per unit volume times its velocity, you get the linear momentum per unit volume by multiplying by the ratio $m_{\mathrm{c}}/q_{\mathrm{c}}$ of current carrier mass over charge. Then the angular momentum is

$$\vec{L} = \int \vec{r}\times\frac{m_{\mathrm{c}}}{q_{\mathrm{c}}}\vec{\jmath}\,\mathrm{d}^3\vec{r} = \frac{2m_{\mathrm{c}}}{q_{\mathrm{c}}}\vec{\mu}$$

D.72.6 Lorentz forces on a current distribution

Next is the derivation of the Lorentz forces on a given current distribution $\vec{\jmath}$ in a constant external magnetic field $\vec{\mathcal{B}}_{\mathrm{ext}}$. The Lorentz force law says that the force $\vec{F}$ on a charge q moving with speed $\vec{v}$ equals

$$\vec{F} = q\vec{v}\times\vec{\mathcal{B}}_{\mathrm{ext}}$$

In terms of a current distribution, the moving charge per unit volume times its velocity is the current density, so the force on a volume element $\mathrm{d}^3\vec{r}$ is:

$$\mathrm{d}\vec{F} = \vec{\jmath} \times \vec{\mathcal{B}}_{\text{ext}}\, \mathrm{d}^3\vec{r}$$

The net force on the current distribution is therefore zero, because according to (D.50) with $m = n = 0$, the integrals of the components of the current distribution are zero.

The moment is not zero, however. It is given by

$$\vec{M} = \int \vec{r} \times \left(\vec{\jmath} \times \vec{\mathcal{B}}_{\text{ext}} \right) \mathrm{d}^3\vec{r}$$

According to the vectorial triple product rule, that is

$$\vec{M} = \int \left(\vec{r} \cdot \vec{\mathcal{B}}_{\text{ext}} \right) \vec{\jmath}\, \mathrm{d}^3\vec{r} - \int \left(\vec{r} \cdot \vec{\jmath} \right) \vec{\mathcal{B}}_{\text{ext}}\, \mathrm{d}^3\vec{r}$$

The second integral is zero because of (D.50) with $m = 1$, $n = 0$. What is left is can be written in index notation as

$$M_i = \int r_i \mathcal{B}_{\text{ext,i}} j_i\, \mathrm{d}^3\vec{r} + \int r_{\bar{\imath}} \mathcal{B}_{\text{ext},\bar{\imath}} j_i\, \mathrm{d}^3\vec{r} + \int r_{\bar{\bar{\imath}}} \mathcal{B}_{\text{ext},\bar{\bar{\imath}}} j_i\, \mathrm{d}^3\vec{r}$$

The first of the three integrals is zero because of (D.50) with $m = 1$, $n = 0$. The other two can be rewritten using (D.51):

$$M_i = -\mu_{\bar{\imath}} \mathcal{B}_{\text{ext},\bar{\imath}} + \mu_{\bar{\imath}} \mathcal{B}_{\text{ext},\bar{\bar{\imath}}}$$

and in vector notation that reads

$$\vec{M} = \vec{\mu} \times \vec{\mathcal{B}}_{\text{ext}}$$

When the (frozen) current distribution is slowly rotated around the axis aligned with the moment vector, the work done is

$$-M\, \mathrm{d}\alpha = -\mu \mathcal{B}_{\text{ext}} \sin\alpha\, \mathrm{d}\alpha = \mathrm{d}(\mu \mathcal{B}_{\text{ext}} \cos\alpha)$$

where α is the angle between $\vec{\mu}$ and $\vec{\mathcal{B}}_{\text{ext}}$. By integration, it follows that the work done corresponds to a change in energy for an energy given by

$$E_{\text{ext}} = -\vec{\mu} \cdot \vec{\mathcal{B}}_{\text{ext}}$$

D.72.7 Field of a current dipole

A current density $\vec{\jmath}$ creates a magnetic field because of Maxwell's second and fourth equations for the divergence and curl of the magnetic field:

$$\nabla \cdot \vec{\mathcal{B}} = 0 \qquad \nabla \times \vec{\mathcal{B}} = \frac{1}{\epsilon_0 c^2} \vec{\jmath}$$

where $\vec{\mathcal{B}}$ vanishes at infinity assuming there is no additional ambient magnetic field.

A magnetic vector potential $\vec{A}$ will now be defined as the solution of the Poisson equation

$$\nabla^2 \vec{A} = -\frac{1}{\epsilon_0 c^2} \vec{\jmath}$$

that vanishes at infinity. Taking the divergence of this equation shows that the divergence of the vector potential satisfies a homogeneous Poisson equation, because the divergence of the current density is zero, with zero boundary conditions at infinity. Therefore the divergence of the vector potential is zero. It then follows that

$$\vec{\mathcal{B}} = \nabla \times \vec{A}$$

because it satisfies the equations for $\vec{\mathcal{B}}$: the divergence of any curl is zero, and the curl of the curl of the vector potential is according to the vectorial triple product its Laplacian, hence the correct curl of the magnetic field.

You might of course wonder whether there might not be more than one magnetic field that has the given divergence and curl and is zero at infinity. The answer is no. The difference between any two such fields must have zero divergence and curl. Therefore the curl of the curl of the difference is zero too, and the vectorial triple product shows that equal to minus the Laplacian of the difference. If the Laplacian of the difference is zero, then the difference is zero, since the difference is zero at infinity (subsection 2). So the solutions must be the same.

Since the integrals of the current density are zero, (D.50) with $m = n = 0$, the asymptotic expansion (13.31) of the Green's function integral shows that at large distances, the components of $\vec{A}$ behave as a dipole potential. Specifically,

$$A_i \sim \frac{1}{4\pi\epsilon_0 c^2 r^3} \sum_{\underline{\imath}=1}^{3} r_{\underline{\imath}} \int \underline{r}_{\underline{\imath}} j_i \, \mathrm{d}^3\vec{\underline{r}}$$

Now the term $\underline{\imath} = i$ in the sum does not give a contribution, because of (D.50) with $m = 1$, $n = 0$. The other two terms are

$$A_i \sim \frac{1}{4\pi\epsilon_0 c^2 r^3} \left[r_{\bar{\imath}} \int \underline{r}_{\bar{\imath}} j_i \, \mathrm{d}^3\vec{\underline{r}} + r_{\bar{\bar{\imath}}} \int \underline{r}_{\bar{\bar{\imath}}} j_i \, \mathrm{d}^3\vec{\underline{r}} \right]$$

with $\bar{\imath}$ following i in the sequence $\ldots 123123 \ldots$ and $\bar{\bar{\imath}}$ preceding it. These two integrals can be rewritten using (D.51) to give

$$A_i \sim -\frac{1}{4\pi\epsilon_0 c^2 r^3} \left[r_{\bar{\imath}} \mu_{\bar{\bar{\imath}}} - r_{\bar{\bar{\imath}}} \mu_{\bar{\imath}} \right]$$

Note that the expression between brackets is just the i-th component of $\vec{r} \times \vec{\mu}$.

The magnetic field is the curl of $\vec{A}$, so

$$\mathcal{B}_i = \frac{\partial A_{\bar{\bar{\imath}}}}{\partial r_{\bar{\imath}}} - \frac{\partial A_{\bar{\imath}}}{\partial r_{\bar{\bar{\imath}}}}$$

and substituting in for the vector potential from above, differentiating, and cleaning up produces

$$\mathcal{B}_i = \frac{3(\vec{\mu} \cdot \vec{r})\vec{r} - \vec{\mu} r^2}{4\pi\epsilon_0 c^2 r^5}$$

This is the same asymptotic field as a charge dipole with strength $\vec{\mu}$ would have.

However, for an ideal current dipole, the delta function at the origin will be different than that derived for a charge dipole in the first subsection. Integrate the magnetic field over a sphere large enough that on its surface, the asymptotic field is accurate:

$$\int \mathcal{B}_i \, \mathrm{d}^3\vec{r} = \int \frac{\partial A_{\bar{\bar{\imath}}}}{\partial r_{\bar{\imath}}} \, \mathrm{d}^3\vec{r} - \int \frac{\partial A_{\bar{\imath}}}{\partial r_{\bar{\bar{\imath}}}} \, \mathrm{d}^3\vec{r}$$

Using the divergence theorem, the right hand side becomes an integral over the surface of the sphere:

$$\int \mathcal{B}_i \, \mathrm{d}^3\vec{r} = \int A_{\bar{\bar{\imath}}} \frac{r_{\bar{\imath}}}{r} \, \mathrm{d}A - \int A_{\bar{\imath}} \frac{r_{\bar{\bar{\imath}}}}{r} \, \mathrm{d}A$$

Substituting in the asymptotic expression for A_i above,

$$\int \mathcal{B}_i \, \mathrm{d}^3\vec{r} = -\frac{1}{4\pi\epsilon_0 c^2 r^4} \left[\int (r_i \mu_{\bar{\imath}} - r_{\bar{\imath}} \mu_i) r_{\bar{\imath}} \, \mathrm{d}A - \int (r_{\bar{\bar{\imath}}} \mu_i - r_i \mu_{\bar{\bar{\imath}}}) r_{\bar{\bar{\imath}}} \, \mathrm{d}A \right]$$

The integrals of $r_i r_{\bar{\imath}}$ and $r_i r_{\bar{\bar{\imath}}}$ are zero, for one because the integrand is odd in r_i. The integrals of $r_{\bar{\imath}} r_{\bar{\imath}}$ and $r_{\bar{\bar{\imath}}} r_{\bar{\bar{\imath}}}$ are each one third of the integral of r^2 because of symmetry. So, noting that the surface area A of the spherical surface is $4\pi r^2$,

$$\int \mathcal{B}_i \, \mathrm{d}^3\vec{r} = \frac{2}{3\epsilon_0 c^2} \mu_i$$

That gives the strength of the delta function for an ideal current dipole.

D.72.8 Biot-Savart law

In the previous section, it was noted that the magnetic field of a current distribution is the curl of a vector potential $\vec{A}$. This vector potential satisfies the Poisson equation

$$\nabla^2 \vec{A} = -\frac{1}{\epsilon_0 c^2} \vec{\jmath}$$

The solution for the vector potential can be written explicitly in terms of the current density using the Green's function integral (13.29):

$$A_i = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{1}{|\vec{r} - \underline{\vec{r}}|} j_i(\underline{\vec{r}}) \, \mathrm{d}^3\underline{\vec{r}}$$

The magnetic field is the curl of $\vec{A}$,

$$\mathcal{B}_i = \frac{\partial A_{\bar{\imath}}}{\partial r_{\bar{\imath}}} - \frac{\partial A_{\bar{\imath}}}{\partial r_{\bar{\imath}}}$$

or substituting in and differentiating under the integral

$$\mathcal{B}_i = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{r_{\bar{\imath}} - \underline{r}_{\bar{\imath}}}{|\vec{r} - \underline{\vec{r}}|^3} j_{\bar{\imath}}(\underline{\vec{r}}) - \frac{r_{\bar{\imath}} - \underline{r}_{\bar{\imath}}}{|\vec{r} - \underline{\vec{r}}|^3} j_{\bar{\imath}}(\underline{\vec{r}}) \, \mathrm{d}^3\underline{\vec{r}}$$

In vector notation that gives the Biot-Savart law

$$\vec{\mathcal{B}} = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{\vec{r} - \underline{\vec{r}}}{|\vec{r} - \underline{\vec{r}}|^3} \times \vec{\jmath} \mathrm{d}^3\underline{\vec{r}}$$

Now assume that the current distribution is limited to one or more thin wires, as it usually is. In that case, a volume element of nonzero current distribution can be written as

$$\vec{\jmath} \mathrm{d}^3\underline{\vec{r}} = I \, \mathrm{d}\underline{\vec{r}}$$

where in the right hand side $\underline{\vec{r}}$ describes the position of the centerline of the wire and I is the current through the wire. More specifically, I is the integral of the current density over the cross section of the wire. The Biot-Savart law becomes

$$\vec{\mathcal{B}} = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{\vec{r} - \underline{\vec{r}}}{|\vec{r} - \underline{\vec{r}}|^3} \times I(\underline{\vec{r}}) \, \mathrm{d}\underline{\vec{r}}$$

where the integration is over all infinitesimal segments $\mathrm{d}\underline{\vec{r}}$ of the wires.

D.73 Orbital motion in a magnetic field

This note derives the energy of a charged particle in an external magnetic field. The field is assumed constant.

According to chapter 13.1, the Hamiltonian is

$$H = \frac{1}{2m}\left(\widehat{\vec{p}} - q\vec{A}\right)^2 + V$$

where m and q are the mass and charge of the particle and the vector potential $\vec{A}$ is related to the magnetic field $\vec{\mathcal{B}}$ by $\vec{\mathcal{B}} = \nabla \times \vec{A}$. The potential energy V is of no particular interest in this note. The first term is, and it can be multiplied out as:

$$H = \frac{1}{2m}\widehat{\vec{p}}^2 - \frac{q}{2m}\left(\widehat{\vec{p}}\cdot\vec{A} + \vec{A}\cdot\widehat{\vec{p}}\right) + \frac{q^2}{2m}\left(\vec{A}\right)^2 + V$$

The middle two terms in the right hand side are the changes in the Hamiltonian due to the magnetic field; they will be denoted as:

$$H_{\mathcal{B}L} \equiv -\frac{q}{2m}\left(\widehat{\vec{p}}\cdot\vec{A} + \vec{A}\cdot\widehat{\vec{p}}\right) \qquad H_{\mathcal{B}D} \equiv \frac{q^2}{2m}\left(\vec{A}\right)^2$$

Now to simplify the analysis, align the z-axis with $\vec{\mathcal{B}}$ so that $\vec{\mathcal{B}} = \hat{k}\mathcal{B}_z$. Then an appropriate vector potential $\vec{A}$ is

$$\vec{A} = -\hat{\imath}\tfrac{1}{2}y\mathcal{B}_z + \hat{\jmath}\tfrac{1}{2}x\mathcal{B}_z.$$

The vector potential is not unique, but a check shows that indeed $\nabla \times \vec{A} = \hat{k}\mathcal{B}_z = \vec{\mathcal{B}}$ for the one above. Also, the canonical momentum is

$$\widehat{\vec{p}} = \frac{\hbar}{\mathrm{i}}\nabla = \hat{\imath}\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial y} + \hat{k}\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial z}$$

Therefore, in the term $H_{\mathcal{B}L}$ above,

$$H_{\mathcal{B}L} = -\frac{q}{2m}(\widehat{\vec{p}}\cdot\vec{A} + \vec{A}\cdot\widehat{\vec{p}}) = -\frac{q}{2m}\mathcal{B}_z\left(x\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial y} - y\frac{\hbar}{\mathrm{i}}\frac{\partial}{\partial x}\right) = -\frac{q}{2m}\mathcal{B}_z\widehat{L}_z$$

the latter equality being true because of the definition of angular momentum as $\vec{r}\times\widehat{\vec{p}}$. Because the z-axis was aligned with $\vec{\mathcal{B}}$, $\mathcal{B}_z\widehat{L}_z = \vec{\mathcal{B}}\cdot\widehat{\vec{L}}$, so, finally,

$$H_{\mathcal{B}L} = -\frac{q}{2m}\vec{\mathcal{B}}\cdot\widehat{\vec{L}}.$$

Similarly, in the part $H_{\mathcal{B}D}$ of the Hamiltonian, substitution of the expression for $\vec{A}$ produces

$$\frac{q^2}{2m}\left(\vec{A}\right)^2 = \frac{q^2}{8m}\mathcal{B}_z^2\left(x^2+y^2\right),$$

or writing it so that it is independent of how the z-axis is aligned,

$$H_{\mathcal{B}D} = \frac{q^2}{8m}\left(\vec{\mathcal{B}}\times\vec{r}\right)^2$$

D.74 Electron spin in a magnetic field

If you are curious how the magnetic dipole strength of the electron can just pop out of the relativistic Dirac equation, this note gives a quick derivation.

First, a problem must be addressed. Dirac's equation, chapter 12.12, assumes that Einstein's energy square root falls apart in a linear combination of terms:

$$H = \sqrt{(mc^2)^2 + \sum_{i=1}^{3} (\widehat{p}_i c)^2} = \alpha_0 mc^2 + \sum_{i=1}^{3} \alpha_i \widehat{p}_i c$$

which works for the 4×4 α matrices given in that section. For an electron in a magnetic field, according to chapter 13.1 you want to replace $\widehat{\vec{p}}$ with $\widehat{\vec{p}} - q\vec{A}$ where $\vec{A}$ is the magnetic vector potential. But where should you do that, in the square root or in the linear combination? It turns out that the answer you get for the electron energy is *not* the same.

If you believe that the Dirac linear combination is the way physics really works, and its description of spin leaves little doubt about that, then the answer is clear: you need to put $\widehat{\vec{p}} - q\vec{A}$ in the linear combination, not in the square root.

So, what are now the energy levels? That would be hard to say directly from the linear form, so square it down to H^2, using the properties of the α matrices, as given in chapter 12.12 and its note. You get, in index notation,

$$H^2 = \big(mc^2\big)^2 I + \sum_{i=1}^{3} \Big((\widehat{p}_i - qA_i)c\Big)^2 I + \sum_{i=1}^{3} [\widehat{p}_{\bar{\imath}} - qA_{\bar{\imath}}, \widehat{p}_{\bar{\bar{\imath}}} - qA_{\bar{\bar{\imath}}}] c^2 \alpha_{\bar{\imath}} \alpha_{\bar{\bar{\imath}}}$$

where I is the four by four unit matrix, $\bar{\imath}$ is the index following i in the sequence 123123..., and $\bar{\bar{\imath}}$ is the one preceding i. The final sum represents the additional squared energy that you get by substituting $\widehat{\vec{p}} - q\vec{A}$ in the linear combination instead of the square root. The commutator arises because $\alpha_{\bar{\imath}}\alpha_{\bar{\bar{\imath}}} + \alpha_{\bar{\bar{\imath}}}\alpha_{\bar{\imath}} = 0$, giving the terms with the indices reversed the opposite sign. Working out the commutator using the formulae of chapter 4.5.4, and the definition of the vector potential $\vec{A}$,

$$H^2 = \big(mc^2\big)^2 I + \sum_{i=1}^{3} \Big((\widehat{p}_i - qA_i)c\Big)^2 I + q\hbar c^2 \mathrm{i} \sum_{i=1}^{3} \mathcal{B}_i \alpha_{\bar{\imath}} \alpha_{\bar{\bar{\imath}}}.$$

By multiplying out the expressions for the α_i of chapter 12.12, using the fundamental commutation relation for the Pauli spin matrices that $\sigma_{\bar{\imath}}\sigma_{\bar{\bar{\imath}}} = \mathrm{i}\sigma_i$,

$$H^2 = \big(mc^2\big)^2 I + \sum_{i=1}^{3} \Big((\widehat{p}_i - qA_i)c\Big)^2 I - q\hbar c^2 \sum_{i=1}^{3} \mathcal{B}_i \begin{pmatrix} \sigma_i & 0 \\ 0 & \sigma_i \end{pmatrix}$$

It it seen that due to the interaction of the spin with the magnetic field, the square energy changes by an amount $-qhc^2\sigma_i\mathcal{B}_i$. Since $\frac{1}{2}\hbar$ times the Pauli spin matrices gives the spin $\widehat{\vec{S}}$, the square energy due to the magnetic field acting on the spin is $-2qc^2\widehat{\vec{S}}\cdot\vec{\mathcal{B}}$.

In the nonrelativistic case, the rest mass energy mc^2 is much larger than the other terms, and in that case, if the change in square energy is $-2qc^2\widehat{\vec{S}}\cdot\vec{\mathcal{B}}$, the change in energy itself is smaller by a factor $2mc^2$, so the energy due to the magnetic field is

$$H_{\mathcal{B}S} = -\frac{q}{m}\widehat{\vec{S}}\cdot\vec{\mathcal{B}} \tag{D.52}$$

which is what was to be proved.

D.75 Solving the NMR equations

To solve the two coupled ordinary differential equations for the spin up and down probabilities, first get rid of the time dependence of the right-hand-side matrix by defining new variables A and B by

$$a = Ae^{\mathrm{i}\omega t/2}, \quad b = Be^{-\mathrm{i}\omega t/2}.$$

Then find the eigenvalues and eigenvectors of the now constant matrix. The eigenvalues can be written as $\pm\mathrm{i}\omega_1/f$, where f is the resonance factor given in the main text. The solution is then

$$\begin{pmatrix} A \\ B \end{pmatrix} = C_1\vec{v}_1 e^{\mathrm{i}\omega_1 t/f} + C_2\vec{v}_2 e^{-\mathrm{i}\omega_1 t/f}$$

where $\vec{v}_1$ and $\vec{v}_2$ are the eigenvectors. To find the constants C_1 and C_2, apply the initial conditions $A(0) = a(0) = a_0$ and $B(0) = b(0) = b_0$ and clean up as well as possible, using the definition of the resonance factor and the Euler formula.

It's a mess.

D.76 Harmonic oscillator revisited

This note rederives the harmonic oscillator solution, but in spherical coordinates. The reason to do so is to obtain energy eigenfunctions that are also eigenfunctions of square angular momentum and of angular momentum in the z-direction. The derivation is very similar to the one for the hydrogen atom given in derivation {D.15}, so the discussion will mainly focus on the differences.

The solutions are again in the form $R(r)Y_l^m(\theta,\phi)$ with the Y_l^m the spherical harmonics. However, the radial functions R are different; the equation for them is now

$$-\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right)+l(l+1)+\frac{2m_{\mathrm{e}}}{\hbar^2}\tfrac{1}{2}m_{\mathrm{e}}\omega^2r^4=\frac{2m_e}{\hbar^2}r^2E$$

The difference from {D.15} is that a harmonic oscillator potential $\frac{1}{2}m_e\omega^2r^2$ has replaced the Coulomb potential. A suitable rescaling is now $r=\rho\sqrt{\hbar/m_{\mathrm{e}}\omega}$, which produces

$$-\frac{1}{R}\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\rho^2\frac{\mathrm{d}R}{\mathrm{d}\rho}\right)+l(l+1)+\rho^4=\rho^2\epsilon$$

where $\epsilon=E/\frac{1}{2}\hbar\omega$ is the energy in half quanta.

Split off the expected asymptotic behavior for large ρ by defining

$$R=e^{-\rho^2/2}f$$

Then f satisfies

$$\rho^2f''+2\rho f'-l(l+1)f=2\rho^3f'+(3-\epsilon)\rho^2f$$

Plug in a power series $f=\sum_p c_p\rho^p$, then the coefficients must satisfy:

$$[p(p+1)-l(l+l)]c_p=[2(p-2)+3-\epsilon]c_{p-2}$$

From that it is seen that the lowest power in the series is $p_{\min}=l$, $p_{\min}=-l-1$ not being acceptable. Also the series must terminate, or blow up will occur. That requires that $\epsilon=2p_{\max}+3$. So the energy must be $(p_{\max}+\frac{3}{2})\hbar\omega$ with $p_{\max}$ an integer no smaller than l, so at least zero.

Therefore, numbering the energy levels from $n=1$ like for the hydrogen level gives the energy levels as

$$E_n=(n+\frac{1}{2})\hbar\omega$$

That are the same energy levels as derived in Cartesian coordinates, as they should be. However, the eigenfunctions are different. They are of the form

$$\psi_{nlm}=e^{-\rho^2/2}P_{nl}(\rho)Y_l^m(\theta,\phi)$$

where P_{nl} is some polynomial of degree $n-1$, whose lowest power of ρ is ρ^l. The value of the azimuthal quantum number l must run up to $n-1$ like for the hydrogen atom. However, in this case l must be odd or even depending on whether $n-1$ is odd or even, or the power series will not terminate.

Note that for even l, the power series proceed in even powers of r. These eigenfunctions are said to have even parity: if you replace r by $-r$, they are unchanged. Similarly, the eigenfunctions for odd l expand in odd powers of r. They are said to have odd parity; if you replace r by $-r$, they change sign.

D.77 Impenetrable spherical shell

To solve the problem of particles stuck inside an impenetrable shell of radius a, refer to addendum {A.6}. According to that addendum, the solutions without unacceptable singularities at the center are of the form

$$\psi_{Elm}(r,\theta,\phi) \propto j_l(p_{rmc}r/\hbar)Y_l^m(\theta,\phi) \qquad p_{rmc} \equiv \sqrt{2m(E-V)} \tag{D.53}$$

where the j_l are the spherical Bessel functions of the first kind, the Y_l^m the spherical harmonics, and p_{rmc} is the classical momentum of a particle with energy E. V_0 is the constant potential inside the shell, which can be taken to be zero without fundamentally changing the solution.

Because the wave function must be zero at the shell $r = a$, $p_{rmc}a/\hbar$ must be one of the zero-crossings of the spherical Bessel functions. Therefore the allowable energy levels are

$$E_{\bar{n}l} = \frac{hbar^2}{2ma^2}\beta_{\bar{n}l}^2 + V_0 \tag{D.54}$$

where $\beta_{\bar{n}l}$ is the $\bar{n}$-th zero-crossing of spherical Bessel function j_l (not counting the origin). Those crossings can be found tabulated in for example [1], (under the guise of the Bessel functions of half-integer order.)

In terms of the count n of the energy levels of the harmonic oscillator, $\bar{n} = 1$ corresponds to energy level $n = l + 1$, and each next value of $\bar{n}$ increases the energy levels by two, so

$$n = l - 1 + 2\bar{n}$$

D.78 Shell model quadrupole moment

The result for one proton is readily available in literature and messy to derive yourself. If you want to give it a try anyway, one way is the following. Note that in spherical coordinates

$$3z^2 - r^2 = 2r^2 - 3r^2\sin^2\theta$$

and the first term produces $2\langle r^2\rangle$ simply by the definition of expectation value. The problem is to get rid of the $\sin^2\theta$ in the second expectation value.

To do so, use chapter 12.8, 2. That shows that the second term is essentially $3\langle r^2\rangle$ modified by factors of the form

$$\langle Y_l^l|sin^2\theta Y_l^l\rangle \quad \text{and} \quad \langle Y_l^{l-1}|sin^2\theta Y_l^{l-1}\rangle$$

where the integration is over the unit sphere. If you use the representation of the spherical harmonics as given in {D.64}, you can relate these inner products to the unit inner products

$$\langle Y_{l+1}^{l+1}|Y_{l+1}^{l+1}\rangle \quad \text{and} \quad \langle Y_{l+1}^{l}|Y_{l+1}^{l}\rangle$$

Have fun.

The expression for the quadrupole moment if there are an odd number $i \geqslant 3$ of protons in the shell would seem to be a very messy exercise. Some text books suggest that the odd-particle shell model implies that the one-proton value applies for any odd number of protons in the shell. However, it is clear from the state with a single hole that this is untrue. The cited result that the quadrupole moment varies linearly with the odd number of protons in the shell comes directly from Krane, [31, p. 129]. No derivation or reference is given. In fact, the restriction to an odd number of protons is not even stated. If you have a reference or a simple derivation, let me know and I will add it here.

D.79 Derivation of perturbation theory

This note derives the perturbation theory results for the solution of the eigenvalue problem $(H_0 + H_1)\psi = E\psi$ where H_1 is small. The considerations for degenerate problems use linear algebra.

First, "small" is not a valid mathematical term. There are no small numbers in mathematics, just numbers that become zero in some limit. Therefore, to mathematically analyze the problem, the perturbation Hamiltonian will be written as

$$H_1 \equiv \varepsilon H_\varepsilon$$

where ε is some chosen number that physically indicates the magnitude of the perturbation potential. For example, if the perturbation is an external electric field, ε could be taken as the reference magnitude of the electric field. In perturbation analysis, ε is assumed to be vanishingly small.

The idea is now to start with a good eigenfunction $\psi_{\vec{n},0}$ of H_0, (where "good" is still to be defined), and correct it so that it becomes an eigenfunction of $H = H_0 + H_1$. To do so, both the desired energy eigenfunction and its energy eigenvalue are expanded in a power series in terms of ε:

$$\psi_{\vec{n}} = \psi_{\vec{n},0} + \varepsilon\psi_{\vec{n},\varepsilon} + \varepsilon^2\psi_{\vec{n},\varepsilon^2} + \ldots$$

$$E_{\vec{n}} = E_{\vec{n},0} + \varepsilon E_{\vec{n},\varepsilon} + \varepsilon^2 E_{\vec{n},\varepsilon^2} + \ldots$$

If ε is a small quantity, then ε^2 will be much smaller still, and can probably be ignored. If not, then surely ε^3 will be so small that it can be ignored. A result that forgets about powers of ε higher than one is called first order perturbation theory. A result that also includes the quadratic powers, but forgets about powers higher than two is called second order perturbation theory, etcetera.

Before proceeding with the practical application, a disclaimer is needed. While it is relatively easy to see that the eigenvalues expand in whole powers of ε, (note that they must be real whether ε is positive or negative), it is much more messy to show that the eigenfunctions must expand in whole powers. In

fact, for degenerate energies $E_{\vec{n},0}$ they only do if you choose good states $\psi_{\vec{n},0}$. See Rellich's lecture notes on Perturbation Theory [Gordon & Breach, 1969] for a proof. As a result the problem with degeneracy becomes that the good unperturbed eigenfunction $\psi_{\vec{n},0}$ is initially unknown. It leads to lots of messiness in the procedures for degenerate eigenvalues described below.

When the above power series are substituted into the eigenvalue problem to be solved,

$$(H_0 + \varepsilon H_\varepsilon)\,\psi_{\vec{n}} = E_{\vec{n}}\psi_{\vec{n}}$$

the net coefficient of *every* power of ε must be equal in the left and right hand sides. Collecting these coefficients and rearranging them appropriately produces:

$$\begin{aligned}
\varepsilon^0 &: \quad (H_0 - E_{\vec{n},0})\psi_{\vec{n},0} = 0 \\
\varepsilon^1 &: \quad (H_0 - E_{\vec{n},0})\psi_{\vec{n},\varepsilon} = -H_\varepsilon\psi_{\vec{n},0} + E_{\vec{n},\varepsilon}\psi_{\vec{n},0} \\
\varepsilon^2 &: \quad (H_0 - E_{\vec{n},0})\psi_{\vec{n},\varepsilon^2} = -H_\varepsilon\psi_{\vec{n},\varepsilon} + E_{\vec{n},\varepsilon}\psi_{\vec{n},\varepsilon} + E_{\vec{n},\varepsilon^2}\psi_{\vec{n},0} \\
\varepsilon^3 &: \quad (H_0 - E_{\vec{n},0})\psi_{\vec{n},\varepsilon^3} = -H_\varepsilon\psi_{\vec{n},\varepsilon^2} + E_{\vec{n},\varepsilon}\psi_{\vec{n},\varepsilon^2} + E_{\vec{n},\varepsilon^2}\psi_{\vec{n},\varepsilon} + E_{\vec{n},\varepsilon^3}\psi_{\vec{n},0} \\
&\vdots \quad \cdots
\end{aligned}$$

These are the equations to be solved in succession to give the various terms in the expansion for the wave function $\psi_{\vec{n}}$ and the energy $E_{\vec{n}}$. The further you go down the list, the better your combined result should be.

Note that all it takes is to solve problems of the form

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},\ldots} = \ldots$$

The equations for the unknown functions are in terms of the unperturbed Hamiltonian H_0, with some additional but in principle knowable terms.

For difficult perturbation problems like you find in engineering, the use of a small parameter ε is essential to get the mathematics right. But in the simple applications in quantum mechanics, it is usually overkill. So most of the time the expansions are written without, like

$$\psi_{\vec{n}} = \psi_{\vec{n},0} + \psi_{\vec{n},1} + \psi_{\vec{n},2} + \ldots$$

$$E_{\vec{n}} = E_{\vec{n},0} + E_{\vec{n},1} + E_{\vec{n},2} + \ldots$$

where you are assumed to just imagine that $\psi_{\vec{n},1}$ and $E_{\vec{n},1}$ are "first order small," $\psi_{\vec{n},2}$ and $E_{\vec{n},2}$ are "second order small," etcetera. In those terms, the successive equations to solve are:

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},0} = 0 \tag{D.55}$$

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},1} = -H_1\psi_{\vec{n},0} + E_{\vec{n},1}\psi_{\vec{n},0} \tag{D.56}$$

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},2} = -H_1\psi_{\vec{n},1} + E_{\vec{n},1}\psi_{\vec{n},1} + E_{\vec{n},2}\psi_{\vec{n},0} \tag{D.57}$$

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},3} = -H_1\psi_{\vec{n},2} + E_{\vec{n},1}\psi_{\vec{n},2} + E_{\vec{n},2}\psi_{\vec{n},1} + E_{\vec{n},3}\psi_{\vec{n},0} \tag{D.58}$$

$\ldots$

Now consider each of these equations in turn. First, (D.55) is just the Hamiltonian eigenvalue problem for H_0 and is already satisfied by the chosen unperturbed solution $\psi_{\vec{n},0}$ and its eigenvalue $E_{\vec{n},0}$. However, the remaining equations are not trivial. To solve them, write their solutions in terms of the other eigenfunctions $\psi_{\underline{\vec{n}},0}$ of the *unperturbed* Hamiltonian H_0. In particular, to solve (D.56), write

$$\psi_{\vec{n},1} = \sum_{\underline{\vec{n}} \neq \vec{n}} c_{\underline{\vec{n}},1}\psi_{\underline{\vec{n}},0}$$

where the coefficients $c_{\underline{\vec{n}},1}$ are still to be determined. The coefficient of $\psi_{\vec{n},0}$ is zero on account of the normalization requirement. (And in fact, it is easiest to take the coefficient of $\psi_{\vec{n},0}$ also zero for $\psi_{\vec{n},2}$, $\psi_{\vec{n},3}$, $\ldots$, even if it means that the resulting wave function will no longer be normalized.)

The problem (D.56) becomes

$$\sum_{\underline{\vec{n}} \neq \vec{n}} c_{\underline{\vec{n}},1}(E_{\underline{\vec{n}},0} - E_{\vec{n},0})\psi_{\underline{\vec{n}},0} = -H_1\psi_{\vec{n},0} + E_{\vec{n},1}\psi_{\vec{n},0}$$

where the left hand side was cleaned up using the fact that the $\psi_{\underline{\vec{n}},0}$ are eigenfunctions of H_0. To get the first order energy correction $E_{\vec{n},1}$, the trick is now to take an inner product of the entire equation with $\langle\psi_{\vec{n},0}|$. Because of the fact that the energy eigenfunctions of H_0 are orthonormal, this inner product produces zero in the left hand side, and in the right hand side it produces:

$$0 = -H_{\vec{n}\vec{n},1} + E_{\vec{n},1} \qquad H_{\vec{n}\vec{n},1} = \langle\psi_{\vec{n},0}|H_1\psi_{\vec{n},0}\rangle$$

And that is exactly the first order correction to the energy claimed in {A.38.1}; $E_{\vec{n},1}$ equals the Hamiltonian perturbation coefficient $H_{\vec{n}\vec{n},1}$. If the problem is not degenerate or $\psi_{\vec{n},0}$ is good, that is.

To get the coefficients $c_{\underline{\vec{n}},1}$, so that you know what is the first order correction $\psi_{\vec{n},1}$ to the wave function, just take an inner product with each of the other eigenfunctions $\langle\psi_{\underline{\vec{n}},0}|$ of H_0 in turn. In the left hand side it only leaves the coefficient of the selected eigenfunction because of orthonormality, and for the same reason, in the right hand side the final term drops out. The result is

$$c_{\underline{\vec{n}},1}(E_{\underline{\vec{n}},0} - E_{\vec{n},0}) = -H_{\underline{\vec{n}}\vec{n},1} \quad \text{for} \quad \underline{\vec{n}} \neq \vec{n} \qquad H_{\underline{\vec{n}}\vec{n},1} = -\langle\psi_{\underline{\vec{n}},0}|H_1\psi_{\vec{n},0}\rangle$$

The coefficients $c_{\underline{\vec{n}},1}$ can normally be computed from this.

Note however that if the problem is degenerate, there will be eigenfunctions $\psi_{\underline{\vec{n}},0}$ that have the same energy $E_{\vec{n},0}$ as the eigenfunction $\psi_{\vec{n},0}$ being corrected. For these the left hand side in the equation above is zero, and the equation cannot in general be satisfied. If so, it means that the assumption that an eigenfunction $\psi_{\vec{n}}$ of the full Hamiltonian expands in a power series in ε starting from $\psi_{\vec{n},0}$ is untrue. Eigenfunction $\psi_{\vec{n},0}$ is bad. And that means that the first order energy correction derived above is simply wrong. To fix the problem, what needs to be done is to identify the submatrix of all Hamiltonian perturbation coefficients in which both unperturbed eigenfunctions have the energy $E_{\vec{n},0}$, i.e. the submatrix

$$\text{all} \quad H_{\vec{n}_i\vec{n}_j,1} \quad \text{with} \quad E_{\vec{n}_i,0} = E_{\vec{n}_j,0} = E_{\vec{n},0}$$

The eigenvalues of this submatrix are the correct first order energy changes. So, if all you want is the first order energy changes, you can stop here. Otherwise, you need to replace the unperturbed eigenfunctions that have energy $E_{\vec{n},0}$. For each orthonormal eigenvector $(c_1, c_2, \ldots)$ of the submatrix, there is a corresponding replacement unperturbed eigenfunction

$$c_1\psi_{\vec{n}_1,0,\text{old}} + c_2\psi_{\vec{n}_2,0,\text{old}} + \ldots$$

You will need to rewrite the Hamiltonian perturbation coefficients in terms of these new eigenfunctions. (Since the replacement eigenfunctions are linear combinations of the old ones, no new integrations are needed.) You then need to reselect the eigenfunction $\psi_{\vec{n},0}$ whose energy to correct from among these replacement eigenfunctions. Choose the first order energy change (eigenvalue of the submatrix) $E_{\vec{n},1}$ that is of interest to you and then choose $\psi_{\vec{n},0}$ as the replacement eigenfunction corresponding to a corresponding eigenvector. If the first order energy change $E_{\vec{n},1}$ is not degenerate, the eigenvector is unique, so $\psi_{\vec{n},0}$ is now good. If not, the good eigenfunction will be some combination of the replacement eigenfunctions that have that first order energy change, and the good combination will have to be figured out later in the analysis. In any case, the problem with the equation above for the $c_{\underline{\vec{n}},1}$ will be fixed, because the new submatrix will be a diagonal one: $H_{\underline{\vec{n}}\vec{n},1}$ will be zero when $E_{\underline{\vec{n}},0} = E_{\vec{n},0}$ and $\underline{\vec{n}} \neq \vec{n}$. The coefficients $c_{\underline{\vec{n}},1}$ for which $E_{\underline{\vec{n}},0} = E_{\vec{n},0}$ remain indeterminate at this stage. They will normally be found at a later stage in the expansion.

With the coefficients $c_{\underline{\vec{n}},1}$ as found, or not found, the sum for the first order perturbation $\psi_{\vec{n},1}$ in the wave function becomes

$$\psi_{\vec{n},1} = -\sum_{E_{\underline{\vec{n}},0}\neq E_{\vec{n},0}} \frac{H_{\underline{\vec{n}}\vec{n},1}}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}}\psi_{\underline{\vec{n}},0} + \sum_{\substack{E_{\underline{\vec{n}},0}=E_{\vec{n},0} \\ \underline{\vec{n}}\neq\vec{n}}} c_{\underline{\vec{n}},1}\psi_{\underline{\vec{n}},0}$$

The entire process repeats for higher order. In particular, to second order

(D.57) gives, writing $\psi_{\vec{n},2}$ also in terms of the unperturbed eigenfunctions,

$$\sum_{\underline{\vec{n}}} c_{\underline{\vec{n}},2}(E_{\underline{\vec{n}},0} - E_{\vec{n},0})\psi_{\underline{\vec{n}},0} = \sum_{E_{\underline{\vec{n}},0}\neq E_{\vec{n},0}} \frac{H_{\underline{\vec{n}}\vec{n},1}}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} (H_1 - E_{\vec{n},1})\,\psi_{\underline{\vec{n}},0} - \sum_{\substack{E_{\underline{\vec{n}},0}=E_{\vec{n},0} \\ \underline{\vec{n}}\neq\vec{n}}} c_{\underline{\vec{n}},1}\,(H_1 - E_{\vec{n},1})\,\psi_{\underline{\vec{n}},0} + E_{\vec{n},2}\psi_{\vec{n},0}$$

To get the second order contribution to the energy, take again an inner product with $\langle\psi_{\vec{n},0}|$. That produces, again using orthonormality, (and diagonality of the submatrix discussed above if degenerate),

$$0 = \sum_{E_{\underline{\vec{n}},0}\neq E_{\vec{n},0}} \frac{H_{\underline{\vec{n}}\vec{n},1}H_{\vec{n}\underline{\vec{n}},1}}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} + E_{\vec{n},2}$$

This gives the second order change in the energy stated in {A.38.1}, if $\psi_{\vec{n},0}$ is good. Note that since H_1 is Hermitian, the product of the two Hamiltonian perturbation coefficients in the expression is just the square magnitude of either.

In the degenerate case, when taking an inner product with a $\langle\psi_{\underline{\vec{n}},0}|$ for which $E_{\underline{\vec{n}},0} = E_{\vec{n},0}$, the equation can be satisfied through the still indeterminate $c_{\underline{\vec{n}},1}$ provided that the corresponding diagonal coefficient $H_{\underline{\vec{n}}\underline{\vec{n}},1}$ of the diagonalized submatrix is unequal to $E_{\vec{n},1} = H_{\vec{n}\vec{n},1}$. In other words, provided that the first order energy change is not degenerate. If that is untrue, the higher order submatrix

$$\text{all} \sum_{E_{\underline{\vec{n}},0}\neq E_{\vec{n},0}} \frac{H_{\vec{n}_i\underline{\vec{n}},1}H_{\underline{\vec{n}}\vec{n}_j,1}}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} \quad \text{with} \quad E_{\vec{n}_i,0} = E_{\vec{n}_j,0} = E_{\vec{n},0} \quad E_{\vec{n}_i,1} = E_{\vec{n}_j,1} = E_{\vec{n},1}$$

will need to be diagonalized, (the rest of the equation needs to be zero). Its eigenvalues give the correct second order energy changes. To proceed to still higher energy, reselect the eigenfunctions following the same general lines as before. Obviously, in the degenerate case the entire process can become very messy. And you may never become sure about the good eigenfunction.

This problem can often be eliminated or greatly reduced if the eigenfunctions of H_0 are also eigenfunctions of another operator A, and H_1 commutes with A. Then you can arrange the eigenfunctions $\psi_{\underline{\vec{n}},0}$ into sets that have the same value for the "good" quantum number a of A. You can analyze the perturbed eigenfunctions in each of these sets while completely ignoring the existence of eigenfunctions with different values for quantum number a.

To see why, consider two example eigenfunctions ψ_1 and ψ_2 of A that have different eigenvalues a_1 and a_2. Since H_0 and H_1 both commute with A, their sum H does, so

$$0 = \langle\psi_2|(HA - AH)\psi_1\rangle = \langle\psi_2|HA\psi_1\rangle + \langle A\psi_2|H\psi_1\rangle = (a_1 - a_2)\langle\psi_2|H|\psi_1\rangle$$

and since $a_1 - a_2$ is not zero, $\langle\psi_2|H|\psi_1\rangle$ must be. Now $\langle\psi_2|H|\psi_1\rangle$ is the amount of eigenfunction ψ_2 produced by applying H on ψ_1. It follows that applying H on an eigenfunction with an eigenvalue a_1 does not produce any eigenfunctions with different eigenvalues a. Thus an eigenfunction of H satisfying

$$H\left(\sum_{a=a_1} c_{\vec{n}}\psi_{\vec{n},0} + \sum_{a\neq a_1} c_{\vec{n}}\psi_{\vec{n},0}\right) = E_{\vec{n}}\left(\sum_{a=a_1} c_{\vec{n}}\psi_{\vec{n},0} + \sum_{a\neq a_1} c_{\vec{n}}\psi_{\vec{n},0}\right)$$

can be replaced by just $\sum_{a=a_1} c_{\vec{n}}\psi_{\vec{n},0}$, since this by itself must satisfy the eigenvalue problem: the Hamiltonian of the second sum does not produce any amount of eigenfunctions in the first sum and vice-versa. (There must always be at least one value of a_1 for which the first sum at $\varepsilon = 0$ is independent of the other eigenfunctions of H.) Reduce every eigenfunction of H to an eigenfunction of A in this way. Now the existence of eigenfunctions with different values of a than the one being analyzed can be ignored since the Hamiltonian does not produce them. In terms of linear algebra, the Hamiltonian has been reduced to block diagonal form, with each block corresponding to a set of eigenfunctions with a single value of a. If the Hamiltonian also commutes with another operator B that the $\psi_{\vec{n},0}$ are eigenfunctions of, the argument repeats for the subsets with a single value for b.

The Hamiltonian perturbation coefficient $\langle\psi_2|H_1|\psi_1\rangle$ is zero whenever two good quantum numbers a_1 and a_2 are unequal. The reason is the same as for $\langle\psi_2|H|\psi_1\rangle$ above. Only perturbation coefficients for which all good quantum numbers are the same can be nonzero.

D.80 Hydrogen ground state Stark effect

This note derives the Stark effect on the hydrogen ground state. Since spin is irrelevant for the Stark effect, it will be ignored.

The unperturbed ground state of hydrogen was derived in chapter 4.3. Following the convention in perturbation theory to append a subscript zero to the unperturbed state, it can be summarized as:

$$H_0\psi_{100,0} = E_{100,0}\psi_{100,0} \qquad H_0 = -\frac{\hbar^2}{2m_{\rm e}}\nabla^2 + V \qquad \psi_{100,0} = \frac{1}{\sqrt{\pi a_0^3}}e^{-r/a_0}$$

where H_0 is the unperturbed hydrogen atom Hamiltonian, $\psi_{100,0}$ the unperturbed ground state wave function, $E_{100,0}$ the unperturbed ground state energy, 13.6 eV, and a_0 is the Bohr radius, 0.53 Å.

The Stark perturbation produces a change $\psi_{100,1}$ in this wave function that satisfies, from (A.243),

$$(H_0 - E_{100,0})\psi_{100,1} = -(H_1 - E_{100,1})\psi_{100,0} \quad H_1 = e\mathcal{E}_{\rm ext}z$$

The first order energy change $E_{100,1}$ is zero and can be dropped. The solution for $\psi_{100,1}$ will now simply be guessed to be $\psi_{100,0}$ times some spatial function f still to be found:

$$(H_0 - E_{100,0})\,(f\psi_{100,0}) = -e\mathcal{E}_{\rm ext} z\psi_{100,0} \qquad H_0 = -\frac{\hbar^2}{2m_{\rm e}}\nabla^2 + V$$

Differentiating out the Laplacian ∇^2 of the product $f\psi_{100,0}$ into individual terms using Cartesian coordinates, the equation becomes

$$f\,(H_0 - E_{100,0})\,\psi_{100,0} - \frac{\hbar^2}{m_{\rm e}}(\nabla f)\cdot(\nabla\psi_{100,0}) - \frac{\hbar^2}{2m_{\rm e}}(\nabla^2 f)\psi_{100,0} = -e\mathcal{E}_{\rm ext} z\psi_{100,0}$$

The first term in this equation is zero since $H_0\psi_{100,0} = E_{100,0}\psi_{100,0}$. Also, now using spherical coordinates, the gradients are, e.g. [41, 20.74, 20.82],

$$\nabla f = \frac{\partial f}{\partial r}\hat{\imath}_r + \frac{1}{r}\frac{\partial f}{\partial\theta}\hat{\imath}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial\phi}\hat{\imath}_\phi \qquad \nabla\psi_{100,0} = -\psi_{100,0}\frac{1}{a_0}\hat{\imath}_r$$

Substituting that into the equation, it reduces to

$$\frac{\hbar^2}{m_{\rm e}}\left(\frac{1}{a_0}\frac{\partial f}{\partial r} - \frac{1}{2}\nabla^2 f\right)\psi_{100,0} = -e\mathcal{E}_{\rm ext} z\psi_{100,0}$$

Now $z = r\cos\theta$ in polar coordinates, and for the r-derivative of f to produce something that is proportional to r, f must be proportional to r^2. (The Laplacian in the second term always produces lower powers of r than the r-derivative and can for now be ignored.) So, to balance the right hand side, f should contain a highest power of r equal to:

$$f = -\frac{m_{\rm e}e\mathcal{E}_{\rm ext}a_0}{2\hbar^2}r^2\cos\theta + \ldots$$

but then, using [41, 20.83], the $\nabla^2 f$ term in the left hand side produces an $e\mathcal{E}_{\rm ext}a_0\cos\theta$ term. So add another term to f for its r-derivative to eliminate it:

$$f = -\frac{m_{\rm e}e\mathcal{E}_{\rm ext}a_0}{2\hbar^2}r^2\cos\theta - \frac{m_{\rm e}e\mathcal{E}_{\rm ext}a_0^2}{\hbar^2}r\cos\theta$$

The Laplacian of $r\cos\theta = z$ is zero so no further terms need to be added. The change $f\psi_{100,0}$ in wave function is therefore

$$\psi_{100,1} = -\frac{m_{\rm e}e\mathcal{E}_{\rm ext}a_0}{2\hbar^2\sqrt{\pi a_0^3}}\left(r^2 + 2a_0 r\right)e^{-r/a_0}\cos\theta$$

(This "small perturbation" becomes larger than the unperturbed wave function far from the atom because of the growing value of r^2. It is implicitly assumed that the electric field terminates before a real problem arises. This is related

to the possibility of the electron tunneling out of the atom if the potential far from the atom is less than its energy in the atom: if the electron can tunnel out, there is strictly speaking no bound state.)

Now according to (A.243), the second order energy change can be found as

$$E_{100,2} = \langle \psi_{100,0} | H_1 \psi_{100,1} \rangle \qquad H_1 = e\mathcal{E}_{\rm ext} r \cos\theta$$

Doing the inner product integration in spherical coordinates produces

$$E_{100,2} = -\frac{9 m_{\rm e} e^2 \mathcal{E}_{\rm ext}^2 a_0^4}{4\hbar^2}$$

D.81 Dirac fine structure Hamiltonian

This note derives the fine structure Hamiltonian of the hydrogen atom. This Hamiltonian fixes up the main relativistic errors in the classical solution of chapter 4.3. The derivation is based on the relativistic Dirac equation from chapter 12.12 and uses nontrivial linear algebra.

According to the Dirac equation, the relativistic Hamiltonian and wave function take the form

$$H_D = m_{\rm e}c^2 \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \sum_{i=1}^{3} c\widehat{p}_i \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix} + V \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \vec{\psi}_D = \begin{pmatrix} \vec{\psi}^p \\ \vec{\psi}^n \end{pmatrix}$$

where $m_{\rm e}$ is the mass of the electron when at rest, c the speed of light, and the σ_i are the 2×2 Pauli spin matrices of chapter 12.10. Similarly the ones and zeros in the shown matrices are 2×2 unit and zero matrices. The wave function is a four-dimensional vector whose components depend on spatial position. It can be subdivided into the two-dimensional vectors $\vec{\psi}^p$ and $\vec{\psi}^n$. The two components of $\vec{\psi}^p$ correspond to the spin up and spin down components of the normal classical electron wave function; as noted in chapter 5.5.1, this can be thought of as a vector if you want. The two components of the other vector $\vec{\psi}^n$ are very small for the solutions of interest. These components would be dominant for states that would have negative rest mass. They are associated with the anti-particle of the electron, the positron.

The Dirac equation is solvable in closed form, but that solution is not something you want to contemplate if you can avoid it. And there is really no need for it, since the Dirac equation is not exact anyway. To the accuracy it has, it can easily be solved using perturbation theory in essentially the same way as in derivation {D.79}. In this case, the small parameter is $1/c$: if the speed of light is infinite, the nonrelativistic solution is exact. And if you ballpark a typical velocity for the electron in a hydrogen atom, it is only about one percent or so of the speed of light.

So, following derivation {D.79}, take the Hamiltonian apart into successive powers of $1/c$ as $H_D = H_{D,0} + H_{D,1} + H_{D,2}$ with

$$H_{D,0} = \begin{pmatrix} m_{\rm e}c^2 & 0 \\ 0 & -m_{\rm e}c^2 \end{pmatrix} \quad H_{D,1} = \sum_{i=1}^{3} \begin{pmatrix} 0 & c\widehat{p}_i\sigma_i \\ c\widehat{p}_i\sigma_i & 0 \end{pmatrix} \quad H_{D,2} = \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix}$$

and similarly for the wave function vector:

$$\vec{\psi}_D = \begin{pmatrix} \vec{\psi}_0^{\,p} \\ \vec{\psi}_0^{\,n} \end{pmatrix} + \begin{pmatrix} \vec{\psi}_1^{\,p} \\ \vec{\psi}_1^{\,n} \end{pmatrix} + \begin{pmatrix} \vec{\psi}_2^{\,p} \\ \vec{\psi}_2^{\,n} \end{pmatrix} + \begin{pmatrix} \vec{\psi}_3^{\,p} \\ \vec{\psi}_3^{\,n} \end{pmatrix} + \begin{pmatrix} \vec{\psi}_4^{\,p} \\ \vec{\psi}_4^{\,n} \end{pmatrix} + \ldots$$

and the energy:

$$E_D = E_{D,0} + E_{D,1} + E_{D,2} + E_{D,3} + E_{D,4} + \ldots$$

Substitution into the Hamiltonian eigenvalue problem $H_D\vec{\psi}_D = E_D\vec{\psi}_D$ and then collecting equal powers of $1/c$ together produces again a system of successive equations, just like in derivation {D.79}:

$$c^2: \quad \left[\begin{pmatrix} m_{\rm e}c^2 & 0 \\ 0 & -m_{\rm e}c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_0^{\,p} \\ \vec{\psi}_0^{\,n} \end{pmatrix} = 0$$

$$c^1: \quad \left[\begin{pmatrix} m_{\rm e}c^2 & 0 \\ 0 & -m_{\rm e}c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_1^{\,p} \\ \vec{\psi}_1^{\,n} \end{pmatrix} =$$
$$- \left[\sum_{i=1}^{3} \begin{pmatrix} 0 & c\widehat{p}_i\sigma_i \\ c\widehat{p}_i\sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_0^{\,p} \\ \vec{\psi}_0^{\,n} \end{pmatrix}$$

$$c^0: \quad \left[\begin{pmatrix} m_{\rm e}c^2 & 0 \\ 0 & -m_{\rm e}c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_2^{\,p} \\ \vec{\psi}_2^{\,n} \end{pmatrix} =$$
$$- \left[\sum_{i=1}^{3} \begin{pmatrix} 0 & c\widehat{p}_i\sigma_i \\ c\widehat{p}_i\sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_1^{\,p} \\ \vec{\psi}_1^{\,n} \end{pmatrix}$$
$$- \left[\begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} - \begin{pmatrix} E_{D,2} & 0 \\ 0 & E_{D,2} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_0^{\,p} \\ \vec{\psi}_0^{\,n} \end{pmatrix}$$

$$c^{-1}: \quad \left[\begin{pmatrix} m_{\rm e}c^2 & 0 \\ 0 & -m_{\rm e}c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix}\right] \begin{pmatrix} \vec{\psi}_3^{\,p} \\ \vec{\psi}_3^{\,n} \end{pmatrix} =$$

$$
\begin{aligned}
&- \left[\sum_{i=1}^{3} \begin{pmatrix} 0 & c\widehat{p}_i\sigma_i \\ c\widehat{p}_i\sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_2^{\,p} \\ \vec{\psi}_2^{\,n} \end{pmatrix} \\
&- \left[\begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} - \begin{pmatrix} E_{D,2} & 0 \\ 0 & E_{D,2} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_1^{\,p} \\ \vec{\psi}_1^{\,n} \end{pmatrix} \\
&+ \begin{pmatrix} E_{D,3} & 0 \\ 0 & E_{D,3} \end{pmatrix} \begin{pmatrix} \vec{\psi}_0^{\,p} \\ \vec{\psi}_0^{\,n} \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
c^{-2}: \quad & \left[\begin{pmatrix} m_{\mathrm{e}}c^2 & 0 \\ 0 & -m_{\mathrm{e}}c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_4^{\,p} \\ \vec{\psi}_4^{\,n} \end{pmatrix} = \\
&- \left[\sum_{i=1}^{3} \begin{pmatrix} 0 & c\widehat{p}_i\sigma_i \\ c\widehat{p}_i\sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_3^{\,p} \\ \vec{\psi}_3^{\,n} \end{pmatrix} \\
&- \left[\begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} - \begin{pmatrix} E_{D,2} & 0 \\ 0 & E_{D,2} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_2^{\,p} \\ \vec{\psi}_2^{\,n} \end{pmatrix} \\
&+ \begin{pmatrix} E_{D,3} & 0 \\ 0 & E_{D,3} \end{pmatrix} \begin{pmatrix} \vec{\psi}_1^{\,p} \\ \vec{\psi}_1^{\,n} \end{pmatrix} + \begin{pmatrix} E_{D,4} & 0 \\ 0 & E_{D,4} \end{pmatrix} \begin{pmatrix} \vec{\psi}_0^{\,p} \\ \vec{\psi}_0^{\,n} \end{pmatrix}
\end{aligned}
$$

$$
c^{-3}: \quad \cdots
$$

The first, order c^2, eigenvalue problem has energy eigenvalues $\pm m_e c^2$, in other words, plus or minus the rest mass energy of the electron. The solution of interest is the physical one with a positive rest mass, so the desired solution is

$$
E_{D,0} = m_{\mathrm{e}}c^2 \qquad \vec{\psi}_0^{\,p} = \text{still arbitrary} \qquad \vec{\psi}_0^{\,n} = 0
$$

Plug that into the order c^1 equation to give, for top and bottom subvectors

$$
0 = E_{D,1}\vec{\psi}_0^{\,p} \qquad -2m_{\mathrm{e}}c^2\vec{\psi}_1^{\,n} = -\sum_i c\widehat{p}_i\sigma_i\vec{\psi}_0^{\,p}
$$

It follows from the first of those that the first order energy change must be zero because $\vec{\psi}_0^{\,p}$ cannot be zero; otherwise there would be nothing left. The second equation gives the leading order values of the secondary components, so in total

$$
E_{D,1} = 0 \qquad \vec{\psi}_1^{\,p} = \text{still arbitrary} \qquad \vec{\psi}_1^{\,n} = \sum_j \frac{1}{2m_{\mathrm{e}}c}\widehat{p}_j\sigma_j\vec{\psi}_0^{\,p}
$$

where the summation index i was renamed to j to avoid ambiguity later.

Plug all that in the order c^0 equation to give

$$0 = -\frac{1}{2m_{\rm e}} \sum_i \sum_j \widehat{p}_i \widehat{p}_j \sigma_i \sigma_j \vec{\psi}_0^p - V \vec{\psi}_0^p + E_{D,2} \vec{\psi}_0^p \qquad \vec{\psi}_2^n = \sum_j \frac{1}{2m_{\rm e}c} \widehat{p}_j \sigma_j \vec{\psi}_1^p$$

The first of these two equations is the nonrelativistic Hamiltonian eigenvalue problem of chapter 4.3. To see that, note that in the double sum the terms with $j \neq i$ pairwise cancel since for the Pauli matrices, $\sigma_i\sigma_j + \sigma_j\sigma_i = 0$ when $j \neq i$. For the remaining terms in which $j = i$, the relevant property of the Pauli matrices is that $\sigma_i\sigma_i$ is one (or the 2×2 unit matrix, really,) giving

$$\frac{1}{2m_{\rm e}} \sum_i \sum_j \widehat{p}_i \widehat{p}_j \sigma_i \sigma_j + V = \frac{1}{2m_{\rm e}} \sum_i \widehat{p}_i^2 + V \equiv H_0$$

where H_0 is the nonrelativistic hydrogen Hamiltonian of chapter 4.3.

So the first part of the order c^0 equation takes the form

$$H_0 \vec{\psi}_0^p = E_{D,2} \vec{\psi}_0^p$$

The energy $E_{D,2}$ will therefore have to be a Bohr energy level E_n and each component of $\vec{\psi}_0^p$ will have to be a nonrelativistic energy eigenfunction with that energy:

$$E_{D,2} = E_n \qquad \vec{\psi}_0^p = \sum_l \sum_m c_{lm+} \psi_{nlm}\uparrow + \sum_l \sum_m c_{lm-} \psi_{nlm}\downarrow$$

The sum multiplying $\uparrow$ is the first component of vector $\vec{\psi}_0^p$ and the sum multiplying $\downarrow$ the second. The nonrelativistic analysis in chapter 4.3 was indeed correct as long as the speed of light is so large compared to the relevant velocities that $1/c$ can be ignored.

To find out the error in it, the relativistic expansion must be taken to higher order. To order c^{-1}, you get for the top vector

$$0 = -(H_0 - E_n)\vec{\psi}_1^p + E_{D,3} \vec{\psi}_0^p$$

Now if $\vec{\psi}_1^p$ is written as a sum of the eigenfunctions of H_0, including $\vec{\psi}_0^p$, the first term will produce zero times $\vec{\psi}_0^p$ since $(H_0 - E_n)\vec{\psi}_0^p = 0$. That means that $E_{D,3}$ must be zero. The expansion must be taken one step further to identify the relativistic energy change. The bottom vector gives

$$\vec{\psi}_3^n = \sum_j \frac{1}{2m_{\rm e}c} \widehat{p}_j \sigma_j \vec{\psi}_2^p + \frac{V - E_n}{2m_{\rm e}c^2} \sum_j \frac{1}{2m_{\rm e}c} \widehat{p}_j \sigma_j \vec{\psi}_0^p$$

To order c^{-2}, you get for the top vector

$$0 = -(H_0 - E_n)\vec{\psi}_2^p - \sum_i \sum_j \widehat{p}_i \sigma_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_j \sigma_j \vec{\psi}_0^p + E_{D,4} \vec{\psi}_0^p$$

and that determines the approximate relativistic energy correction.

Now recall from derivation {D.79} that if you do a nonrelativistic expansion of an eigenvalue problem $(H_0 + H_1)\psi = E\psi$, the equations to solve are (D.55) and (D.56);

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},0} = 0 \qquad (H_0 - E_{\vec{n},0})\psi_{\vec{n},1} = -(H_1 - E_{\vec{n},1})\psi_{\vec{n},0}$$

The first equation was satisfied by the solution for $\vec{\psi}_0^p$ obtained above. However, the second equation presents a problem. Comparison with the final Dirac result suggests that the fine structure Hamiltonian correction H_1 should be identified as

$$H_1 \stackrel{?}{=} \sum_i \sum_j \widehat{p}_i \sigma_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_j \sigma_j$$

but that is not right, since E_n is not a physical operator, but an energy eigenvalue for the selected eigenfunction. So mapping the Dirac expansion straightforwardly onto a classical one has run into a snag.

It is maybe not that surprising that a two-dimensional wave function cannot correctly represent a truly four-dimensional one. But clearly, whatever is selected for the fine structure Hamiltonian H_1 must at least get the energy eigenvalues right. To see how this can be done, the operator obtained from the Dirac equation will have to be simplified. Now for any given i, the sum over j includes a term $j = i$, a term $j = \bar{\imath}$, where $\bar{\imath}$ is the number following i in the cyclic sequence $\ldots 123123 \ldots$, and it involves a term $j = \bar{\bar{\imath}}$ where $\bar{\bar{\imath}}$ precedes i in the sequence. So the Dirac operator falls apart into three pieces:

$$H_1 \stackrel{?}{=} \sum_i \widehat{p}_i \sigma_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_i \sigma_i + \sum_i \widehat{p}_i \sigma_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_{\bar{\imath}} \sigma_{\bar{\imath}} + \sum_i \widehat{p}_i \sigma_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_{\bar{\bar{\imath}}} \sigma_{\bar{\bar{\imath}}}$$

or using the properties of the Pauli matrices that $\sigma_i \sigma_i = 1$, $\sigma_i \sigma_{\bar{\imath}} = {\rm i}\sigma_{\bar{\bar{\imath}}}$, and $\sigma_i \sigma_{\bar{\bar{\imath}}} = -{\rm i}\sigma_{\bar{\imath}}$ for any i,

$$H_1 \stackrel{?}{=} \sum_i \widehat{p}_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_i + {\rm i} \sum_i \widehat{p}_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_{\bar{\imath}} \sigma_{\bar{\bar{\imath}}} - {\rm i} \sum_i \widehat{p}_i \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_{\bar{\bar{\imath}}} \sigma_{\bar{\imath}} \tag{D.59}$$

The approach will now be to show first that the final two terms are the spin-orbit interaction in the fine structure Hamiltonian. After that, the much more tricky first term will be discussed. Renotate the indices in the last two terms as follows:

$$H_{1,\text{spin-orbit}} = {\rm i} \sum_i \widehat{p}_{\bar{\imath}} \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_{\bar{\bar{\imath}}} \sigma_i - {\rm i} \sum_i \widehat{p}_{\bar{\bar{\imath}}} \frac{V - E_n}{4m_{\rm e}^2 c^2} \widehat{p}_{\bar{\imath}} \sigma_i$$

Since the relative order of the subscripts in the cycle was maintained in the renotation, the sums still contain the exact same three terms, just in a different

order. Take out the common factors;

$$H_{1,\text{spin-orbit}} = \frac{\mathrm{i}}{4m_\mathrm{e}^2c^2}\sum_i [\widehat{p}_{\bar{\imath}}(V-E_n)\widehat{p}_{\bar{\bar{\imath}}} - \widehat{p}_{\bar{\bar{\imath}}}(V-E_n)\widehat{p}_{\bar{\imath}}]\,\sigma_i$$

Now according to the generalized canonical commutator of chapter 4.5.4:

$$\widehat{p}_i(V-E_n) = (V-E_n)\widehat{p}_i - \mathrm{i}\hbar\frac{\partial(V-E_n)}{\partial r_i}$$

where E_n is a constant that produces a zero derivative. So $\widehat{p}_{\bar{\imath}}$, respectively $\widehat{p}_{\bar{\bar{\imath}}}$ can be taken to the other side of $V-E_n$ as long as the appropriate derivatives of V are added. If that is done, $(V-E_n)\widehat{p}_{\bar{\imath}}\widehat{p}_{\bar{\bar{\imath}}}$ and $-(V-E_n)\widehat{p}_{\bar{\bar{\imath}}}\widehat{p}_{\bar{\imath}}$ cancel since linear momentum operators commute. What is left are just the added derivative terms:

$$H_{1,\text{spin-orbit}} = \frac{\hbar}{4m_\mathrm{e}^2c^2}\sum_i \left[\frac{\partial V}{\partial r_{\bar{\imath}}}\widehat{p}_{\bar{\bar{\imath}}} - \frac{\partial V}{\partial r_{\bar{\bar{\imath}}}}\widehat{p}_{\bar{\imath}}\right]\sigma_i$$

Note that the errant eigenvalue E_n mercifully dropped out. Now the hydrogen potential V only depends on the distance r from the origin, as $1/r$, so

$$\frac{\partial V}{\partial r_i} = -\frac{V}{r^2}r_i$$

and plugging that into the operator, you get

$$H_{1,\text{spin-orbit}} = -\frac{\hbar V}{4m_\mathrm{e}^2c^2r^2}\sum_i [r_{\bar{\imath}}\widehat{p}_{\bar{\bar{\imath}}} - r_{\bar{\bar{\imath}}}\widehat{p}_{\bar{\imath}}]\,\sigma_i$$

The term between the square brackets can be recognized as the ith component of the angular momentum operator; also the Pauli spin matrix σ_i is defined as $\widehat{S}_i/\frac{1}{2}\hbar$, so

$$H_{1,\text{spin-orbit}} = -\frac{V}{2m_\mathrm{e}^2c^2r^2}\sum_i \widehat{L}_i\widehat{S}_i$$

Get rid of c^2 using $|E_1| = \frac{1}{2}\alpha^2 m_\mathrm{e}c^2$, of V using $V = -2|E_1|a_0/r$, and m_e using $|E_1| = \hbar^2/2m_\mathrm{e}a_0^2$ to get the spin-orbit interaction as claimed in the section on fine structure.

That leaves the term

$$\sum_i \widehat{p}_i\frac{V-E_n}{4m_\mathrm{e}^2c^2}\widehat{p}_i$$

in (D.59). Since $V = H_0 - \widehat{p}^2/2m_\mathrm{e}$, it can be written as

$$\sum_i \widehat{p}_i\frac{H_0-E_n}{4m_\mathrm{e}^2c^2}\widehat{p}_i - \frac{(\widehat{p}^2)^2}{8m_\mathrm{e}^3c^2}$$

The final term is the claimed Einstein correction in the fine structure Hamiltonian, using $|E_1| = \frac{1}{2}\alpha^2 m_e c^2$ to get rid of c^2.

The first term,

$$H_{1,\text{Darwin}} \stackrel{?}{=} \sum_i \widehat{p}_i \frac{H_0 - E_n}{4m_e^2 c^2} \widehat{p}_i$$

is the sole remaining problem. It cannot be transformed into a decent physical operator. The objective is just to get the energy correction right. And to achieve that requires only that the Hamiltonian perturbation coefficients are evaluated correctly at the E_n energy level. Specifically, what is needed is that

$$H_{\underline{\vec{n}}\vec{n},1,\text{Darwin}} \equiv \langle \psi_{\underline{\vec{n}},0} | H_{1,\text{Darwin}} \psi_{\vec{n},0} \rangle = \frac{1}{4m_e^2 c^2} \sum_i \langle \psi_{\underline{\vec{n}},0} | \widehat{p}_i (H_0 - E_n) \widehat{p}_i \psi_{\vec{n},0} \rangle$$

for any arbitrary pair of unperturbed hydrogen energy eigenfunctions $\psi_{\underline{\vec{n}},0}$ and $\psi_{\vec{n},0}$ with energy E_n. To see what that means, the leading Hermitian operator $\widehat{p}_i$ can be taken to the other side of the inner product, and in half of that result, $H_0 - E_n$ will also be taken to the other side:

$$H_{\underline{\vec{n}}\vec{n},1,\text{Darwin}} = \frac{1}{8m_e^2 c^2} \sum_i \left(\langle \widehat{p}_i \psi_{\underline{\vec{n}},0} | (H_0 - E_n) \widehat{p}_i \psi_{\vec{n},0} \rangle + \langle (H_0 - E_n) \widehat{p}_i \psi_{\underline{\vec{n}},0} | \widehat{p}_i \psi_{\vec{n},0} \rangle \right)$$

Now if you simply swap the order of the factors in $(H_0 - E_n)\widehat{p}_i$ in this expression, you get zero, because both eigenfunctions have energy E_n. However, swapping the order of $(H_0 - E_n)\widehat{p}_i$ brings in the generalized canonical commutator $[V, \widehat{p}_i]$ that equals $i\hbar \partial V / \partial r_i$. Therefore, writing out the remaining inner product you get

$$H_{\underline{\vec{n}}\vec{n},1,\text{Darwin}} = \frac{-\hbar^2}{8m_e^2 c^2} \sum_i \int_{\text{all } \vec{r}} \frac{\partial V}{\partial r_i} \frac{\partial \psi^*_{\underline{\vec{n}},0} \psi_{\vec{n},0}}{\partial r_i} \, \mathrm{d}^3 \vec{r}$$

Now, the potential V becomes infinite at $r = 0$, and that makes mathematical manipulation difficult. Therefore, assume for now that the nuclear charge e is not a point charge, but spread out over a very small region around the origin. In that case, the inner product can be rewritten as

$$H_{\underline{\vec{n}}\vec{n},1,\text{Darwin}} = \frac{-\hbar^2}{8m_e^2 c^2} \sum_i \int_{\text{all } \vec{r}} \left[\frac{\partial}{\partial r_i} \left(\frac{\partial V}{\partial r_i} \psi^*_{\underline{\vec{n}},0} \psi_{\vec{n},0} \right) - \frac{\partial^2 V}{\partial r_i^2} \psi^*_{\underline{\vec{n}},0} \psi_{\vec{n},0} \right] \mathrm{d}^3 \vec{r}$$

and the first term integrates away since $\psi^*_{\underline{\vec{n}},0} \psi_{\vec{n},0}$ vanishes at infinity. In the final term, use the fact that the derivatives of the potential energy V give e times the electric field of the nucleus, and therefore the second order derivatives give e times the divergence of the electric field. Maxwell's first equation (13.5) says that that is e/ϵ_0 times the nuclear charge density. Now if the region of nuclear charge is allowed to contract back to a point, the charge density must still integrate to the net proton charge e, so the charge density becomes $e\delta^3(\vec{r})$

where $\delta^3(\vec{r})$ is the three-dimensional delta function. Therefore the Darwin term produces Hamiltonian perturbation coefficients as if its Hamiltonian is

$$H_{1,\text{Darwin}} = \frac{\hbar^2 e^2}{8 m_e^2 c^2 \epsilon_0} \delta^3(\vec{r})$$

Get rid of c^2 using $|E_1| = \frac{1}{2}\alpha^2 m_e c^2$, of e^2/ϵ_0 using $e^2/4\pi\epsilon_0 = 2|E_1|a_0$, and m_e using $|E_1| = \hbar^2/2m_e a_0^2$ to get the Darwin term as claimed in the section on fine structure. It will give the right energy correction for the nonrelativistic solution. But you may rightly wonder what to make of the implied wave function.

D.82 Classical spin-orbit derivation

This note derives the spin-orbit Hamiltonian from a more intuitive, classical point of view than the Dirac equation mathematics.

Picture the magnetic electron as containing a pair of positive and negative magnetic monopoles of a large strength q_m. The very small distance from negative to positive pole is denoted by $\vec{d}$ and the product $\vec{\mu} = q_m \vec{d}$ is the magnetic dipole strength, which is finite.

Next imagine this electron smeared out in some orbit encircling the nucleus with a speed $\vec{v}$. The two poles will then be smeared out into two parallel "magnetic currents" that are very close together. The two currents have opposite directions because the velocity $\vec{v}$ of the poles is the same while their charges are opposite. These magnetic currents will be encircled by electric field lines just like the electric currents in figure 13.15 were encircled by magnetic field lines.

Now assume that seen from up very close, a segment of these currents will seem almost straight and two-dimensional, so that two-dimensional analysis can be used. Take a local coordinate system such that the z-axis is aligned with the negative magnetic current and in the direction of positive velocity. Rotate the xy-plane around the z-axis so that the positive current is to the right of the negative one. The picture is then just like figure 13.15, except that the currents are magnetic and the field lines electric. In this coordinate system, the vector from negative to positive pole takes the form $\vec{d} = d_x \hat{\imath} + d_z \hat{k}$.

The magnetic current strength is defined as $q'_m v$, where q'_m is the moving magnetic charge per unit length of the current. So, according to table 13.2 the negative current along the z-axis generates a two-dimensional electric field whose potential is

$$\varphi_\ominus = -\frac{q'_m v}{2\pi\epsilon_0 c^2}\theta = -\frac{q'_m v}{2\pi\epsilon_0 c^2} \arctan\left(\frac{y}{x}\right)$$

To get the field of the positive current a distance d_x to the right of it, shift x and change sign:

$$\varphi_\oplus = \frac{q'_m v}{2\pi\epsilon_0 c^2} \arctan\left(\frac{y}{x - d_x}\right)$$

If these two potentials are added, the difference between the two arctan functions can be approximated as $-d_x$ times the x derivative of the unshifted arctan. That can be seen from either recalling the very definition of the partial derivative, or from expanding the second arctan in a Taylor series in x. The bottom line is that the monopoles of the moving electron generate a net electric field with a potential

$$\varphi = \frac{q'_m d_x v}{2\pi\epsilon_0 c^2} \frac{y}{x^2+y^2}$$

Now compare that with the electric field generated by a couple of opposite electric line charges like in figure 13.12, a negative one along the z-axis and a positive one above it at a position $y = d_c$. The electric dipole moment per unit length of such a pair of line charges is by definition $\vec{\wp}' = q' d_c \, \hat{\jmath}$, where q' is the electric charge per unit length. According to table 13.1, a single electric charge along the z-axis creates an electric field whose potential is

$$\varphi = \frac{q'}{2\pi\epsilon_0} \ln \frac{1}{r} = -\frac{q'}{4\pi\epsilon_0} \ln \left(x^2+y^2\right)$$

For an electric dipole consisting of a negative line charge along the z-axis and a positive one above it at $y = d_c$, the field is then

$$\varphi = -\frac{q'}{4\pi\epsilon_0} \ln \left(x^2+(y-d)^2\right) + \frac{q'}{4\pi\epsilon_0} \ln \left(x^2+y^2\right)$$

and the difference between the two logarithms can be approximated as $-d_c$ times the y-derivative of the unshifted one. That gives

$$\varphi = \frac{q' d_c}{2\pi\epsilon_0} \frac{y}{x^2+y^2}$$

Comparing this with the potential of the monopoles, it is seen that the magnetic currents create an electric dipole in the y-direction whose strength $\vec{\wp}'$ is $q'_m d_x v/c^2 \, \hat{\jmath}$. And since in this coordinate system the magnetic dipole moment is $\vec{\mu}\,' = q'_m(d_x \hat{\imath} + d_z \hat{k})$ and the velocity $v\hat{k}$, it follows that the generated electric dipole strength is

$$\vec{\wp}' = -\vec{\mu}\,' \times \vec{v}/c^2$$

Since both dipole moments are per unit length, the same relation applies between the actual magnetic dipole strength of the electron and the electric dipole strength generated by its motion. The primes can be omitted.

Now the energy of the electric dipole is $-\vec{\wp} \cdot \vec{\mathcal{E}}$ where $\vec{\mathcal{E}}$ is the electric field of the nucleus, $e\vec{r}/4\pi\epsilon_0 r^3$ according to table 13.1. So the energy is:

$$\frac{e}{4\pi\epsilon_0 c^2} \frac{1}{r^3} \vec{r} \cdot (\vec{\mu} \times \vec{v})$$

and the order of the triple product of vectors can be changed and then the angular momentum can be substituted:

$$-\frac{e}{4\pi\epsilon_0 c^2}\frac{1}{r^3}\vec{\mu}\cdot(\vec{r}\times\vec{v}) = -\frac{e}{4\pi\epsilon_0 c^2 m_{\rm e}}\frac{1}{r^3}\vec{\mu}\cdot\vec{L}$$

To get the correct spin-orbit interaction, the magnetic dipole moment $\vec{\mu}$ used in this expression must be the classical one, $-e\vec{S}/2m_{\rm e}$. The additional factor $g_e = 2$ for the energy of the electron in a magnetic field does not apply here. There does not seem to be a really good reason to give for that, except for saying that the same Dirac equation that says that the additional g-factor is there in the magnetic interaction also says it is not in the spin-orbit interaction. The expression for the energy becomes

$$\frac{e^2}{8\pi\epsilon_0 m_{\rm e}^2 c^2}\frac{1}{r^3}\vec{S}\cdot\vec{L}$$

Getting rid of c^2 using $|E_1| = \frac{1}{2}\alpha^2 m_{\rm e}c^2$, of e^2/ϵ_0 using $e^2/4\pi\epsilon_0 = 2|E_1|a_0$, and of $m_{\rm e}$ using $|E_1| = \hbar^2/2m_{\rm e}a_0^2$, the claimed expression for the spin-orbit energy is found.

D.83 Expectation powers of r for hydrogen

This note derives the expectation values of the powers of r for the hydrogen energy eigenfunctions ψ_{nlm}. The various values to be be derived are:

$$\begin{aligned}
&\ldots\\
&\langle\psi_{nlm}|(a_0/r)^3\psi_{nlm}\rangle = \frac{1}{l(l+\frac{1}{2})(l+1)n^3}\\
&\langle\psi_{nlm}|(a_0/r)^2\psi_{nlm}\rangle = \frac{1}{(l+\frac{1}{2})n^3}\\
&\langle\psi_{nlm}|(a_0/r)\psi_{nlm}\rangle = \frac{1}{n^2}\\
&\langle\psi_{nlm}|1\psi_{nlm}\rangle = 1\\
&\langle\psi_{nlm}|(r/a_0)\psi_{nlm}\rangle = \frac{3n^2-l(l+1)}{2}\\
&\langle\psi_{nlm}|(r/a_0)^2\psi_{nlm}\rangle = \frac{n^2(5n^2-3l(l+1)+1)}{2}\\
&\ldots
\end{aligned} \tag{D.60}$$

where a_0 is the Bohr radius, about 0.53 Å. Note that you can get the expectation value of a more general function of r by summing terms, provided that the

function can be expanded into a Laurent series. Also note that the value of m does not make a difference: you can combine ψ_{nlm} of different m values together and it does not change the above expectation values. And watch it, when the power of r becomes too negative, the expectation value will cease to exist. For example, for $l = 0$ the expectation values of $(a_0/r)^3$ and higher powers are infinite.

The trickiest to derive is the expectation value of $(a_0/r)^2$, and that one will be done first. First recall the hydrogen Hamiltonian from chapter 4.3,

$$H = -\frac{\hbar^2}{2m_{\rm e}r^2}\left\{\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right\} - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r}$$

Its energy eigenfunctions of given square and z angular momentum and their energy are

$$\psi_{nlm} = R_{nl}(r)Y_l^m(\theta,\phi) \qquad E_n = -\frac{\hbar^2}{2n^2m_{\rm e}a_0^2} \qquad a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_{\rm e}e^2}$$

where the Y_l^m are called the spherical harmonics.

When this Hamiltonian is applied to an eigenfunction ψ_{nlm}, it produces the exact same result as the following "dirty trick Hamiltonian" in which the angular derivatives have been replaced by $l(l+1)$:

$$H_{\rm DT} = -\frac{\hbar^2}{2m_{\rm e}r^2}\left\{\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) - l(l+1)\right\} - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r}$$

The reason is that the angular derivatives are essentially the square angular momentum operator of chapter 4.2.3. Now, while in the hydrogen Hamiltonian the quantum number l has to be an integer because of its origin, in the dirty trick one l can be allowed to assume any value. That means that you can differentiate the Hamiltonian and its eigenvalues E_n with respect to l. And that allows you to apply the Hellmann-Feynman theorem of section A.38.1:

$$\frac{\partial E_{n,\rm DT}}{\partial l} = \left\langle \psi_{nlm} \middle| \frac{\partial H_{\rm DT}}{\partial l}\psi_{nlm} \right\rangle$$

(Yes, the eigenfunctions ψ_{nlm} are good, because the purely radial $H_{\rm DT}$ commutes with both $\widehat{L}_z$ and $\widehat{L}^2$, which are angular derivatives.) Substituting in the dirty trick Hamiltonian,

$$\frac{\partial E_{n,\rm DT}}{\partial l} = \frac{\hbar^2(2l+1)}{2m_{\rm e}a_0^2}\left\langle \psi_{nlm} \middle| \left(\frac{a_0}{r}\right)^2 \psi_{nlm} \right\rangle$$

So, if you can figure out how the dirty trick energy changes with l near some desired integer value $l = l_0$, the desired expectation value of $(a_0/r)^2$ at that integer value of l follows. Note that the eigenfunctions of $H_{\rm DT}$ can still be taken

to be of the form $R_{nl}(r)Y_{l_0}^m(\theta,\phi)$, where $Y_{l_0}^m$ can be divided out of the eigenvalue problem to give $H_{\rm DT}R_{nl} = E_{DT}R_{nl}$. If you skim back through chapter 4.3 and its note, you see that that eigenvalue problem was solved in derivation {D.15}. Now, of course, l is no longer an integer, but if you skim through the note, it really makes almost no difference. The energy eigenvalues are still $E_{n,\rm DT} = -\hbar^2/2n^2m_{\rm e}a_0^2$. If you look near the end of the note, you see that the requirement on n is that $n = q+l+1$ where q must remain an integer for valid solutions, hence must stay constant under small changes. So $\mathrm{d}n/\mathrm{d}l = 1$, and then according to the chain rule the derivative of $E_{\rm DT}$ is $\hbar^2/n^3m_{\rm e}a_0^2$. Substitute it in and there you have that nasty expectation value as given in (D.60).

All other expectation values of $(r/a_0)^q$ for integer values of q may be found from the "Kramers relation," or "(second) Pasternack relation:"

$$4(q+1)\langle q\rangle - 4n^2(2q+1)\langle q-1\rangle + n^2q[(2l+1)^2-q^2]\langle q-2\rangle = 0 \qquad \text{(D.61)}$$

where $\langle q\rangle$ is shorthand for the expectation value $\langle\psi_{nlm}|(r/a_0)^q\psi_{nlm}\rangle$.

Substituting $q = 0$ into the Kramers-Pasternack relation produces the expectation value of a_0/r as in (D.60). It may be noted that this can instead be derived from the virial theorem of chapter 7.2, or from the Hellmann-Feynman theorem by differentiating the hydrogen Hamiltonian with respect to the charge e. Substituting in $q = 1, 2, \ldots$ produces the expectation values for r/a_0, $(r/a_0)^2$, ... Substituting in $q = -1$ *and* the expectation value for $(a_0/r)^2$ from the Hellmann-Feynman theorem gives the expectation value for $(a_0/r)^3$. The remaining negative integer values $q = -2, -3, \ldots$ produce the remaining expectation values for the negative integer powers of r/a_0 as the $\langle q-2\rangle$ term in the equation.

Note that for a sufficiently negative powers of r, the expectation value becomes infinite. Specifically, since ψ_{nlm} is proportional to r^l, {D.15}, it can be seen that $\langle q-2\rangle$ becomes infinite when $q = -2l-1$. When that happens, the coefficient of the expectation value in the Kramers-Pasternack relation becomes zero, making it impossible to compute the expectation value. The relationship can be used until it crashes and then the remaining expectation values are all infinite.

The remainder of this note derives the Kramers-Pasternack relation. First note that the expectation values are defined as

$$\langle q\rangle \equiv \langle\psi_{nlm}|(r/a_0)^q\psi_{nlm}\rangle = \int_{\text{all }\vec{r}} (r/a_0)^q|\psi_{nlm}|^2\,\mathrm{d}^3\vec{r} = \int_{\text{all }\vec{r}} (r/a_0)^q|R_{nl}Y_l^m|^2\,\mathrm{d}^3\vec{r}$$

When this integral is written in spherical coordinates, the integration of the square spherical harmonic over the angular coordinates produces one. So, the expectation value simplifies to

$$\langle q\rangle = \int_{r=0}^{\infty} (r/a_0)^q R_{nl}^2 r^2\,\mathrm{d}r$$

To simplify the notations, a nondimensional radial coordinate $\rho = r/a_0$ will be used. Also, a new radial function $f \equiv \sqrt{a_0^3}\rho R_{nl}$ will be defined. In those terms, the expression above for the expectation value shortens to

$$\langle q \rangle = \int_0^\infty \rho^q f^2 \, \mathrm{d}\rho$$

To further shorten the notations, from now on the limits of integration and $\mathrm{d}\rho$ will be omitted throughout. In those notations, the expectation value of $(r/a_0)^q$ is

$$\langle q \rangle = \int \rho^q f^2$$

Also note that the integrals are improper. It is to be assumed that the integrations are from a very small value of r to a very large one, and that only at the end of the derivation, the limit is taken that the integration limits become zero and infinity.

According to derivation {D.15}, the function R_{nl} satisfies in terms of ρ the ordinary differential equation.

$$-\rho^2 R_{nl}'' - 2\rho R_{nl}' + \left[l(l+1) - 2\rho + \frac{1}{n^2}\rho^2 \right] R_{nl} = 0$$

where primes indicate derivatives with respect to ρ. Substituting in $R_{nl} = f/\sqrt{a_0^3}\rho$, you get in terms of the new unknown function f that

$$f'' = \left[\frac{1}{n^2} - \frac{2}{\rho} + \frac{l(l+1)}{\rho^2} \right] f \tag{D.62}$$

Since this makes f'' proportional to f, forming the integral $\int \rho^q f'' f$ produces a combination of terms of the form $\int \rho^{\text{power}} f^2$, hence of expectation values of powers of ρ:

$$\int \rho^q f'' f = \frac{1}{n^2} \langle q \rangle - 2\langle q-1 \rangle + l(l+1)\langle q-2 \rangle \tag{D.63}$$

The idea is now to apply integration by parts on $\int \rho^q f'' f$ to produce a different combination of expectation values. The fact that the two combinations must be equal will then give the Kramers-Pasternack relation.

Before embarking on this, first note that since

$$\int \rho^q f f' = \int \rho^q \left(\tfrac{1}{2} f^2\right)' = \rho^q \tfrac{1}{2} f^2 \Big| - \int q \rho^{q-1} \tfrac{1}{2} f^2,$$

the latter from integration by parts, it follows that

$$\int \rho^q f f' = \frac{1}{2} \rho^q f^2 \Big| - \frac{q}{2} \langle q-1 \rangle \tag{D.64}$$

This result will be used routinely in the manipulations below to reduce integrals of that form.

Now an obvious first integration by parts on $\int \rho^q f'' f$ produces

$$\int \rho^q f\, f'' = \rho^q f f' \Big| - \int (\rho^q f)'\, f' = \rho^q f f' \Big| - \int q\rho^{q-1} f f' - \int \rho^q f' f'$$

The first of the two integrals reduces to an expectation value of ρ^{q-2} using (D.64). For the final integral, use another integration by parts, but make sure you do not run around in a circle because if you do you will get a trivial expression. What works is integrating ρ^q and differentiating $f'f'$:

$$\int \rho^q f f'' = \rho^q f f' \Big| - \frac{q}{2}\rho^{q-1} f^2 \Big| + \frac{q(q-1)}{2}\langle q-2\rangle - \frac{\rho^{q+1}}{q+1} f'^2 \Big| + 2\int \frac{\rho^{q+1}}{q+1} f' f'' \tag{D.65}$$

In the final integral, according to the differential equation (D.62), the factor f'' can be replaced by powers of ρ times f:

$$2\int \frac{\rho^{q+1}}{q+1} f' f'' = 2\int \frac{\rho^{q+1}}{q+1}\left[\frac{1}{n^2} - \frac{2}{\rho} + \frac{l(l+1)}{\rho^2}\right] f f'$$

and each of the terms is of the form (D.64), so you get

$$2\int \frac{\rho^{q+1}}{q+1} f' f'' = \frac{1}{(q+1)n^2}\rho^{q+1} f^2 \Big| - \frac{2}{q+1}\rho^q f^2 \Big| + \frac{l(l+1)}{q+1}\rho^{q-1} f^2 \Big|$$
$$-\frac{1}{n^2}\langle q\rangle + \frac{2q}{q+1}\langle q-1\rangle - \frac{l(l+1)(q-1)}{q+1}\langle q-2\rangle$$

Plugging this into (D.65) and then equating that to (D.63) produces the Kramers-Pasternack relation. It also gives an additional right hand side

$$\rho^q f f' \Big| - \frac{q\rho^{q-1}}{2} f^2 \Big| - \frac{\rho^{q+1}}{q+1} f'^2 \Big| + \frac{\rho^{q+1}}{(q+1)n^2} f^2 \Big| - \frac{2\rho^q}{q+1} f^2 \Big| + \frac{l(l+1)\rho^{q-1}}{q+1} f^2 \Big|$$

but that term becomes zero when the integration limits take their final values zero and infinity. In particular, the upper limit values always become zero in the limit of the upper bound going to infinity; f and its derivative go to zero exponentially then, beating out any power of ρ. The lower limit values also become zero in the region of applicability that $\langle q-2\rangle$ exists, because that requires that $\rho^{q-1} f^2$ is for small ρ proportional to a power of ρ greater than zero.

The above analysis is not valid when $q = -1$, since then the final integration by parts would produce a logarithm, but since the expression is valid for any other q, not just integer ones you can just take a limit $q \to -1$ to cover that case.

D.84 Band gap explanation derivations

To see mathematically how the results of note {N.9} were obtained requires knowledge of linear algebra. If you are unaware of it, definitely skip the below derivation.

First define the "growth matrix G that gives the values of ψ, ψ' at $x = d_x$ given the values at $x = 0$:

$$\begin{pmatrix} \psi(d_x) \\ \psi'(d_x) \end{pmatrix} = G \begin{pmatrix} \psi(0) \\ \psi'(0) \end{pmatrix}$$

Simply take the initial conditions to be 1,0 and 0,1 respectively, and find the solutions at d_x to find the two columns of G.

Since the potential is the same in all atomic cell, matrix G describes the change over any cell, not just the first one. And for a periodic solution for a box with N_x "atoms," after N_x applications of G the original values of ψ, ψ' must be obtained. According to linear algebra, and assuming that the two eigenvalues of G are unequal, that means that at least one eigenvalue of G raised to the power N_x must be 1.

Now matrix G must have unit determinant, because for the two basic solutions with 1,0 and 0,1 initial conditions,

$$\psi_1 \psi_2' - \psi_1' \psi_2 = \text{constant} = 1$$

for all x. The quantity in the left hand side is called the Wronskian of the solutions. To verify that it is indeed constant, take ψ_1 times the Hamiltonian eigenvalue problem for ψ_2 minus ψ_2 times the one for ψ_1 to get

$$0 = \psi_1 \psi_2'' - \psi_2 \psi_1'' = (\psi_1 \psi_2' - \psi_2 \psi_1')'$$

According to linear algebra, if G has unit determinant then the product of its two eigenvalues is 1. Therefore, if its eigenvalues are unequal and real, their magnitude is unequal to 1. One will be less than 1 in magnitude and the other greater than 1. Neither can produce 1 when raised to the power N_x, so there are no periodic solutions. Energies that produce such matrices G are in the band gaps.

If the eigenvalues of G are complex conjugates, they must have magnitude 1. In that case, the eigenvalues can always be written in the form

$$e^{\mathrm{i}k_x d_x} \quad \text{and} \quad e^{-\mathrm{i}k_x d_x}$$

for *some* value of k_x. For either eigenvalue raised to the power N_x to produce 1, $N_x k_x d_x$ must be a whole multiple of 2π. That gives the same wave number values as for the free-electron gas.

To see when the eigenvalues of G have the right form, consider the sum of the eigenvalues. This sum is called the trace. If the eigenvalues are real and

unequal, and their product is 1, then the trace of G must be greater than 2 in magnitude. (One way of seeing that for positive eigenvalues is to multiply out the expression $(\sqrt{\lambda_1}-\sqrt{\lambda_2})^2 > 0$. For negative ones, add two minus signs in the square roots.) Conversely, when the eigenvalues are complex conjugates, their sum equals $2\cos(k_x d_x)$ according to the Euler formula (2.5). That is less than 2 in magnitude. So the condition for valid periodic eigenfunctions becomes

$$\text{trace}(G) = 2\cos(k_x d_x) \qquad k_x d_x = \frac{n_x}{N_x} 2\pi$$

From the fact that periodic solutions with twice the crystal period exist, (the ones at the band gaps), it is seen that the values of the trace must be such that the cosine runs through the entire gamut of values. Indeed when the trace is plotted as a function of the energy, it oscillates in value between minima less than -2 and maxima greater than 2. Each segment between adjacent minima and maxima produces one energy band. At the gap energies

$$v_x^{\rm p} = \frac{{\rm d}E_x^{\rm p}}{{\rm d}\hbar k_x} = \frac{1}{\hbar}\frac{{\rm d}2\cos(k_x d_x)}{{\rm d}k_x} \Big/ \frac{{\rm dtrace}(G)}{{\rm d}E_x^{\rm p}} = 0$$

because the cosine is at its ± 1 extrema at the gap energies. So the velocity becomes zero at the ends of the bands.

Identification of the eigenfunctions using the growth matrix G is readily put on a computer. A canned zero finder can be used to find the energies corresponding to the allowed values of the trace.

Appendix N

Notes

This appendix collects various notes on the material. This sort of material is often given in footnotes at the bottom of the text. However, such a footnote is distracting. You tend to read them even if they are probably not really that important to you. Also, footnotes have to be concise, or they make a mess of the main text.

N.1 Why this book?

With the current emphasis on nanotechnology, quantum mechanics is becoming increasingly essential to engineering students. Yet, the typical quantum mechanics texts for physics students are not written in a style that most engineering students would likely feel comfortable with. Furthermore, an engineering education provides very little real exposure to modern physics, and introductory quantum mechanics books do little to fill in the gaps. The emphasis tends to be on the computation of specific examples, rather than on discussion of the broad picture. Undergraduate physics students may have the luxury of years of further courses to pick up a wide physics background, engineering graduate students not really. In addition, the coverage of typical introductory quantum mechanics books does not emphasize understanding of the larger-scale quantum system that a density functional computation, say, would be used for.

Hence this book, written by an engineer for engineers. As an engineering professor with an engineering background, this is the book *I* wish I would have had when I started learning real quantum mechanics a few years ago. The reason I like this book is not because I wrote it; the reason I wrote this book is because I like it.

This book is not a popular exposition: quantum mechanics can only be described properly in the terms of mathematics; suggesting anything else is crazy. But the assumed background in this book is just basic undergraduate calculus and physics as taken by all engineering undergraduates. There is no intention to

teach students proficiency in the clever manipulation of the mathematical machinery of quantum mechanics. For those engineering graduate students who may have forgotten some of their undergraduate calculus by now, there are some quick and dirty reminders in the notations. For those students who may have forgotten some of the details of their undergraduate physics, frankly, I am not sure whether it makes much of a difference. The ideas of quantum mechanics are that different from conventional physics. But the general ideas of classical physics are assumed to be known. I see no reason why a bright undergraduate student, having finished calculus and physics, should not be able to understand this book. A certain maturity might help, though. There are a lot of ideas to absorb.

My initial goal was to write something that would "read like a mystery novel." Something a reader would not be able to put down until she had finished it. Obviously, this goal was unrealistic. I am far from a professional writer, and this is quantum mechanics, after all, not a murder mystery. But I have been told that this book is very well written, so maybe there is something to be said for aiming high.

To prevent the reader from getting bogged down in mathematical details, I mostly avoid nontrivial derivations in the text. Instead I have put the outlines of these derivations in notes at the end of this document: personally, I enjoy checking the correctness of the mathematical exposition, and I would not want to rob my students of the opportunity to do so too. In fact, the chosen approach allows a lot of detailed derivations to be given that are skipped in other texts to reduce distractions. Some examples are the harmonic oscillator, orbital angular momentum, and radial hydrogen wave functions, Hund's first rule, and rotation of angular momentum. And then there are extensive derivations of material not even included in other introductory quantum texts.

While typical physics texts jump back and forward from issue to issue, I thought that would just be distracting for my audience. Instead, I try to follow a consistent approach, with as central theme the method of separation-of-variables, a method that most mechanical graduate students have seen before already. It is explained in detail anyway. To cut down on the issues to be mentally absorbed at any given time, I purposely avoid bringing up new issues until I really need them. Such a just-in-time learning approach also immediately answers the question why the new issue is relevant, and how it fits into the grand scheme of things.

The desire to keep it straightforward is the main reason that topics such as Clebsch-Gordan coefficients (except for the unavoidable introduction of singlet and triplet states) and Pauli spin matrices have been shoved out of the way to a final chapter. My feeling is, if I can give my students a solid understanding of the basics of quantum mechanics, they should be in a good position to learn more about individual issues by themselves when they need them. On the other hand, if they feel completely lost in all the different details, they are not likely

to learn the basics either.

That does not mean the coverage is incomplete. All topics that are conventionally covered in basic quantum mechanics courses are present in some form. Some are covered in much greater depth. And there is a lot of material that is not usually covered. I include significant qualitative discussion of atomic and chemical properties, Pauli repulsion, the properties of solids, Bragg reflection, and electromagnetism, since many engineers do not have much background on them and not much time to pick it up. The discussion of thermal physics is much more elaborate than you will find in other books on quantum mechanics. It includes all the essentials of a basic course on classical thermodynamics, in addition to the quantum statistics. I feel one cannot be separated from the other, especially with respect to the second law. While mechanical engineering students will surely have had a course in basic thermodynamics before, a refresher cannot hurt. Unlike other books, this book also contains a chapter on numerical procedures, currently including detailed discussions of the Born-Oppenheimer approximation, the variational method, and the Hartree-Fock method. Hopefully, this chapter will eventually be completed with a section on density-functional theory. (The Lennard-Jones model is covered earlier in the section on molecular solids.) The motivation for including numerical methods in a basic exposition is the feeling that after a century of work, much of what can be done analytically in quantum mechanics has been done. That the greatest scope for future advances is in the development of improved numerical methods.

Knowledgeable readers may note that I try to stay clear of abstract mathematics when it is not needed. For example, I try to go slow on the more abstract vector notation permeating quantum mechanics, usually phrasing such issues in terms of a specific basis. Abstract notation may seem to be completely general and beautiful to a mathematician, but I do not think it is going to be intuitive to a typical engineer. The discussion of systems with multiple particles is centered around the physical example of the hydrogen molecule, rather than particles in boxes. The discussion of solids in chapter 10 avoids the highly abstract Dirac comb (delta functions) mathematical model in favor of a physical discussion of more realistic one-dimensional crystals. The Lennard-Jones potential is derived for two atoms instead of harmonic oscillators.

The book tries to be as consistent as possible. Electrons are grey tones at the initial introduction of particles, and so they stay through the rest of the book. Nuclei are red dots. Occupied quantum states are red, empty ones grey. That of course required all figures to be custom made. They are not intended to be fancy but consistent and clear. I also try to stay consistent in notations throughout the book, as much as is possible without deviating too much from established usage.

When I derive the first quantum eigenfunctions, for a pipe and for the harmonic oscillator, I make sure to emphasize that they are not *supposed* to look like anything that we told them before. It is only natural for students to want

to relate what we told them before about the motion to the completely different story we are telling them now. So it should be clarified that (1) no, they are not going crazy, and (2) yes, we will eventually explain how what they learned before fits into the grand scheme of things.

Another difference of approach in this book is the way it treats classical physics concepts that the students are likely unaware about, such as canonical momentum, magnetic dipole moments, Larmor precession, and Maxwell's equations. They are largely "derived" in quantum terms, with no appeal to classical physics. I see no need to rub in the student's lack of knowledge of specialized areas of classical physics if a satisfactory quantum derivation is readily given.

This book is not intended to be an exercise in mathematical skills. Review questions are targeted towards understanding the ideas, with the mathematics as simple as possible. I also try to keep the mathematics in successive questions uniform, to reduce the algebraic effort required. There is an absolute epidemic out there of quantum texts that claim that "the only way to learn quantum mechanics is to do the exercises," and then those exercises turn out to be, by and large, elaborate exercises in integration and linear algebra that take excessive time and have nothing to do with quantum mechanics. Or worse, they are often basic theory. (Lazy authors that claim that basic *theory* is an "exercise" avoid having to cover that material themselves and also avoid having to come up with a *real* exercise.) Yes, I too did waste a lot of time with these. And then, when you are done, the answer teaches you nothing because you are unsure whether there might not be an algebraic error in your endless mass of algebra, and even if there is no mistake, there is no hint that it means what you think it means. All that your work has earned you is a 75/25 chance or worse that you now "know" something that is not true. Not in this book.

Finally, this document faces the very real conceptual problems of quantum mechanics head-on, including the collapse of the wave function, the indeterminacy, the nonlocality, and the symmetrization requirements. The usual approach, and the way I was taught quantum mechanics, is to shove all these problems under the table in favor of a good sounding, but upon examination self-contradictory and superficial story. Such superficiality put me off solidly when they taught me quantum mechanics, culminating in the unforgettable moment when the professor told us, seriously, that the wave function *had* to be symmetric with respect to exchange of bosons *because* they are all truly the same, and then, when I was popping my eyes back in, continued to tell us that the wave function is *not* symmetric when fermions are exchanged, which are all truly the same. I would not do the same to my own students. And I really do not see this professor as an exception. Other introductions to the ideas of quantum mechanics that I have seen left me similarly unhappy on this point. One thing that really bugs me, none had a solid discussion of the many worlds interpretation. This is obviously not because the results would be incorrect, (they have not been contradicted for half a century,) but simply because the

teachers just do not like these results. I do not like the results myself, but basing teaching on what the teacher would *like* to be true rather on what the evidence indicates *is* true remains absolutely unacceptable in my book.

N.2 History and wish list

- Aug. 26, 2018. Version 5.63 alpha.

  The main new thing is a correction in section 6.23 on Semiconductors. In the detailed explanation of how n-type semiconductors work, I wrote somewhere "conduction band" where I meant "valence band." This of course does not improve clarity. Even after fixing, I thought the discussion was still confusing, so I rewrote the corresponding paragraphs from scratch.
- July 3, 2018. Version 5.62 alpha.

  Mainly rewrites of Hartree-Fock derivation {D.54}. If I cannot follow my own reasoning, it is pretty bad.
- April 27, 2018. Version 5.61 alpha.

  Rewrote section 9.1.1 for readability. Also rewrote addendum {A.7}. Other minor rewrites.

  Reposting because an (ununderstood) latex2html bug appeared out of nothing with the (tricky) way -1 is formatted on the web pages. This bug is not repeatable *and it did not happen for the colorized version nor for the remake.* I am putting in some precautions anyway.
- April 20, 2018. Version 5.60 alpha.

  Rewrote the introductory Hartree-Fock section 9.3.1 to be more readable and more independent of earlier chapters. Expanded on the description of Hartree-Fock correlation energy in 9.3.5.4 and its note N.18. Incomplete, but I am posting it now because I found and fixed some ugly errors.
- Mar. 4, 2018. Version 5.59 alpha.

  Corrected a typo (j instead of l for orbital angular momentum) at the start of section 12.8.

  In Multiple-Particle Systems:

  In the periodic table, added the recently assigned names for element numbers 113, 115, 117, 118. Now every element in the table has a real name.

  In Macroscopic systems:

  Completely rewrote note N.9 explaining how band gaps arise if you add a bit of a periodic potential to the free electron gas. Corrected "the crystal spacing is a half-integer multiple of the [Bragg] wave

lengths" into "double the atom spacing is a whole multiple of the wave lengths". ("half-integer" seems to exclude "whole integer.")

In Time Evolution:

Cleaned up the links in the pdf to the animations on the web. Also, weblinks now go to the same server as the pdf is downloaded from.

Completely rewrote addendum A.17 on the virial theorem, as I now expect I will need it when I cover density functional theory.

In Classical and Quantum Thermodynamics:

More rewrites in "Specific Heats".

In Nuclei:

Added the Z vs N form of all Z vs ΔN graphs as external pdf files.

Showed the nuclear decay processes first in a "Chart of the Nuclides" (Z vs N) form.

Explained how the Z vs ΔN form clearly illustrates that nucleons of the same type like to pair up, while the Z vs N form does not.

In subsection 14.5.1, added a graph showing half-lifes in addition to the nuclear decay processes.

Changed spin colors to make the individual spins easier to recognize.

Left out the unused colors in the legends for the same reason.

In the spin plots, nuclei whose spin has reservations are now shown in the expected color instead of yellow, with a light check mark (a cross if none yet). That increases information content a lot. All nuclei now get a mark in the spin plots except those with unknown spin (yellow squares). Added a mark for nuclei in which an imperfect odd-nucleon model lowers the spin by one unit.

For the parity plots, used light colors instead of yellow crosses to indicate which parities with reservations are predicted OK or not. In the odd-even respectively even-odd parity graphs, shell model parity lines are now only shown at the odd-even, respectively even-odd locations. Both sets of odd-A parity lines are now shown in the odd-odd parity graphs.

In the beta decay rate plots, added the stable nuclei to fill in the holes with known data.

In gamma decay, I converted subsection N.36 into a note. Which it should always have been, as it is not part of standard theory but an hypothesis. Also rewrote it quite a bit to improve the presentation.

- Nov 1, 2016. Version 5.57 alpha. Corrected a number of errors and poor phrasings pointed out by various readers, as in the acknowledgements.

Corrected "molecular mass" in "The New Variables" into "molar mass". Cannot complain about others if I do it myself.

Yes, I need to get back to doing some more serious writing on this book.

- Feb 4, 2015. Version 5.57 alpha. Corrected some rather horrible typos pointed out by Rob Vossen.
- Nov 14, 2013. Version 5.56 alpha. The book has been converted to be processed by l2h instead of LaTeX2HTML-FU. Web page hyphenation and bad-math-break prevention are now done by l2h, instead of inside the latex source. And the Wordperfect and MS Word grammar checkers can now be applied. (In document pieces, to be sure. Not on all 1,600 pages at once!) Removed "we" from the notations. More rewriting of {A.9}. Added Flerovium and Livermorium to the periodic table. Corrected some errors pointed out by kind readers.
- June 15, 2012. Version 5.55 alpha. Added hyperspheres to the notations. Finished {A.8} and {D.22}. Rewrote {A.9}. These may need some more editing. The problems of the previous version with the index have been fixed. Various small problems fixed.

  Various corrections of this version were posted the next few days. They only differ in formatting of the index and notations. Also some very minor editing, part of which may not have been posted. But now the formatting problems really seem fixed. I hope.
- June 11, 2012. Version 5.54 alpha. Added the screened Poisson equation to {D.2}. Rewrote {A.8} on positive unique ground states and added an explanatory figure. Needs more work. Added an entry on spherical coordinates to the notations. Cleaned up and expanded {D.14} a bit. Editing of {A.42}. Improved the values of the physical constants in the notations. Improved alignment of inline math for Internet Explorer. Took inline equations apart and added tuned relation symbols. Created a better key image for key points. Improved appearance of the key point and question lists. Bolded math labels in the notations and added some space behind the labels. In the web version symbols are no longer on preceding lines. In the pdf, links are now enclosed by non-offensive thin grey lines. Urls are now enclosed by non-offensive thin black lines. In html, very long words, including all 14 characters long or more. will now hyphenate if needed. That includes one-dimensional, two-dimensional, three-dimensional, four-dimensional, n-dimensional. accomplishments, acknowledgments, anticommutators, antisymmetrically, antisymmetrization, ... Also in html, stupid line breaks at inline math images should no longer occur.

  Please note that surely ten thousands of small changes have been made. Some are bound to have introduced a problem, especially

in formatting. I do think the esthetics of the web pages has been improved tremendously. Even more so for Internet Explorer, I guess. (But I use linux.)

Added note: the zipped version of this date has some minor format corrections, but also an error fix in {D.22} (which still needs more work.)

Added note: unfortunately, now the index looks like hell.

- May 4, 2012. Version 5.53 alpha. Corrections and extensions to fundamental forces, section 7.5.2. Editing of {A.42}.
- Apr. 25, 2012. Version 5.52 alpha. Editing of {A.42}. Corrected figures 3.3 and N.2. Very minor changes the next day.
- Feb. 15, 2012. Version 5.51 alpha. Improved discussion of field operators in {A.15.9}. Added a brief explanation of the Casimir force, {A.23.4}. Noted that pure particle and pure antiparticle states preserve norm, {D.32} and {A.15.9}. Corrected that time reversal is antiunitary, not unitary, {A.19.2} Greatly expanded the discussion of the Fourier inversion theorem and Parseval, {A.26}. Added separate derivation of the Green's function of the Poisson equation, {D.2}. Added an example of variational calculus, {A.2}. Added fields to classical Lagrangian and Hamiltonian analysis, {A.1.5}. Added a new addendum on forces by particle exchange, {A.22}. While already big, this one will need to be expanded in a later version. Added that lone systems have definite spin and parity to chapter 7.3. Added draft explanation of the OPEP to {A.42}.
- Dec. 15, 2011. Version 5.50 alpha. Further rewrites and correction of a couple of glaring errors in Quantum Field Theory in a Nanoshell {A.15}. I think this addendum is now much better than it was originally. Removed interpretation of the photon wave function again. Rewrites in 8.7
- Dec. 12, 2011. Version 5.49 alpha. Corrected previous history item. The Dirac γ matrices are now defined, {A.36}. Rest mass is now just m, not m_0. Added note about Majorana neutrinos to {A.44}. Added reference for gauge symmetries in {A.19.5}. Put in some disclaimers in {A.42}. Added an interpretation of the photon wave function. There is now a much needed draft rewrite of Quantum Field Theory in a Nanoshell {A.15}.
- Nov. 18, 2011. Version 5.48 alpha. Added a complete comparison of the mixed dipole and quadrupole gamma decays, figure 14.65. Minor corresponding rewrites. Nowadays NuDat 2 no longer tells you that you are requesting too much data and simply gives you partial data. That explained why M1 and E2 transitions stopped so quickly. It has been corrected. Added a description of neutrinos, {A.44}.

- Nov. 14, 2011. Version 5.47 alpha. Corrected the definition of "interval" in special relativity to be author-dependent. Due to a program error, in the comparison of gamma decay with data the program was actually selecting the nuclei to be as much the same as possible, instead of as much different as possible. It has been corrected. Not that it makes much of a visual difference. The text has been rewritten a bit too. In the "cage of Faraday" proposal, stupid me forgot about measured electromagnetic moments. That seems to kill off Meisner pretty well. It has been rewritten.
- Nov. 9, 2011. Version 5.46 alpha. Modified the "cage of Faraday" proposal a bit.
- Nov. 8, 2011. Version 5.45 alpha. Added the cage-of-Faraday proposal, chapter N.36.
- Nov. 8, 2011. Version 5.44 alpha. Added an example without charge independence to isospin to clarify the need for it. Corrected a typo in a formula and minor editing. Added examples to E0 internal conversion and edited the text a bit in chapter 14.20.6. Added a comparison of the single-particle theory of gamma decay with experimental data to chapter 14.20.5.
- Oct. 25, 2011. Version 5.43 alpha. Further editing of isospin. Seems to be OK for now. Some minor editing of quantum field theory.
- Oct. 17, 2011. Version 5.42 alpha. Rewrote the discussion for isospin. The original discussion was only defensible for a very crude model of the deuteron. The current discussion is better but needs more work.
- Oct. 12, 2011. Version 5.41 alpha. I finally got around to taking out some very dubious statements on gamma decay in 14.20 that have been bothering me for years. The relief is tremendous. I even managed to find a ballpark for the "missing" E0 transition.
- Oct. 03, 2011. Version 5.40 alpha. Some minor corrections and improvements on emission of radiation.
- Sep. 30, 2011. Version 5.39 alpha. Some minor corrections and improvements on emission of radiation. Being doing other stuff.
- Sep. 26, 2011. Version 5.38 alpha. Adds a table of hydrogen radial correction factors. Deals better with the magnetic transitions of nonrelativistic hydrogen.
- Sep. 16, 2011. Version 5.37 alpha. Minor changes and corrections in the previous items. Expanded the table with Weisskopf/Moszkowski correction factors.
- Sep. 12, 2011. Version 5.36 alpha. Minor changes in section 8.4. Added addendum {A.25} on multipole transitions, with corresponding derivations and notes {D.43}, {N.13}, and {N.14}. Took out

claim from Wikipedia that the 21 cm line is highly forbidden and replaced it with low energy.

- Aug. 10, 2011. Version 5.35 alpha. Completed addenda A.23 and A.24 Added section 8.4.
- Aug. 5, 2011. Version 5.34 alpha. Significant rewrite of section 7.4.3. Doing the electric transitions first really helps getting rid of all the ifs and buts. Some rewrite of addendum N.10. Improved the discussion of two-photon zero angular momentum in section 7.4.4. On second thought, bringing up the alpha particle in N.10 was a bad idea. The worst part is that the alpha decay has an exponential dependence on angular momentum due to tunneling. That is less of an issue in beta decay, but then there is the blasted neutrino. So the particle is now a spinless nonrelativistic photon. Gee. The Lorentz gauge is now the Lorenz gauge. The Coulomb gauge is no longer misrepresented. Rewrote addenda A.23 on second quantization and A.24 on quantum derivation of spontaneous emission. They are incomplete; I am posting drafts since I have to go to work anyway.
- July 27, 2011. Version 5.33 alpha. The electric and magnetic fields are now $\mathcal{E}$ and $\mathcal{B}$. Hopefully. Based on [24, p. 240] and a similar message in a physics news group, took the bold step of defining a "type 2 wave function" for the photon. Some rewrite of addendum A.14. Correspondingly added notes A.21 and D.36 on the photon wave function and its derivation. They are incomplete; I am posting drafts since I have to go to work anyway.
- July 20, 2011. Version 5.32 alpha. Corrected statement on parity of the photon. Whatever I learn on the web does *not* hold up well. Added addendum A.20 on the spin of vector particles.
- July 18, 2011. Version 5.31 alpha. Rewrote note physics of the fundamental commutators. Added an addendum with Maxwell's wave equations. Added the Klein-Gordon equation. Or at least, added a separate addendum for it, to eventually replace scattered discussions of it elsewhere.
- July 12, 2011. Version 5.30 alpha. Normalized appendices to allow question lists in them. Moved WKB to an addendum. Minor rewrites in time evolution. Added missing links to derivations in relativity.
- July 05, 2011. Version 5.29 alpha. Moved derivations from addenda to derivations. Added deuteron data from Argonne v18. Slight rewrites for relativity. Relocated some addenda. Cleaned up adiabatic theorem addendum and improved derivation. The Heisenberg formulation has been moved to an addendum. There are now direct links to the animated figures on the web in the pdf. The unsteady particle in the pipe has been animated. Made J the symbol of

generic angular momentum, with much effort. In 3D scattering, D is a private symbol of Griffiths according to his book on elementary particles, not a standard symbol as I thought. It has been removed. Rewrote intro to two-state systems. Changed action in relativistic case to be "stationary," rather than "least." Dropped claim obtained from Internet of no population inversion for semiconductor lasers. Gee. Major rewrite of the first half of time evolution.

- June 2, 2011. Version 5.28 alpha. Draft rewrite of the section on modeling the deuteron. Draft rewrite of the addendum on nuclear forces.
- May 30, 2011. Version 5.27 alpha. Inverted history list. Textual formulae make less use of images. Textual superscripts like powers and degree signs have been joined to their root with nobr tags. Made table formatting more consistent. Line length in html (at recommended browser width) has been reduced a bit. It is still longer than in the pdf version, but about the same as Lamport's LaTeX book. Maximum formulae, table, and figure lengths are now consistent in html and equal to the recommended browser width. This required a 3% reduction in displayed formulae size. Table sizes must be explicitly set due to buggy LaTeX2HTML code. Center environments have been replaced by centering to eliminate redundant space, usually after the caption. Fixed formatting problems in table of contents and lists of figures, tables. Pdf links to figures and tables do now show the actual thing. Advanced angular momentum has been made a separate chapter. HTML table of contents has been cleaned up. HTML truncation of the two electron Clebsch Gordon figure has been fixed. Actually, it has NOT been fixed. :() A new attempt has been implemented. Expanded the description of fundamental forces in the section on particle exchange. Draft rewrite of the first two sections of nuclei. Made the section on nuclear forces a more limited addendum. Also eliminated a bad mistake in it.
- Apr. 25, 2011. Version 5.26 alpha. Some further minor rewrites of the section on index notation. Spell check and minor rewrites of the relativity chapter in general.
- Apr. 18, 2011. Version 5.25 alpha. Rewrote section on index notation, added an addendum about it.
- Mar. 14, 2011. Version 5.24 alpha. Bisected addenda on quantum field theory and perturbation theory. Started just a bit cleaning up the chapter on nuclei. One tiny step at a time.
- Mar. 14, 2011. Version 5.23 alpha. Brought structure into the 124 notes. Relocated the chapter "Additional Topics" in the addenda notes. Made relativity a separate Part. Minor rewrites in the intro to nuclei.

- Mar. 8, 2011. Version 5.22 alpha. Noted and fixed a blunder on cosmic redshift. Sometimes I do not seem to think at all. Everyone else seems to make the same blunder??
- Mar. 7, 2011. Version 5.21 alpha. All remaining low-resolution bitmap graphics are gone. Spherical Hankel and Bessel functions have been taken out of the note on 3D scattering and given their own note. The same for the integral Schrödinger equation. The remaining note on 3D scattering has been given a much needed clean up. Some rewrites in the note on symmetries.
- Feb. 7, 2011. Version 5.20 alpha. Some changes in the section and note on conservation laws and symmetries. A large fraction of the remaining low-resolution bitmap graphics has been replaced by higher quality vector graphics. The note on special relativity has been turned into a chapter.
- Jan. 16, 2011. Version 5.19 alpha. Very minor changes in the section on conservation laws and symmetries.
- Jan. 3, 2011. Version 5.18 alpha. Revised section on conservation laws and symmetries. Slight cosmetic improvements in some figures.
- Nov. 30, 2010. Version 5.17 alpha. Moved hydrogen and helium to the end of the periodic table; I simply got tired of saying "except hydrogen" and "except helium." Rewrote the subsection on ionic conduction.
- Nov. 16, 2010. Version 5.16 alpha. Second and for now final version of the subsection on typical metals and insulators.
- Nov. 12, 2010. Version 5.15 alpha. First rewrite of the subsection on typical metals and insulators.
- Nov. 1, 2010. Version 5.14 alpha. Various minor rewrites in the chapter on macroscopic systems.
- Oct. 11, 2010. Version 5.13 alpha. Technical modifications to allow links in the pdf. Many people seem to use the pdf for reading instead of the web pages.
- Oct. 4, 2010. Version 5.12 alpha. Various minor rewrites, including for subsection 6.22.5. A number of poor phrasings pointed out by Ramaswami Sastry Vedam corrected.
- Sep. 13, 2010. Version 5.11 alpha. Main change is the addition of subsection 6.22.5 giving an introduction to the band theory of three-dimensional crystals.
- Aug. 30, 2010. Version 5.10 alpha. Added spectra of actual materials to the section on crystal momentum. Fixed an erroneous statement about the presentation of spectra. Fixed an error where spins were listed, but not included in the transformation in the section on conservation laws and symmetries.

- Aug. 23, 2010. Version 5.09 alpha. Rewrites of p-n junction and transistor.
- Aug. 9, 2010. Version 5.08 alpha. Third law moved to notes. Edits in the chapter on macroscopic systems. Draft proof of the Onsager relations added and immediately removed. Gee. It will not be back.
- July 28, 2010. Version 5.07 alpha. Better description of the third law. Some rewrites and error corrections in thermoelectrics.
- July 23, 2010. Version 5.06 alpha. Some rewrites and error corrections in thermoelectrics.
- July 19, 2010. Version 5.05 alpha. Some error corrections in thermoelectrics and a discussion of the Onsager relations added.
- July 16, 2010. Version 5.04 alpha. Some rewrites and error corrections.
- July 13, 2010. Version 5.03 alpha. Various rewrites and error corrections. Also a new periodic table.
- June 6, 2010. Version 5 alpha. Lots of spelling and grammar corrections, minor rewrites, and additions of summaries during teaching a 3 hour DIS on "quantum literacy." Added a chapter on nuclei. Added sections and tables on angular momentum of shells. Alpha decay has been moved to the new chapter on nuclei. Forbidden decays are now included. Various program improvements/tuning. Corrected "effective" mass (for a two-body system) into "reduced." Added a chapter on macroscopic systems to Part II. Much of this chapter has been scavenged from Part III. It is supposed to provide some more practical knowledge in various areas. It was inspired by the DIS mentioned above, which showed you cannot do much of Part III in a single semester. The semiconductor discussion in the chapter is all new.
- March 22, 2009. Version 4.2 alpha. Spin matrices for systems greater than spin one half are now discussed. Classical Lagrangian and Hamiltonian dynamics is now covered in a note. Special relativity is now covered in a note. There is now a derivation of the hydrogen dipole selection rules and more extensive discussion of forbidden transitions. Angular momentum and parity conservation in transitions are now discussed. The Gamow theory data are now corrected for nuclear versus atomic mass. There is no perceivable difference, however. The alignment bars next to the electromagnetic tables in the web version should have been eliminated.
- Jan. 1, 2009. Version 4.0 alpha reorders the book into two parts to achieve a much better book structure. The changed thinking justifies a new version. Parts of the lengthy preface have been moved to the notes. The background sections have been combined in their own chapter to reduce distraction in part II. There is now a derivation of

effective mass in a note. A few more commutators were added to the reference. There is a note on Fourier transforms and the Parseval equality. The stupid discussion of group velocity has been replaced by a better (even more stupid?) one. Two of the gif animations were erroneous (the nondelta function tunneling and the rapid drop potential) and have been corrected. High resolution versions of the animations have been added. Time-dependent perturbation theory is now concisely covered. WKB theory is now covered. Alpha decay is now included. The adiabatic theorem is now covered. Three-dimensional scattering is now covered, in a note. Fixed a mistyped shelf number energy in the thermo chapter. The derivations of the Dirac equation and the gyromagnetic ratio of electron spin have been moved to the notes. Note D.71 now gives the full derivation of the expectation Lorentz force. The direction of the magnetic field in the figure for Maxwell's fourth law was corrected. A section on electrostatic solutions has been added. The description on electrons in magnetic fields now includes the diamagnetic contribution. The section on Stern-Gerlach was moved to the electromagnetic section where it belongs. Electron split experiments have been removed completely. There is now a full coverage of time-independent small perturbation theory, including the hydrogen fine structure. Natural frequency is now angular frequency. Gee. The Planck formula is now the Planck-Einstein relation. The Euler identity is now the apparently more common Euler formula. Black body as noun, blackbody as compound adjective.

- Jan. 19, 2009. Version 4.1 alpha. There is now a discussion of the Heisenberg picture. The horribly written, rambling, incoherent, section on nearly-free electrons that has been bothering me for years has been rewritten into two much better sections. There is now a discussion on the quantization of the electromagnetic field, including photon spin and spontaneous emission. The Rayleigh formula is now derived. The perturbation expansion of eigenfunctions now refers to Rellich's book to show that it really works for degenerate eigenfunctions.
- July 14, 2008. Version 3 beta 4.2 expands the section on unsteady two-state systems to include a full discussion of "time-dependent perturbation theory," read emission and absorption of radiation. Earlier versions just had a highly condensed version since I greatly dislike the derivations in typical textbooks that are full of nontrivial assumptions for which no justification is, or can be, given at all.
- July 2, 2008. Version 3 beta 4.1 adds a new, "advanced," chapter on basic and quantum thermodynamics. An advanced section on the fundamental ideas underlying quantum field theory has also been

added. The discussion of the lambda transition of helium versus Bose-Einstein condensation has been rewritten to reflect the considerable uncertainty. Uniqueness has been added to the note on the hydrogen molecular ion ground state properties. Added a missing 2π in the Rayleigh-Jeans formula.

- April 7, 2008. Version 3 beta 4 adds key points and exercises added to chapter 4, with the usual rewrites to improve clarity. The Dirichlet completeness proof of the Fourier modes has been moved from the solution manual to the notes. The actual expressions for the hydrogen molecular ion integrals are now given in the note. The London force derivation has been moved to the notes. The subsection of ferromagnetism has been rewritten to more clearly reflect the uncertainty in the field, and a discussion of Hund's rules added.
- Dec. 20, 2007. Version 3 beta 3.4 cleans up the format of the "notes." No more need for loading an interminable web page of 64 notes all at the same time over your phone line to read 20 words. It also corrects a few errors, one important one pointed out by Johann Joss. It also also extends some further griping about correlation energy to all three web locations. You may surmise from the lack of progress that I have been installing Linux on my home PC. You are right.
- Sept. 9, 2007. Version 3 beta 3.3 mainly adds sections on solids, that have been combined with rewritten free and nearly-free electrons sections into a full chapter on solids. The rest of the old chapter on examples of multiple particle systems has been pushed back into the basic multiple particle systems chapter. A completely nonsensical discussion in a paragraph of the free-electron gas section was corrected; I cannot believe I have read over that several times. I probably was reading what I wanted to say instead of what I said. The alternative name "twilight terms" has been substituted for "exchange terms." Many minor changes.
- July 19, 2007. Version 3 beta 3.2 adds a section on Hartree-Fock. It took forever. My main regret is that most of them who wasted my time in this major way are probably no longer around to be properly blasted. Writing a book on quantum mechanics by an engineer for engineers is a minefield of having to see through countless poor definitions and dubious explanations. It takes forever. In view of the fact that many of those physicist were probably supported by tax payers much of the time, it should not be such an absolute mess!

  There are some additions on Born-Oppenheimer and the variational formulation that were in the Hartree-Fock section, but that I took out, since they seemed to be too general to be shoved away inside an application. Also rewrote section 5.7 and subsection 5.9.2 to be consistent, and in particular in order to have a single consistent

notation. Zero point energy (the vacuum kind) is back. What the heck.

- May 21, 2007. An updated version 3 beta 3.1 to correct a poorly written subsection on quantum confinement for the particle in a pipe. Thanks to Swapnil Jain for pointing out the problem. I do not want people to get lost so early in the game, so I made it a priority correction. In general, I do think that the sections added later to the document are not of the same quality as the original with regard to writing style. The reason is simple. When I wrote the original, I was on a sabbatical and had plenty of time to think and rethink how it would be clearest. The later sections are written during the few spare hours I can dig up. I write them and put them in. I would need a year off to do this as it really should be done.
- May 5, 2007. There are now lists of key points and review questions for chapter 3. That makes it the 3 beta 3 version. Various other fixes, like spectral line broadening, Helium's refusal to take on electrons, and countless other less than ideal phrasings. And full solutions of the harmonic oscillator, spherical harmonics, and hydrogen wave function ODEs, Mandelshtam-Tamm energy-time uncertainty, (all in the notes.) A dice is now a die, though it sounds horrible to me. Zero point energy went out again as too speculative.
- April 2, 2007. There are now lists of key points and review questions for chapter 2. That makes it the 3 beta 2 version. So I guess the final beta version will be 3 beta 6. Various other fixes. I also added, probably unwisely, a note about zero point energy.
- Mid Feb.,2007. There are now lists of key points and review questions for chapter 1. Answers are in the new solution manual.
- Mid Jan., 2007. Added sections on confinement and density of states, a commutator reference, a section on unsteady perturbed two state systems, and an advanced chapter on angular momentum, the Dirac equation, the electromagnetic field, and NMR. Fixed a dubious phrasing about the Dirac equation and other minor changes.
- Mid April, 2006. Various minor fixes. Also I changed the format from the "article" to the "book" style.
- Mid Feb., 2006. A new version was posted. Main differences are correction of a number of errors and improved descriptions of the free-electron and band spectra. There is also a rewrite of the many worlds interpretation to be clearer and less preachy.
- May 11 2005. I got cold feet on immediately jumping into separation of variables, so I added a section on a particle in a pipe.
- May 4, 2005. A revised version was posted. I finally read the paper by Everett, III on the many worlds interpretation, and realized that I had to take the crap out of pretty much all my discussions. I also

rewrote everything to try to make it easier to follow. I added the motion of wave packets to the discussion and expanded the one on Newtonian motion.

- Nov 27, 2004. A revised version was posted, fixing a major blunder related to a nasty problem in using classical spring potentials for more than a single particle. The fix required extensive changes. This version also added descriptions of how the wave function of larger systems is formed.
- Oct 24, 2004. The first version of this manuscript was posted.

Part I is a draft.

Part II is mostly in a fairly good shape. But there are a few recent additions that probably could do with another look. Of course, fairly good is not the same as good. Chapter 7 is poor.

In Part III various parts sure could do with a few more rewrites. For example, the thermodynamics chapter is quite embarrassing. The chapter on nuclei is an incomplete absolute mess.

The shape of addenda and notes is pretty much like their root chapters.

Somewhat notably missing at this time:

1. Electron split experiments. Do engineers need it?? Some mention is now in the basic ideas chapter.
2. Quantum electrodynamics. (The full relativistic theory.) Do engineers need it??
3. Density-functional theory.
4. Mössbauer effect.
5. Superfluidity (but there is not really a microscopic theory.)
6. Superconductivity.

N.3 Nature and real eigenvalues

The major difference between real and complex numbers is that real numbers can be ordered from smaller to larger. So you might speculate that the fact that the numbers of our world are real may favor a human tendency towards simplistic rankings where one item is "worse" or "better" than the other. What if your grade for a quantum mechanics test was $55 + 90\mathrm{i}$ and someone else had a $70 + 65\mathrm{i}$? It would be logical in a world in which the important operators would not be Hermitian.

N.4 Are Hermitian operators really like that?

A mathematician might choose to phrase the problem of Hermitian operators having or not having eigenvalues and eigenfunctions in a suitable space of permissible functions and then find, with some justification, that some operators

in quantum mechanics, like the position or momentum operators do not have any permissible eigenfunctions. Let alone a complete set. The approach of this text is to simply follow the formalism anyway, and then fix the problems that arise as they arise.

More generally, what this book tells you about operators is absolutely true for systems with a finite number of variables, but gets mathematically suspect for infinite systems. The functional analysis required to do better is well beyond the scope of this book and the abstract mathematics a typical engineer would ever want to have a look at.

In any case, when problems are discretized to a finite one for numerical solution, the problem no longer exists. Or rather, it has been reduced to figuring out how the numerical solution approaches the exact solution in the limit that the problem size becomes infinite.

N.5 Why boundary conditions are tricky

You might well ask why you cannot have a wave function that has a change in wave function value at the ends of the pipe. In particular, you might ask what is wrong with a wave function that is a nonzero constant inside the pipe and zero outside it. Since the second derivative of a constant is zero, this (incorrectly) appears to satisfy the Hamiltonian eigenvalue problem with an energy eigenvalue equal to zero.

The problem is that this wave function has "jump discontinuities" at the ends of the pipe where the wave function jumps from the constant value to zero. (Graphically, the function is "broken" into separate pieces at the ends.) Suppose you approximate such a wave function with a smooth one whose value merely drops down steeply rather than jumps down to zero. The steep fall-off produces a first order derivative that is very large in the fall-off regions, and a second derivative that is much larger still. Therefore, including the fall-off regions, the average kinetic energy is not close to zero, as the constant part alone would suggest, but actually almost infinitely large. And in the limit of a real jump, such eigenfunctions produce infinite energy, so they are not physically acceptable.

The bottom line is that jump discontinuities in the wave function are not acceptable. However, the correct solutions will have jump discontinuities in the *derivative* of the wave function, where it jumps from a nonzero value to zero at the pipe walls. Such discontinuities in the derivative correspond to "kinks" in the wave function. These kinks are acceptable; they naturally form when the walls are made more and more impenetrable. Jumps are wrong, but kinks are fine. (Don't break the wave function, but crease it all you like.)

For more complicated cases, it may be less trivial to figure out what singularities are acceptable or not. In general, you want to check the "expectation

value," as defined later, of the energy of the almost singular case, using integration by parts to remove difficult-to-estimate higher derivatives, and then check that this energy remains bounded in the limit to the fully singular case. That is mathematics far beyond what this book wants to cover, but in general you want to make singularities as minor as possible.

N.6 Is the variational approximation best?

Clearly, "best" is a subjective term. If you are looking for the wave function within a definite set that has the most accurate expectation value of energy, then minimizing the expectation value of energy will do it. This function will also approximate the true eigenfunction shape the best, in some technical sense {A.7}. (There are many ways the best approximation of a function can be defined; you can demand that the maximum error is as small as possible, or that the average magnitude of the error is as small as possible, or that a root-mean-square error is, etcetera. In each case, the "best" answer will be different, though there may not be much of a practical difference.)

But given a set of approximate wave functions like those used in finite element methods, it may well be possible to get much better results using additional mathematical techniques like Richardson extrapolation. In effect you are then deducing what happens for wave functions that are beyond the approximate ones you are using.

N.7 Shielding approximation limitations

In the helium atom, if you drop the shielding approximation for the remaining electron in the ionized state, as common sense would suggest, the ionization energy would become negative! This illustrates the dangers of mixing models at random. This problem might also be why the discussion in [25] is based on the zero shielding approximation, rather than the full shielding approximation used here.

But zero shielding does make the base energy levels of the critical outer electrons of heavy atoms very large, proportional to the square of the atomic number. And that might then suggest the question: if the energy levels explode like that, why doesn't the ionization energy or the electronegativity? And it makes the explanation why helium would not want another electron more difficult. Full shielding puts you in the obviously more desirable starting position of the additional electron not being attracted, and the already present electrons being shielded from the nucleus by the new electron. And how about the size of the atoms imploding in zero shielding?

Overall, this book prefers the full shielding approach. Zero shielding would predict the helium ionization energy to be 54.4 eV, which really seems worse

than 13.6 eV when compared to the exact value of 24.6 eV. On the other hand, zero shielding does give a fair approximation of the actual total energy of the atom; 109 eV instead of an exact value of 79. Full shielding produces a poor value of 27 eV for the total energy; the total energy is proportional to the *square* of the effective nucleus strength, so a lack of full shielding will increase the total energy very strongly. But also importantly, full shielding avoids the reader's distraction of having to rescale the wave functions to account for the nonunit nuclear strength.

If eventually X-ray spectra need to be covered in this book, a description of "hot" relativistic inner electrons would presumably fix any problem well.

N.8 Why the s states have the least energy

The probability of being found near the nucleus, i.e. the origin, is determined by the magnitude of the relevant hydrogen wave function $|\psi_{nlm}|^2$ near the origin. Now the power series expansion of ψ_{nlm} in terms of the distance r from the origin starts with power r^l, (D.8). For small enough r, a p, (i.e. ψ_{n1m}), state involving a factor r will be much smaller than an s, (ψ_{n0m}), state without such a factor. Similarly a d, (ψ_{n2m}), state involving a factor r^2 will be much less still than a p state with just single factor r, etcetera. So states of higher angular momentum quantum number l stay increasingly strongly out of the immediate vicinity of the nucleus. This reflects in increased energy since the nuclear attraction is much greater close the nucleus than elsewhere in the presence of shielding.

N.9 Explanation of the band gaps

Chapter 6.21 showed that the spectra of the electrons of solids have "band gaps;" energy ranges for which there are no quantum states for the electrons. These band gaps were qualitatively explained as the remnants of the discrete electron energy states of the individual atoms. These discrete energy states spread out when multiple atoms start interacting, but not necessarily enough to completely remove the gaps.

However, if you start from the free-electron gas point of view, it is much less clear why and when addition of just a bit of crystal potential would suddenly pop up band gaps out of nothing. If you are curious, this note is for you.

To understand what is going on, the Kronig & Penney model will be used. The "crystal" is again taken to be one-dimensional. The potential consists again of a sequence of straight dips, as was shown in green in 6.22. The dips represent the attraction of the atoms on the individual atomic electrons. However, to allow an easier comparison with the free-electron gas solutions, this time the dips will taken far less deep than before. Think of it as a model for a metal, where the outer electrons are only very weakly bound to their atomic cores.

For these shallower "atomic" dips, and for a crystal consisting of very many "atoms," the energy levels are as shown to the left in figure N.1. Note that for the higher energies, this is generally speaking very similar to the energy levels for the free-electron spectrum shown to the left. That should be expected; why would the shallow potential energy dips have much of an effect when the kinetic energy of the electron considered is very large? But even for high energy levels, there are still occasional thin gaps. At these gaps, the electron velocity plunges to zero. Why are these gaps there?

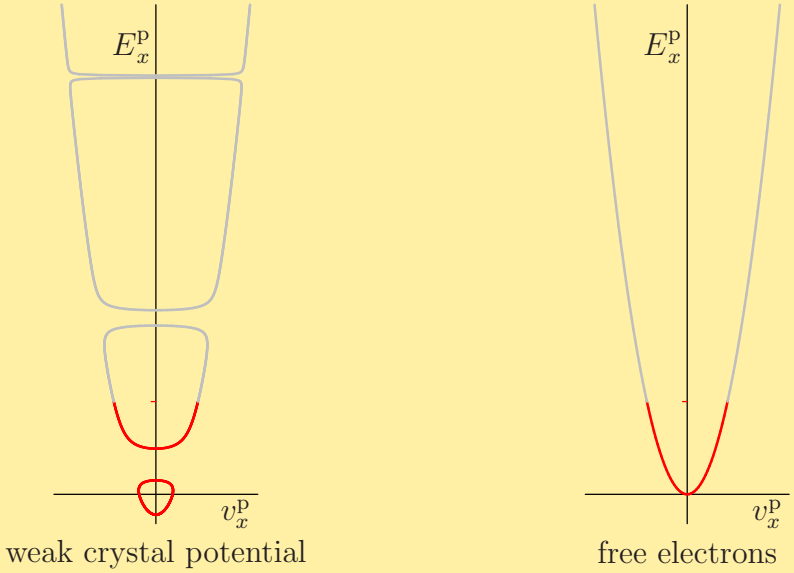


Figure N.1: Spectrum for a weak potential.

To qualitatively understand what is going on, from here on it will be assumed that the periodic "crystal" consists of just 12 "atoms," (rather than, say, a million). Mathematically, after twelve atoms, the quantum wave function becomes the same as it was initially and the solution repeats. You may think of the twelve atoms as physically being arranged in a ring shape.

To make things easier to understand, it is also desirable to switch from the complex "Bloch wave" wave functions to the equivalent real ones. These real wave functions may be found as the real and imaginary parts of the Bloch waves. That is easiest for the free-electron gas, where the Bloch waves are simply complex exponentials; the Euler identity says

$$e^{{\rm i}k_x x} = \cos(k_x x) + {\rm i}\sin(k_x x)$$

So for the free-electron gas, the real wave functions are $\cos(k_x x)$ and $\sin(k_x x)$, ignoring an unimportant normalization constant. As before, the wave number k_x is a measure of the "crystal momentum" $p_{{\rm cm},x} = \hbar k_x$, which is turn related to the electron velocity $v_x^{\rm p}$ through the energy.

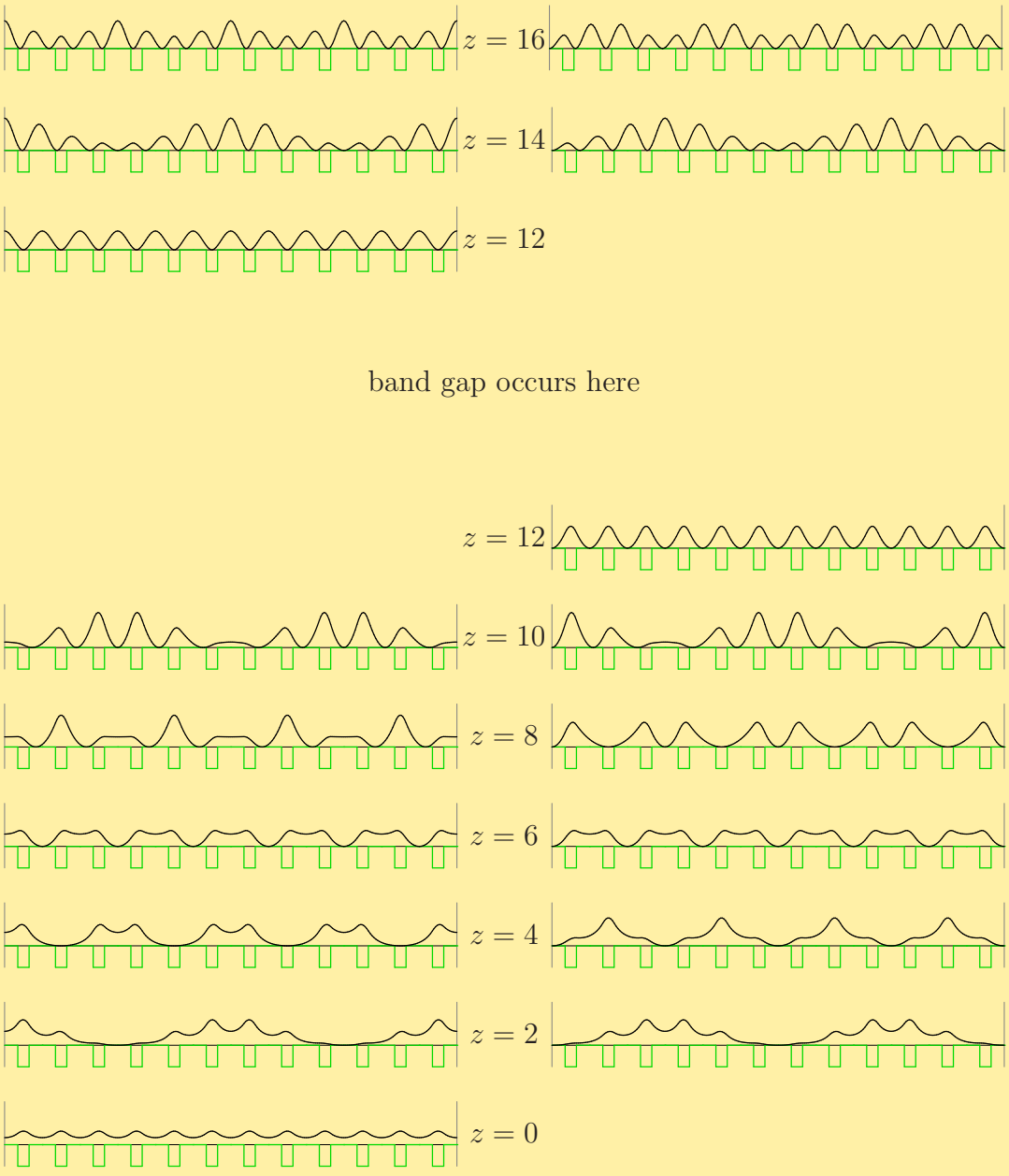


Figure N.2: The 17 real wave functions of lowest energy for a small one-dimensional periodic box with only 12 atomic cells. Black curves show the square wave function, which gives the relative probability of finding the electron at that location.

For the Kronig-Penney model, the real wave functions are more complex than simple sines and cosines, but can be found the same way.

Note that normally there are two different wave functions for each value of the wave number k_x. The one exception is for the ground state of lowest energy where k_x is zero. In terms of the free electron gas, $\sin(0x) = 0$. Zero is not a valid wave function. Remember that the square magnitude of a wave function gives the probability of finding the electron. So if a wave function would be zero, there would not be any chance of finding the electron anywhere. So there would be no electron.

For the free-electron gas, that leaves as only ground state wave function $\cos(0x) = 1$ times some constant. That is just a constant. And since the square of a constant is still a constant, that means that the probability of finding the electron is the same everywhere in the period.

For the Kronig-Penney case, the situation is a less simple. Consider first the ground state, shown in the picture at the bottom of figure N.2. (In figure N.2 the height of a wave function picture illustrates the relative amount of energy of that wave function. So the ground state picture is at the bottom.) The square magnitude of the ground state wave function, shown as the black line, is no longer a constant. It is higher than average at the dips in the potential, at the "atoms." It is lower than average in between "atoms." So the electron is somewhat more likely to be found near an atom that attracts it than in between atoms. The electron reduces its potential energy that way. But it cannot do this without limit; if the electron is only found at the atoms, the reduced uncertainty in position increases the kinetic energy more than the potential energy is lowered. The best compromise is given by the black line at the bottom of figure N.2.

To understand the energy states above the ground state, a key concept of the general mathematical properties of real one-dimensional wave functions is needed: *The more zero crossings in the wave function, the higher the energy.* Qualitatively, the reason is not that hard to understand. The more zero crossings, the more wildly the wave function swings back and forward between positive and negative values, raising the kinetic energy. In figure N.2 the number of zero crossings is listed as z. Note that by squaring the wave functions, the zero crossings become touching zero, not crossing it.

Note further that only even numbers of zero crossings z appear. A periodic wave function must return to the same sign at the end of the period as at the start, and that is only possible if the number of zero crossings is even.

Note next that in almost all cases, there are two *different* wave functions of the *same* energy at a given number of zero crossings. That is because if you have one wave function at a given z, you can simply shift it over by one atomic cell, and you have another wave function of the same energy. This second wave function is *almost* always a different one. In particular, it can only be the same wave function if the z zeros are still in the same place. But if you have 12 atomic

cells and, say, $z = 8$ zeros, then some of the 12 atomic cells must have zeros and other ones not. So the shifted wave function cannot possibly have all its zeros in the same place. So the shifted wave function is a second, different, wave function of the same energy. So there are two different wave functions with the same z and energy.

The only way this can possibly fail, and does, is if each atomic cell has the same number of zeros as its next neighbor. So every atomic cell must have the same number of zeros in it.

That, then, is why it is possible at all that there is only one wave function in the ground state. In the ground state there are no zeros, so every atomic cell has the same number, none. Indeed, looking closer, in the ground state the wave function is identical in every atomic cell. Mathematically, for $k_x = 0$, the exponential part of the Bloch wave is just a trivial constant, making the complete Bloch wave the same for all atomic cells. So shifting the wave function over one atomic cell gives you back the exact same thing.

The next possibility that the shifted wave function does not give a different one occurs when every atomic cell has one zero crossing. For a "crystal" of 12 atomic cells, that requires that there are $z = 12$ zero crossings. This happens when the wavenumber $k_x = \pi/d_x$ where d_x is the atomic cell size. Then the exponential part of the Bloch wave function in any atomic cell is identical to that in the next atomic cell except for a mere minus sign. So the shifted wave function is only different by a minus sign. This means it is physically equivalent to the original one. Not a separate wave function.

But even if shifting the wave function does not give you a second one, still there *must* be two different eigenfunctions for each even number of zeros z greater than zero. In particular, the wave functions in figure N.2 were obtained in two ways. For the wave functions in the left-hand column, it was assumed that the derivative of the wave function is zero at the start of the period (like it is for the $\cos(k_x x)$ free-electron gas solutions). For the wave functions in the right-hand column, it was assumed that the wave function itself is zero at the start of the period (like it is for the $\sin(k_x x)$ free-electron gas solutions). These are two different solutions; they cannot be equivalent because the right-hand wave function has a zero at the start of the period, but the left-hand one does not.

For one, this explains why for the ground state where $z = 0$, there is no right-hand wave function. If you start out with a zero crossing, you must have at least one of them. It also explains why a wave function in the right column is not just the wave function in the left column shifted by an atomic cell. The two wave functions were separately computed. In almost all cases, the right-hand wave function is then a *combination* of the shifted and unshifted left-hand wave functions, the combination that is zero at the start of the period. In about half the cases, that turns out to be the left wave function shifted by a quarter period, in the other half of the cases it is just all different. The energy of the

two solutions is still the same.

But for the special case of $z = 12$ zeros for the 12 atomic cells, figure N.2 shows that the left and right wave functions are physically fundamentally different. In the left-hand wave function, the electron is most likely to be found in the region of high potential energy between atoms. All the peaks in wave function are there. In the right-hand wave function, the electron is most likely to be found in the region of low potential energy at the atoms. The peaks are there. That means that the left hand wave function has a lot more potential energy than the right-hand one. So the two wave functions do *not* have the same energy in this case. We have a band gap when the number of zeros is exactly the same as the number of atoms.

Similarly there will be a band gap at $z = 24$, where there are two zero crossings in each atomic cell, etcetera. The band gaps occur at whole multiples of the number of atomic cells. And there are 12 energy states in each band. For a physically realistic number of atomic cells, call it a million instead of 12, there are a million energy states between band gaps, effectively forming a continuum band between the gaps.

One thing that may still be counter-intuitive is why the right-hand $z = 12$ wave function has higher, rather than much lower energy than the $z = 10$ ones. In particular, the peaks in the right-hand wave function at $z = 12$ are all perfectly aligned with the atomic locations. That greatly reduces the potential energy. But in the $z = 10$ case the peaks are not aligned with the atoms. While the $z = 10$ case has some advantage in kinetic energy with less zero crossings, that advantage is small. That would not be able to explain it if the right-hand $z = 12$ would really have a big advantage in potential energy over the $z = 10$ states.

To see why it is possible, look more closely at the $z = 10$ case in figure N.2. It is true that a significant fraction of the peaks in wave function are in between atoms instead of on top of atoms. However, the physics modulates the *height* of the peaks so that the big peaks are the ones on top of the atoms, and the small peaks the ones in between atoms. That still has the effect that the electron is most likely found at an atom, still greatly reducing the potential energy. That eliminates the apparent advantage of the right-hand $z = 12$ state in potential energy.

Also note that for $z = 10$, the number of zeros is still close to the number of atoms. So the distance between peaks is still almost the same as the distance between atoms, (especially for a million atoms instead of 12). So if, say, a peak is pretty much on top of an atom, the neighboring peaks are too. Therefore the modulation of peak amplitudes can be done in a way that slowly varies along the length of the crystal. So it does not add a big amount to the kinetic energy.

However, because of the modulation, the $z = 10$ wave functions do give up almost all their small kinetic energy advantage compared to the right-hand $z = 12$ case. That means that the two energies become very close together. Since

the change in energy is a measure of the electron propagation velocity $v_x^{\rm p}$, that velocity plunges to zero. Which is exactly what you see in figure N.1.

Similarly for the $z = 14$ wave functions, the peaks are modulated so that the electron is most likely to be found in between atoms, just like for the left-hand $z = 12$ wave function. So the $z = 14$ wave functions have about the same potential energy as the left-hand $z = 12$ one. And because the $z = 12$ wave function is so effective in raising its potential energy, you would expect that the energy difference with the $z = 14$ case would be relatively small, producing small electron velocity. And that is indeed what happens.

So the only finite energy gaps occur when the number of zeros is a whole multiple of the number of atoms. And the gap is between the two states with that number of zeros.

And between the states immediately above and below the gaps, the energy difference is even smaller than elsewhere in the band. That makes the electron velocity $v_x^{\rm p}$ zero at the edges of the bands

Since the wave functions at the edges of the bands have zero propagation velocity, electrons in these states cannot move through the crystal. Now an implicit result of the analysis above is that for these states, a whole multiple of the Bloch wave length must equal double the atomic spacing. The Bloch exponential can change sign going from one atomic cell to the next, then return to the original sign at the next cell, but nothing more. If you train a beam of electrons with a wave length like that onto the crystal, the beam cannot propagate and will be totally reflected. That is in fact a key result of the Bragg reflection theory of wave mechanics, (10.16) in chapter 10.7.2. Thus Bragg theory can provide an intuitive justification for some of the features of the band structure.

If you want to see mathematically that the propagation velocity is indeed zero at the band gaps, and you know linear algebra, you can find the derivation in {D.84}. That also explains how the wave function figures in figure N.2 were made.

N.10 A less fishy story

This note gives a simple model for the emission of a particle like a photon. It is assumed that the emitted particle has a typical quantum wave length λ that is large compared to the typical size R of the atom or nucleus that does the emitting. The purpose of the model is to show that in that case, the particle will very likely come out with zero orbital angular momentum but has some probability of nonzero angular momentum.

First, photon wave functions are messy and not that easy to make sense of, {A.21.7}. The photon would be much simpler if it did not have spin and was nonrelativistic. A reasonable wave function for a hypothetical spinless nonrela-

tivistic photon coming out of the center of the emitter with typical wave length λ would be

$$\psi = \frac{1}{\lambda^{3/2}} f\left(\frac{r^2}{\lambda^2}\right)$$

where r is the distance from the center. (The various factors λ have been added to make the function f independent of the photon wave length λ despite the corresponding spatial scale and the normalization requirement.)

The above wave function has no preferred direction in the emission, making it spherically symmetric. It depends only on the distance r from the center of the emitter. That means that the wave function has zero orbital angular momentum. Recall that zero angular momentum corresponds to the spherical harmonic Y_0^0, which is independent of the angular position, chapter 4.2.

There are various reasons to give why you would want the wave function of a particle coming out of the origin to have zero angular momentum. For one, since it comes out of a featureless point, there should not be a preferred direction. Or in terms of classical physics, if it had angular momentum then it would have to have infinite velocity at the origin. The similar quantum idea is that the relevant wave functions for a particle moving away from the origin, the Hankel functions of the first kind, blow up very strongly at the origin if they have angular momentum, {A.6}. But it is really better to describe the emitted particle in terms of the Bessel functions of the first kind. These have zero probability of the particle being at the origin if the angular momentum is not zero. And a particle should not be created at a point where it has zero probability of being.

Of course, a spherically symmetric quantum wave function also means that the particle is moving away from the emitter equally in all directions. Following the stated ideas of quantum mechanics, this will be true until the position of the particle is "measured." Any macroscopic surroundings cannot reasonably remain uncommitted to exactly where the outgoing particle is for very long.

Now consider the same sort of emission, but from a point in the emitter a bit away from the center. For simplicity, assume the emission point to be at $R\hat{k}$, where R is the typical size of the emitter and $\hat{k}$ is the unit vector along the chosen z-axis. In that case the wave function is

$$\psi = \frac{1}{\lambda^{3/2}} f\left(\frac{(\vec{r} - R\hat{k})^2}{\lambda^2}\right)$$

Using Taylor series expansion, that becomes

$$\psi = \frac{1}{\lambda^{3/2}} f\left(\frac{r^2}{\lambda^2}\right) - \frac{R}{\lambda}\frac{1}{\lambda^{3/2}} f'\left(\frac{r^2}{\lambda^2}\right) 2\frac{r}{\lambda}\frac{z}{r} + \ldots$$

In the second term, z/r is the spherical harmonic Y_1^0, table 4.3. This term has angular momentum quantum number $l = 1$. So there is now uncertainty in momentum. And following the stated ideas of quantum mechanics, the probability

for $l = 1$ is given by the square magnitude of the coefficient of the (normalized) eigenfunction.

That makes the probability for $l = 1$ proportional to $(R/\lambda)^2$. If you carried out the Taylor series to the next order, you would end up with a $(z/r)^2$ term, which, combined with a spherically symmetric contribution, makes up the spherical harmonic Y_2^0. It then follows that the probability for $l = 2$ is of order $(R/\lambda)^4$. And so on. Under the assumed condition that the emitter size R is much less than the quantum wave length λ of the emitted particle, the probabilities for nonzero angular momentum are small and decrease rapidly even further with increasing l.

N.11 Better description of two-state systems

An approximate definition of the states ψ_1 and ψ_2 would make the states $\psi_{\rm L}$ and $\psi_{\rm H}$ only approximate energy eigenstates. But they can be made exact energy eigenfunctions by defining $(\psi_1 + \psi_2)/\sqrt{2}$ and $(\psi_1 - \psi_2)/\sqrt{2}$ to be the exact symmetric ground state and the exact antisymmetric state of second lowest energy. The precise "basic" wave function ψ_1 and ψ_2 can then be reconstructed from that.

Note that ψ_1 and ψ_2 themselves are not energy eigenstates, though they might be so by approximation. The errors in this approximation, even if small, will produce the wrong result for the time evolution. (The small differences in energy drive the *nontrivial* part of the unsteady evolution.)

N.12 Second quantization in other books

The approach to second quantization followed in this book is quite different from what you will find in other basic quantum mechanics or advanced physics books. This book simply sticks to its guns. Right at the beginning, this book said that observable properties are the eigenvalues of Hermitian operators. And that these act on particle wave functions. These same rules are then used to quantize the electromagnetic field.

What other books do is write down various classical wave solutions to Maxwell's equations. Then these books reach deep inside these messy equations, cross out certain coefficients, and scribble in new ones. The new ones have operators in them and undetermined coefficients. The undetermined coefficients are then determined by examining the energy of the wave and comparing it with a harmonic oscillator, as analyzed using quantum field theory.

This book, however, greatly dislikes writing down classical solutions. A general student may not be familiar with these solutions. Or have long forgotten them. And it seems quite doubtful that even physics students are really

familiar with the messy electric and magnetic multipole fields of classical electromagnetics. The approach in this book is to skip classical physics and give a self-contained and reasonable quantum derivation wherever possible. (Which means almost always.)

This book detests reaching into the middle of equations known to be wrong, and then crossing out things and writting in new things, all the while waving your hands a lot. The method of science is to make certain fundamental assumptions and then take them to their logical conclusion, whatever it may be. Not messing around until you get something that seems the right answer. And a book on science should showcase the methods of science.

Then there is the problem that the classical waves are inherently time-dependent. The Schrödinger approach, however, is to put the time dependence in the wave function. For good reasons. That means that starting from the classical waves, you have two options, both ugly. You can suddenly switch to the Heisenberg representation, which is what everybody does. Or you can try to unextract the time dependence and put it on an explicit wave function.

And things get even uglier because the entire approach depends essentially on a deep familiarity with a different problem; the quantum-field description of the harmonic oscillator.

In fact, it may be noted that in early versions, this book did really try to give an understandable description of second quantization using the usual approach. The result was an impenetrable mess.

N.13 Combining angular momentum factors

Angular momenta from different sources can be combined using the Clebsch-Gordan coefficients of chapter 12.7. For example, you can combine orbital and spin angular momenta of a particle that way, or the angular momenta of different particles.

But sometimes you need to multiply angular momentum states of the same source together. For example, you may need to figure out a product of spherical harmonics, chapter 4.2.3, like

$$Y_{l_1}^{m_1} Y_{l_2}^{m_2}$$

where the position coordinates refer to the same particle. If you multiply a few examples from table 4.3 together, you quickly see that the combinations are not given by the Clebsch-Gordan coefficients of figure 12.6.

One way to understand the angular momentum properties of the above product qualitatively is to consider the case that the second spherical harmonic takes the coordinates of an imagined second particle. Then you can use the normal procedures to figure out the properties of the two-particle system. And you can get the corresponding properties of the original one-particle system by re-

stricting the two-particle wave function coordinates to the subset in which the particle coordinates are equal.

Note in doing so that angular momentum properties are directly related to the effect of coordinate system rotations, {A.19}. Coordinate system rotations maintain equality of particle coordinates; they stay within the subset. But inner products for the two-particle system will obviously be different from those for the reduced one-particle system.

And Clebsch-Gordan coefficients are in fact inner products. They are the inner product between the corresponding horizontal and vertical states in figures 12.5 and 12.6. The correct coefficients for the product above are still related to the Clebsch-Gordan ones, though you may find them in terms of the equivalent Wigner "3j" coefficients.

Wigner noticed a number of problems with the Clebsch-Gordan coefficients:

1. Clebsch and Gordan, and not Wigner, get credit for them.
2. They are far too easy to type.
3. They have an intuitive interpretation.

So Wigner changed the sign on one of the variables, took out a common factor, and reformatted the entire thing as a matrix. In short

$$\begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \equiv \frac{(-1)^{j_1-j_2-m_3}}{\sqrt{2j_3+1}} \langle j_3 \, -m_3 || j_1 \, m_1 \rangle |j_2 \, m_2\rangle \tag{N.1}$$

Behold, the spanking new "Wigner 3j symbol." Thus Wigner succeeded by his hard work in making physics a bit more impenetrable still than before. A big step for physics, a small step back for mankind.

The most important thing to note about this symbol/coefficient is that it is zero unless

$$m_1 + m_2 + m_3 = 0 \quad \text{and} \quad |j_1 - j_2| \leqslant j_3 \leqslant j_1 + j_2$$

The right-hand conditions are the so-called triangle inequalities. The ordering of the j-values does not make a difference in these inequalities. You can swap the indices 1, 2, and 3 arbitrarily around.

If all three m values are zero, then the symbol is zero if the sum of the j values is odd. If the sum of the j values is even, the symbol is not zero unless the triangle inequalities are violated.

If you need an occasional value for such a symbol that you cannot find in figure 12.5 and 12.6 or more extensive tables elsewhere, there are convenient calculators on the web, ⟦16⟧. There is also software available to evaluate them. Note further that {D.65} gives an explicit expression.

In literature, you may also encounter the so-called Wigner "6j" and "9j" symbols. They are typically written as

$$\begin{Bmatrix} j_1 & j_2 & j_3 \\ l_1 & l_2 & l_3 \end{Bmatrix} \qquad \begin{Bmatrix} j_{11} & j_{12} & j_{13} \\ j_{21} & j_{22} & j_{23} \\ j_{31} & j_{32} & j_{33} \end{Bmatrix}$$

They appear in the combination of angular momenta. If you encounter one, there are again calculators on the web. The most useful thing to remember about 6j symbols is that they are zero unless each of the four triads $j_1 j_2 j_3$, $j_1 l_2 l_3$, $l_1 j_2 l_3$, and $l_1 l_2 j_3$ satisfies the triangle inequalities.

The 9j symbol changes by at most a sign under swapping of rows, or of columns, or transposing. It can be expressed in terms of a sum of 6j symbols. There are also 12j symbols, if you cannot get enough of them.

These symbols are needed to do advanced computations but these are far outside the scope of this book. And they are very abstract, definitely not something that the typical engineer would want to get involved in in the first place. All that can be done here is to mention a few key concepts. These might be enough keep you reading when you encounter them in literature. Or at least give a hint where to look for further information if necessary.

The basic idea is that it is often necessary to know how things change under rotation of the axis system. Many derivations in classical physics are much simpler if you choose your axes cleverly. However, in quantum mechanics you face problems such as the fact that angular momentum vectors are not normal vectors, but are quantized. Then the appropriate way of handling rotations is through the so-called Wigner-Eckart theorem. The above symbols then pop up in various places.

For example, they allows you to do such things as figuring out the derivatives of the harmonic polynomials $\mathcal{Y}_l^m$ of table 4.3, and to define the vector spherical harmonics $\vec{Y}_{JlM}$ that generalize the ordinary spherical harmonics to vectors. A more advanced treatment of vector bosons, {A.20}, or photon wave functions of definite angular momentum, {A.21.7}, would use these. And so would a more solid derivation of the Weisskopf and Moszkowski correction factors in {A.25.8}. You may also encounter symbols such as $\mathcal{D}^{(j)}_{m'm}$ for matrix elements of finite rotations.

All that will be done here is give the derivatives of the harmonic polynomials $r^l Y_l^m$, since that formula is not readily available. Define the following complex coordinates x_μ for $\mu = -1,\ 0,\ 1$:

$$\mu = -1: \quad x_{-1} = \frac{x - \mathrm{i}y}{\sqrt{2}} \qquad \mu = 0: \quad x_0 = z \qquad \mu = 1: \quad x_1 = -\frac{x + \mathrm{i}y}{\sqrt{2}}$$

Then

$$\begin{aligned} \frac{\partial r^l Y_l^m}{\partial x_\mu} &= (-1)^{\mu+1} \sqrt{\frac{l(2l+1)^2}{(2l-1)}} \langle l-1\ m-\mu || l\ m \rangle |1\ -\mu\rangle r^{l-1} Y_{l-1}^{m-\mu} \\ &= C_{\mu l m} r^{l-1} Y_{l-1}^{m-\mu} \end{aligned}$$

where the inner product of kets is a Clebsch-Gordan coefficient and

$$C_{-1lm} = \sqrt{\frac{(2l+1)(l-m)(l-m-1)}{2(2l-1)}}$$

$$C_{0lm} = \sqrt{\frac{(2l+1)2(l+m)(l-m)}{2(2l-1)}}$$

$$C_{1lm} = \sqrt{\frac{(2l+1)(l+m)(l+m-1)}{2(2l-1)}}$$

or zero if the final magnetic quantum number is out of bounds.

If just having a rough idea of what the various symbols are is not enough, you will have to read up on them in a book like [13]. There are a number of such books, but this particular book has the redeeming feature that it lists some practical results in a usable form. Some highlights: general expression for the Clebsch-Gordan coefficients on p. 45; *wrong* definition of the 3j coefficient on p. 46, (one of the rare mistakes; correct is above), symmetry properties of the 3j symbol on p. 47; 3j symbols with all m values zero on p. 50; list of alternate notations for the Clebsch-Gordan and 3j coefficients on p.52, (yes, of course it is a long list); integral of the product of three spherical harmonics, like in the electric multipole matrix element, on p. 63; the correct expression for the product of two spherical harmonics with the same coordinates, as discussed above, on p. 63; effect of nuclear orientation on electric quadrupole moment on p. 78; *wrong* derivatives of spherical harmonics times radial functions on p. 69, 80, (another 5 rare mistakes in the second part of the expression alone, see above for the correct expression for the special case of harmonic polynomials); plane wave in spherical coordinates on p. 81. If you do try to read this book, note that γ stands for "other quantum numbers," as well as the Euler angle. That is used well before it is defined on p. 33. If you are not psychic, it can be distracting.

N.14 The electric multipole problem

There is a big problem with electric multipole transitions in nuclei. Electric multipole transitions arise from a matrix element of the form

$$H_{21} = \sum_i \langle \psi_{\mathrm{L}} | \vec{A}_i^{\mathrm{E}\ell *} \cdot \widehat{\vec{p}}_i | \psi_{\mathrm{H}} \rangle$$

Here ψ_{L} is the final atomic or nuclear state and ψ_{H} the initial one. The sum is over the atomic or nuclear particles, with i the particle index. Also

$$\widehat{\vec{p}}_i = \frac{\hbar}{\mathrm{i}} \nabla_i \qquad \vec{A}_i^{\mathrm{E}\ell} = \nabla_i \times \vec{r}_i \times \nabla_i j_\ell(kr_i) Y_\ell^m(\theta_i, \phi_i)$$

where j_ℓ is a spherical Bessel function and Y_ℓ^m a spherical harmonic.

Under the approximation that the atomic or nuclear size is small compared to the wave length of the emitted or absorbed photon, this may be approximated.

In that approximation. the single-particle electric matrix element is commonly described as proportional to

$$\langle \psi_{\rm L} | r_i^\ell Y_{\ell,i}^m | \psi_{\rm H} \rangle \qquad Y_{\ell,i}^m \equiv Y_\ell^m(\theta_i, \phi_i)$$

where $r_i^\ell Y_{\ell,i}^m$ is a harmonic polynomial. However, the correct inner product is

$$\langle \psi_{\rm L} | r_i^\ell Y_{\ell,i}^m + [V/\hbar\omega, r_i^\ell Y_{\ell.i}^m] | \psi_{\rm H} \rangle$$

Here $\hbar\omega$ is the energy of the photon. The additional commutator term is not mentioned in any other basic textbook on nuclei that this author knows of. A derivation is in {D.43}.

If the potential depends only on position, the commutator is zero, and there is no problem. That is a valid approximation for the outer electrons in atoms. But nuclear potentials include significant momentum terms. These do not commute. One example is the spin-orbit term in the shell model. Now consider the size of the commutator term above compared to the first term. For a ballpark, note that it does not make much difference whether the orbital angular momentum in the spin-orbit term acts on the wave function $\psi_{\rm H}$ or on the harmonic polynomial. So the relative size of the commutator term ballparks to the ratio of the spin-orbit energy to the photon energy. That ratio can be big. For example, in the so-called "islands of isomerism" transitions, one state has enough spin-orbit energy to cross a major shell boundary. But the energy release $\hbar\omega$ stays within the same shell and could be quite small.

As a check on this ballpark, consider the simplest possible electric multipole transition:

$$\psi_{\rm L} = R_{\rm L}(r_i) Y_{0,i}^0 \uparrow \qquad \psi_{\rm H} = R_{\rm H}(r_i) Y_{\ell,i}^\ell \uparrow \qquad m = \ell \qquad V = V_0 - V_{\rm so} f(r_i) \frac{1}{\hbar^2} \widehat{\vec{L}}_i \cdot \widehat{\vec{S}}_i$$

Here the part V_0 of the potential V represents terms that only depend on position, or do not involve the position coordinates of particle i. This part commutes with the harmonic polynomial. Also f is a function of radial position of order 1. Then the constant $V_{\rm so}$ gives the magnitude of the spin orbit energy. In the above case, only the z components of the angular momentum operators give a contribution. Then it is easily seen that the commutator term produces a contribution that is larger in magnitude than the other term by a factor of order $V_{\rm so}\ell/\hbar\omega$.

Note also that the effect is especially counter-intuitive for electric dipole transitions. It would seem logical to think that such transitions could be approximately described by a straightforward interaction of the particles with the electric field. However, the commutator need not be zero. So the electric field could be dwarfed by a larger additional field. That field is then a consequence of the fact that quantum mechanics uses the vector potential rather than the classical electric field.

Which brings up the next problem. The commutator term will not be there if you use the gauge property of the electromagnetic field to approximate the leading order electric field through an electrostatic potential. So should the commutator be there or not to get the best solution? As far as this author can see, there is no way to tell. And certainly for higher multipole orders you cannot even ignore the problem by using the electrostatic potential.

(Note that the real problem is the fact that you get different answers depending on how you select the gauge. If the nuclear Hamiltonian (A.169) respected the quantum gauge property of {A.19.5}, the commutator would be zero. That can be seen by substituting in the gauge property: it shows that for any particle the potential must commute with $e^{-\mathrm{i}q_i\chi_i/\hbar}$, and the exponential is a completely arbitrary function of the particle coordinates. But the fact that the commutator *should* be zero does not take away the fact that it *is not* zero. Presumably, if you described the nucleus in terms of the individual quarks, you could write down a potential that respected the gauge property. But using quarks is not an option. The reality is that you must use protons and neutrons. And in those terms, a potential that does a decent job of describing the nucleus will simply not respect the gauge property. Yes, this does suggest that describing gamma decay using protons and neutrons instead of quarks is an inherently fishy procedure.)

This author knows not a single reference that gives a decent description of the above issue. Many simply do not mention the problem at all and just omit the commutator. Some simply state that the potential is assumed to depend on the particle positions only, like [33]. Surely it ought to be mentioned explicitly that the leading electric multipole operator as listed may well be no good at all? If it is listed, people will assume that is is meaningful unless stated otherwise. As [33] notes, "significant information regarding nuclear wave functions can be obtained from a comparison of experimental γ-decay transition probabilities with theoretical values calculated on basis of specific models of the nucleus." If one is not aware of the possibility of the additional commutator as a leading order effect, one might incorrectly conclude that a nuclear wave function is poor where the real problem is the ignored commutator.

At least [11, p.9-172] and [5] can be read to say that there might be a nontrivial problem. The original relativistic derivation of Stech, [44], mentions the issue, but no ballpark is given. (It is however noted that the spin-orbit term might be significant for magnetic transitions. The purely nonrelativistic analysis used here does not show such an effect. The present author suspects that the difference is that the relativistic derivation of Stech inherently assumes that the gauge property is valid. Surely there must be ways to do the relativistic analysis such that, in say the electric dipole case, both the results with and without commutator are reproduced.)

To be sure, the author no longer believes that this is a potential explanation why E1 transitions are so much slower than the Weisskopf estimate. The deviations in chapter 14.20.5 figures 14.63 and 14.64 seem much too big and

systematic to be explained by this mechanism. And it does not seem to address the concerns about mixed transitions that are mentioned there.

N.15 A tenth of a googol in universes

There is an oft-cited story going around that the many worlds interpretation implies the existence of 10^{99} worlds, and this number apparently comes from Everett, III himself. It is often used to argue that the many-worlds interpretation is just not credible. However, the truth is that the existence of infinitely many worlds, (or practically speaking infinitely many of them, maybe, if space and time would turn out to be discrete and finite), is a basic requirement of quantum mechanics itself, regardless of interpretation. Everett, III cannot be blamed for that, just for coming up with the ludicrous number of 10^{99} to describe infinity.

N.16 A single Slater determinant is not exact

The simplest example that illustrates the problem with representing a general wave function by a single Slater determinant is to try to write a general two-variable function $F(x,y)$ as a Slater determinant of two functions f_1 and f_2. You would write

$$F(x,y) = \frac{a}{\sqrt{2}}\Big(f_1(x)f_2(y) - f_2(x)f_1(y)\Big)$$

A general function $F(x,y)$ cannot be written as a combination of the *same* two functions $f_1(x)$ and $f_2(x)$ at *every* value of y. However well chosen the two functions are.

In fact, for a general antisymmetric function F, a single Slater determinant can get F right at only two nontrivial values $y = y_1$ and $y = y_2$. (Nontrivial here means that functions $F(x,y_1)$ and $F(x,y_2)$ should not just be multiples of each other.) Just take $f_1(x) = F(x,y_1)$ and $f_2(x) = F(x,y_2)$. You might object that in general, you have

$$F(x,y_1) = c_{11}f_1(x) + c_{12}f_2(x) \qquad F(x,y_2) = c_{21}f_1(x) + c_{22}f_2(x)$$

where c_{11}, c_{12}, c_{21}, and c_{22} are some constants. (They are f_1 or f_2 values at y_1 or y_2, to be precise). But if you plug these two expressions into the Slater determinant formed with $F(x,y_1)$ and $F(x,y_2)$ and multiply out, you get the Slater determinant formed with f_1 and f_2 within a constant, so it makes no difference.

If you add a second Slater determinant, you can get F right at two more y values y_3 and y_4. Just take the second Slater determinant's functions to be $f_1^{(2)} = \Delta F(x,y_3)$ and $f_2^{(2)} = \Delta F(x,y_4)$, where ΔF is the deviation between the

true function and what the first Slater determinant gives. Keep adding Slater determinants to get more and more y-values right. Since there are infinitely many y-values to get right, you will in general need infinitely many determinants.

You might object that maybe the deviation ΔF from the single Slater determinant must be zero for some reason. But you can use the same ideas to explicitly construct functions F that show that this is untrue. Just select two arbitrary but different functions f_1 and f_2 and form a Slater determinant. Now choose two locations y_1 and y_2 so that $f_1(y_1), f_2(y_1)$ and $f_1(y_2), f_2(y_2)$ are not in the same ratio to each other. Then add additional Slater determinants whose functions $f_1^{(2)}, f_2^{(2)}, f_1^{(3)}, f_2^{(3)}, \ldots$ you choose so that they are zero at y_1 and y_2. The so constructed function F is different from just the first Slater determinant. However, if you try to describe this F by a single determinant, then it could only be the first determinant since that is the only single determinant that gets y_1 and y_2 right. So a single determinant cannot get F right.

N.17 Generalized orbitals

This note has a brief look at generalized orbitals of the form

$$\psi_n^{\rm p}(\vec{r}) = \psi_{n+}^{\rm s}(\vec{r})\uparrow(S_z) + \psi_{n-}^{\rm s}(\vec{r})\downarrow(S_z).$$

For such orbitals, the expectation energy can be worked out in exactly the same way as in {D.52}, except without simplifying the spin terms. The energy is

$$\begin{aligned}
\langle E \rangle \;=\; & \sum_{n=1}^{I} \left\langle \psi_n^{\rm p} \middle| h^{\rm e} \middle| \psi_n^{\rm p} \right\rangle \\
& + \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n}\neq n}}^{I} \left\langle \psi_n^{\rm p} \psi_{\underline{n}}^{\rm p} \middle| v^{\rm ee} \middle| \psi_n^{\rm p} \psi_{\underline{n}}^{\rm p} \right\rangle \\
& - \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n}\neq n}}^{I} \left\langle \psi_n^{\rm p} \psi_{\underline{n}}^{\rm p} \middle| v^{\rm ee} \middle| \psi_{\underline{n}}^{\rm p} \psi_n^{\rm p} \right\rangle
\end{aligned}$$

To multiply out to the individual spin terms, it is convenient to normalize the spatial functions, and write

$$\psi_n^{\rm p} = c_{n+}\psi_{n+,0}^{\rm s}\uparrow + c_{n-}\psi_{n-,0}^{\rm s}\downarrow,$$

$$\langle \psi_{n+,0}^{\rm s} | \psi_{n+,0}^{\rm s} \rangle = \langle \psi_{n-,0}^{\rm s} | \psi_{n-,0}^{\rm s} \rangle = 1, \quad |c_{n+}|^2 + |c_{n-}|^2 = 1$$

In that case, the expectation energy multiplies out to

$$\begin{aligned}
\langle E \rangle \;=\; & \sum_{n=1}^{I} \left\langle \psi^{\rm s}_{n+,0} \middle| h^{\rm e} \middle| \psi^{\rm s}_{n+,0} \right\rangle |c_{n+}|^2 + \sum_{n=1}^{I} \left\langle \psi^{\rm s}_{n-,0} \middle| h^{\rm e} \middle| \psi^{\rm s}_{n-,0} \right\rangle |c_{n-}|^2 \\
& + \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n}\neq n}}^{I} \Big(\left\langle \psi^{\rm s}_{n+,0}\psi^{\rm s}_{\underline{n}+,0} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{n+,0}\psi^{\rm s}_{\underline{n}+,0} \right\rangle \\
& \qquad\qquad - \left\langle \psi^{\rm s}_{n+,0}\psi^{\rm s}_{\underline{n}+,0} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}+,0}\psi^{\rm s}_{n+,0} \right\rangle \Big) |c_{n+}|^2 |c_{\underline{n}+}|^2 \\
& + \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n}\neq n}}^{I} 2 \left\langle \psi^{\rm s}_{n+,0}\psi^{\rm s}_{\underline{n}-,0} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{n+,0}\psi^{\rm s}_{\underline{n}-,0} \right\rangle |c_{n+}|^2 |c_{\underline{n}-}|^2 \\
& + \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n}\neq n}}^{I} \Big(\left\langle \psi^{\rm s}_{n-,0}\psi^{\rm s}_{\underline{n}-,0} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{n-,0}\psi^{\rm s}_{\underline{n}-,0} \right\rangle \\
& \qquad\qquad - \left\langle \psi^{\rm s}_{n-,0}\psi^{\rm s}_{\underline{n}-,0} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}-,0}\psi^{\rm s}_{n-,0} \right\rangle \Big) |c_{n-}|^2 |c_{\underline{n}-}|^2 \\
& - \tfrac{1}{2} \sum_{n=1}^{I} \sum_{\substack{\underline{n}=1 \\ \underline{n}\neq n}}^{I} 2\Re \Big(\left\langle \psi^{\rm s}_{n+,0}\psi^{\rm s}_{\underline{n}-,0} \middle| v^{\rm ee} \middle| \psi^{\rm s}_{\underline{n}+,0}\psi^{\rm s}_{n-,0} \right\rangle c^*_{n+} c_{n-} c^*_{\underline{n}-} c_{\underline{n}+} \Big)
\end{aligned}$$

where $\Re$ stands for the real part of its argument.

Now assume you have a normal unrestricted Hartree-Fock solution, and you try to lower its ground-state energy by selecting, for example, a spin-up orbital $\psi^{\rm s}_m\uparrow \equiv \psi^{\rm s}_{m+,0}\uparrow$ and adding some amount of spin down to it. First note then that the final sum above is zero, since at least one of c_{n+}, c_{n-}, $c_{\underline{n}-}$, and $c_{\underline{n}+}$ must be zero: all states except m are still either spin-up or spin-down, and m cannot be both n and $\underline{n} \neq n$. With the final sum gone, the energy is a linear function of $|c_{m-}|^2$, with $|c_{m+}|^2 = 1 - |c_{m-}|^2$. The minimum energy must therefore occur for either $|c_{m-}|^2 = 0$, the original purely spin up orbital, or for $|c_{m-}|^2 = 1$. (The latter case means that the unrestricted solution with the opposite spin for orbital m must have less energy, so that the spin of orbital m was incorrectly selected.) It follows from this argument that for correctly selected spin states, the energy cannot be lowered by replacing a single orbital with a generalized one.

Also note that for small changes, $|c_{m-}|^2$ is quadratically small and can be ignored. So the variational condition of zero change in energy is satisfied for all small changes in orbitals, even those that change their spin states. In other

words, the unrestricted solutions are solutions to the full variational problem $\delta\langle E\rangle = 0$ for generalized orbitals as well.

Since these simple arguments do not cover finite changes in the spin state of more than one orbital, they do not seem to exclude the possibility that there might be additional solutions in which two or more orbitals are of mixed spin. But since either way the error in Hartree-Fock would be finite, there may not be much justification for dealing with the messy problem of generalized orbitals with dubious hopes of improvement. Procedures already exist that guarantee improvements on standard Hartree-Fock results.

N.18 "Correlation energy"

The error in Hartree-Fock is due to the single-determinant approximation only. A term like "Hartree-Fock error" or "single-determinantal error" is therefore both precise, and immediately understandable by a general audience.

However, it is called "correlation energy," and to justify that term, it would have to be both clearer and equally correct mathematically. It fails both requirements miserably. The term correlation energy is clearly confusing and distracting for nonspecialist. But in addition, there does not seem to be any theorem that proves that an independently defined correlation energy is identical to the Hartree-Fock single determinant error. That would not just make the term correlation energy disingenuous, it would make it wrong.

Instead of finding a rigorous theorem, you are lucky if standard textbooks, e.g,. [30, 34, 46] and typical web references, offer a vague qualitative story why Hartree-Fock underestimates the repulsions if a pair of electrons gets very close. That is a symptom of the disease of having an incomplete function representation, it is not the disease itself. Low-parameter function representations have general difficulty with representing localized effects, whatever their physical source. If you make up a system where the Coulomb force vanishes both at short and at long distance, such correlations do not exist, and Hartree-Fock would still have a finite error.

The kinetic energy is not correct either; what is the correlation in that? Some sources, like [30] and web sources, seem to suggest that these are "indirect" results of having the wrong correlation energy, whatever correlation energy may be. The idea is apparently, *if* you would have the electron-electron repulsions exact, you would compute the correct kinetic energy too. That is just like saying, *if* you computed the correct kinetic energy term, you would compute the correct potential too, so let's rename the Hartree-Fock error "kinetic energy interaction."

Even *if* you computed the potential energy correctly, you would still have to convert the wave function to single-determinantal form before evaluating the kinetic energy, *otherwise it is not Hartree-Fock,* and that would produce

a finite error. Phrased differently, there is absolutely no way to get a general wave function correct with a finite number of single-electron functions, *whatever* corrections you make to the potential energy.

Szabo and Ostlund [46, p. 51ff,61] state that it is called correlation energy since "the motion of electrons with opposite spins is not correlated within the Hartree-Fock approximation." That is incomprehensible, for one thing since it seems to suggest that Hartree-Fock is exact for excited states with all electrons in the same spin state, which would be ludicrous. In addition, the electrons do not have motion; a stationary wave function is computed, and they do not have spin; all electrons occupy all the states, spin up and down. It is the orbitals that have spin, and the spin-up and spin-down orbitals are most definitely correlated.

However, the authors do offer a "clarification;" they take a Slater determinant of two opposite spin orbitals, compute the probability of finding the two electrons at given positions and find that it is correlated (unless the spatial orbitals are equal). They then declare that these *correlated* positions mean that the "motion" of the two electrons is *un*correlated.

The unrestricted Hartree-Fock solution of the dissociated hydrogen molecule is of this type. This solution was discussed in the introductory section 9.3.1. Since if one electron is around the left proton, the other is around the right one, and vice versa, normal people would call the positions of the electrons strongly correlated. And even while "motion" is not defined on quantum scales, to any engineer it would seem ludicrous to claim that strongly correlated positions would produce uncorrelated "motion" if "motion" existed.

Koch and Holthausen, [30, pp.22-23], address the same two electron example as Szabo and Ostlund, but do not have the same problem of finding the electron probabilities correlated. For example, if the spin-independent probability of finding the electrons at positions $\vec{r}_1$ and $\vec{r}_2$ in the dissociated hydrogen molecule is

$$\tfrac{1}{2}|\psi_{\rm l}(\vec{r}_1)|^2|\psi_{\rm r}(\vec{r}_2)|^2 + \tfrac{1}{2}|\psi_{\rm r}(\vec{r}_1)|^2|\psi_{\rm l}(\vec{r}_2)|^2$$

then, Koch and Holthausen explain to us, the second term must be the same as the first. After all, if the two terms were different, the electrons would be distinguishable: electron 1 would be the one that selected $\psi_{\rm l}$ in the first term that Koch and Holthausen wrote down in their book. So, the authors conclude, the second term above is the same as the first, making the probability of finding the electrons equal to twice the first term, $|\psi_{\rm l}(\vec{r}_1)|^2|\psi_{rmr}(\vec{r}_2)|^2$. That is an uncorrelated product probability.

However, the assumption that electrons are indistinguishable with respect to mathematical formulae in books is highly controversial. Many respected references, and this book too, only see an empirical requirement that the *wave function*, *not books*, is antisymmetric with respect to exchange of any two electrons. And the wave function *is* antisymmetric even when the two terms above are not the same.

Wikipedia, [[21]], Hartree-Fock entry June 2007, lists electron correlation, (defined here vaguely as "effects" arising from from the mean-field approximation, i.e. using the same $v^{\rm HF}$ operator for all electrons) as an approximation made *in addition* to using a single Slater determinant. Sorry, but Hartree-Fock gives the best single-determinantal approximation; there is *no* additional approximation made. The mean "field" approximation is a consequence of the single determinant, not an additional approximation. Then this reference proceeds to declare this correlation energy the most important of the set, in other words, more important that the single-determinant approximation! And again, even if the potential energy *was* computed exactly, instead of using the $v^{\rm HF}$ operator, and only the kinetic energy was computed using a Slater determinant, there would still be a finite error. It would therefore appear then that the name correlation energy is sufficiently impenetrable and poorly defined that even the experts cannot necessarily figure it out.

Consider for a second the ground state of two electrons around a massive nucleus. Because of the strength of the nucleus, the Coulomb interaction between the two electrons can to first approximation be ignored. A reader of the various vague qualitative stories listed above may then be forgiven for assuming that Hartree-Fock should not have any error. But only the unrestricted Hartree-Fock solution with those nasty, "uncorrelated" (true in this case), opposite-spin "electrons" (orbitals) is the one that gets the energy right. A unrestricted solution in terms of those perfect, correlated, aligned-spin "electrons" gets the energy all wrong, since one orbital will have to be an excited one. In short the "correlation energy" (error in energy) that, we are told, is due to the "motion" of electrons of opposite spins not being "correlated" is in this case 100% due to the motion of aligned-spin orbitals being correlated. Note that both solutions get the spin wrong, but we are talking about energy.

And what happened to the word "error" in "correlation energy error?" If you did a finite difference or finite element computation of the ground state, you would not call the error in energy "truncation energy;" it would be called "truncation error" or "energy truncation error." Why does one suspect that the appropriate and informative word "error" did not sound "hot" enough to the physicists involved?

Many sources refer to a reference, (Löwdin, P.-E., 1959, Adv. Chem. Phys., 2, 207) instead of providing a solid justification of this widely-used key term themselves. If one takes the trouble to look up the reference, does one find a rigorously defined correlation energy and a proof it is identical in magnitude to the Hartree-Fock error?

Not exactly. One finds a vague qualitative story about some perceived "holes" whose mathematically rigorous definition remains restricted to the center point of one of them. However, the lack of a defined hole size is not supposed to deter the reader from agreeing wholeheartedly with all sorts of claims about the size of their effects. Terms like "main error,", "small error," "large correla-

tion error" (qualified by "certainly"), "vanish or be very small," (your choice), are bandied around, *even though there is no small parameter that would allow any rigorous mathematical definition of small or big.*

Then the author, who has already noted earlier that the references cannot agree on what the heck correlation energy is supposed to mean in the first place, states "In order to get at least a formal definition of the problem, . . . " and proceeds to *redefine* the Hartree-Fock error to be the "correlation energy." In other words, since correlation energy at this time seems to be a pseudo-scientific concept, let's just cross out the correct name Hartree-Fock error, and write in "correlation energy!"

To this author's credit, he does keep the word error in "correlation error in the wave function" instead of using "correlation wave function." But somehow, that particular term does not seem to be cited much in literature.

N.19 Ambiguities in electron affinity

The International Union of Pure and Applied Chemistry (IUPAC) Gold Book defines electron affinity as "Energy required to detach an electron from the singly charged negative ion [. . .] The equivalent more common definition is the energy released ($E_{\text{initial}} - E_{\text{final}}$) when an additional electron is attached to a neutral atom or molecule." This is also the definition given by Wikipedia. Chemguide says "The first electron affinity is the energy released when 1 mole of gaseous atoms each acquire an electron to form 1 mole of gaseous 1- ions." HyperPhysics says "The electron affinity is a measure of the energy change when an electron is added to a neutral atom to form a negative ion." Encyclopedia Brittanica says "in chemistry, the amount of energy liberated when an electron is added to a neutral atom to form a negatively charged ion." Chemed.chem.purdue.edu says "The electron affinity of an element is the energy given off when a neutral atom in the gas phase gains an extra electron to form a negatively charged ion."

Another definition that can be found: "Electron affinity is the energy released when an electron is added to the valence shell of a gas-phase atom." Note the additional requirement here that the electron be added to the *valence shell* of the atom. It may make a difference.

First note that it is not self-evident that a stable negative ion exists. Atoms, even inert noble gasses, can be weakly bound together by Van der Waals/London forces. You might think that similarly, a distant electron could be weakly bound to an atom or molecule through the dipole strength it induces in the atom or molecule. The atom's or molecule's electron cloud would move a bit away from the distant electron, allowing the nucleus to exert a larger attractive force on the distant electron than the repulsive force by the electron cloud. Remember that according to the variational principle, the energy of the atom or molecule

does not change due to small changes in wave function, while the dipole strength does. So the electron would be weakly bound.

It sounds logical, but there is a catch. A theoretical electron at rest at infinity would have an infinitely large wave function blob. If it moves slightly towards the attractive side of the dipole, it would become somewhat localized. The associated kinetic energy that the uncertainty principle requires, while small at large distances, still dwarfs the attractive force by the induced dipole which is still smaller at large distances. So the electron would not be bound. Note that if the atom or molecule itself already has an inherent dipole strength, then if you ballpark the kinetic energy, you find that for small dipole strength, the kinetic energy dominates and the electron will not be bound, while for larger dipole strength, the electron will move in towards the electron cloud with increasing binding energy, presumably until it hits the electron cloud.

In the case that there is no stable negative ion, the question is, what to make of the definitions of electron affinity above. If there is a requirement that the additional electron be placed in the valence shell, there would be energy needed to do so for an unstable ion. Then the electron affinity would be negative. If there is however no requirement to place the electron in the valence shell, you could make the negative value of the electron affinity arbitrarily small by placing the electron in a sufficiently highly-excited state. Then there would be no meaningful value of the electron affinity, except maybe zero.

Various reputed sources differ greatly about what to make of the electron affinities if there is no stable negative ion. The CRC Handbook of Chemistry and Physics lists noble gasses, metals with filled s shells, and nitrogen all as "not stable" rather than giving a negative electron affinity for them. That seems to agree with the IUPAC definition above, which does not require a valence shell position. However, the Handbook does give a small negative value for ytterbium. A 2001 professional review paper on electron affinity mentioned that it would not discuss atoms with negative electron affinities, seemingly implying that they do exist.

Quite a lot of web sources list specific negative electron affinity values for atoms and molecules. For example, both Wikipedia and HyperPhysics give specific negative electron affinity values for benzene. Though one web source based on Wikipedia (!) claims the opposite.

Also note that references, like Wikipedia and HyperPhysics, differ over how the sign of electron affinity should be defined, making things even more confusing. Wikipedia however agrees with the IUPAC Gold Book on this point: if a stable ion exist, there is a positive affinity. Which makes sense; if you want to specify a negative value for a stable ion, you should not give it the name "affinity."

Wikipedia (July 2007) also claims: "All elements have a positive electron affinity, but older texts mistakenly report that some elements such as inert gases have negative [electron affinity], meaning they would repel electrons. This

is not recognized by modern chemists." However, this statement is very hard to believe in view of all the authoritative sources, like the CRC Handbook above, that explicitly claim that various elements do not form stable ions, and often give explicit negative values for the electron affinity of various elements. If the 2007 Handbook would after all these years still misstate the affinity of many elements, would not by now a lot of people have demanded their money back? It may be noted that Wikipedia lists Ytterbium as blank, and the various elements listed as not stable by the CRC handbook as stars, in other words, Wikipedia itself does not even list the positive values it claims.

N.20 Why Floquet theory should be called so

At about the same time as Floquet, Hill appears to have formulated similar ideas. However, he did not publish them, and the credit of publishing a publicly scrutinizable exposure fairly belongs to Floquet.

Note that there is much more to Floquet theory than what is discussed here. If you have done a course on differential equations, you can see why, since the simplest case of periodic coefficients is constant coefficients. Constant coefficient equations may have exponential solutions that do not have purely imaginary arguments, and they may include algebraic factors if the set of exponentials is not complete. The same happens to the variable coefficient case, with additional periodic factors thrown in. But these additional solutions are not relevant to the discussed periodic crystals. They can be relevant to describing simple crystal boundaries, though.

N.21 Superfluidity versus BEC

Many texts and most web sources suggest quite strongly, without explicitly saying so, that the so-called "lambda" phase transition at 2.17 K from normal helium I to superfluid helium II indicates Bose-Einstein condensation.

One reason given that is that the temperature at which it occurs is comparable in magnitude to the temperature for Bose-Einstein condensation in a corresponding system of noninteracting particles. However, that argument is very weak; the similarity in temperatures merely suggests that the main energy scales involved are the classical energy $k_{\rm B}T$ and the quantum energy scale formed from $\hbar^2/2m$ and the number of particles per unit volume. There are likely to be other processes that scale with those quantities besides macroscopic amounts of atoms getting dumped into the ground state.

Still, there is not much doubt that the transition is due to the fact that helium atoms are bosons. The isotope ^{3}He that is missing a neutron in its nucleus does not show a transition to a superfluid until 2.5 mK. The three orders of magnitude difference can hardly be due to the minor difference in mass;

the isotope does condense into a normal liquid at a comparable temperature as plain helium, 3.2 K versus 4.2 K. Surely, the vast difference in transition temperature to a superfluid is due to the fact that normal helium atoms are bosons, while the missing spin ½ neutron in ^{3}He atoms makes them fermions. (The eventual superfluid transition of ^{3}He at 2.5 mK occurs because at extremely low temperatures very small effects allow the atoms to combine into pairs that act as bosons with net spin one.)

While the fact that the helium atoms are bosons is apparently essential to the lambda transition, the conclusion that the transition should therefore be Bose-Einstein condensation is simply not justified. For example, Feynman [18, p. 324] shows that the boson character has a dramatic effect on the *excited* states. (Distinguishable particles and spinless bosons have the same ground state; however, Feynman shows that the existence of low energy excited states that are not phonons is prohibited by the symmetrization requirement.) And this effect on the *excited* states is a key part of superfluidity: it requires a finite amount of energy to excite these states and thus mess up the motion of helium.

Another argument that is usually given is that the specific heat varies with temperature near the lambda point just like the one for Bose-Einstein condensation in a system of noninteracting bosons. This is certainly a good point if you pretend not to see the dramatic, glaring, differences. In particular, the Bose-Einstein specific heat is *finite* at the Bose-Einstein temperature, while the one at the lambda point is *infinite.*. How much more different can you get? In addition, the specific heat curve of helium below the lambda point has a *logarithmic singularity* at the lambda point. The specific heat curve of Bose-Einstein condensation for a system with a unique ground state stays *analytical* until the condensation terminates, since at that point, out of the blue, nature starts enforcing the requirement that the number of particles in the ground state cannot be negative, {D.57}.

Tilley and Tilley [47, p. 37] claim that the qualitative correspondence between the curves for the number of atoms in the ground state in Bose-Einstein condensation and the fraction of superfluid in a two-fluid description of liquid helium "are sufficient to suggest that T_λ marks the onset of Bose-Einstein condensation in liquid ^{4}He." Sure, if you think that a curve reaching a maximum of one exponentially has a similarity to one that reaches a maximum of one with infinite curvature. And note that this compares two completely different quantities. It does not compare curves for particles in the ground state for both systems. It is quite generally believed that the condensate fraction in liquid helium, unlike that in true Bose-Einstein condensation, does not reach one at zero temperature in the first place, but only about 10% or so, [47, pp. 62-66].

Since the specific heat curves are completely different, Occam's razor would suggest that helium has some sort of different phase transition at the lambda point. However, Tilley and Tilley [47, pp. 62-66] present data, their figure 2.17, that suggests that the number of atoms in the ground state does indeed

increase from zero at the lambda point, if various models are to be believed and one does not demand great accuracy. So, the best available knowledge seems to be that Bose-Einstein condensation, whatever that means for liquid helium, does occur at the lambda point. But the fact that many sources see "evidence" of condensation where none exists is worrisome: obviously, the desire to believe despite the evidence is strong and widespread, and might affect the objectivity of the data.

Snoke & Baym point out (in the introduction to *Bose-Einstein Condensation*, Griffin, A., Snoke, D.W., & Stringari, S., Eds, 1995, Cambridge, p. 4), that the experimental signal of a Bose-Einstein condensate is taken to be a delta function for the occupation number of the particles [particle state?] with zero momentum, associated with long-range phase coherence of the wave function. It is not likely to be unambiguously verified any time soon. The actual evidence for the occurrence of Bose-Einstein condensation is in the agreement of theoretical models and experimental data, including also models for the specific heat anomaly. However, Sokol points out in the same volume, (p. 52): "At present, however, liquid helium is the only system where the existence of an experimentally obtainable Bose condensed phase is *almost universally accepted*" [emphasis added].

The question whether Bose-Einstein condensation occurs at the lambda point seems to be academic anyway. The following points can be distilled from Schmets and Montfrooij [39]:

1. Bose-Einstein condensation is a property of the ground state, while superfluidity is a property of the excited states.
2. Ideal Bose-Einstein condensates are *not* superfluid.
3. Below 1 K, essentially 100% of the helium atoms flow without viscosity, even though only about 7% is in the ground state.
4. In fact, there is no reason why a system could not become a superfluid even if only a very small fraction of the atoms were to form a condensate.

The statement that no Bose-Einstein condensation occurs for photons applies to systems in thermal equilibrium. In fact, Snoke & Baym, as mentioned above, use lasers as an example of a condensate that is not superfluid.

N.22 The mechanism of ferromagnetism

It should be noted that in solids, not just spatial antisymmetry, but also symmetry can give rise to spin alignment. In particular, in many ferrites, there is an opposite spin coupling between the iron atoms and the oxygen ones. If two iron atoms are opposite in spin to the same oxygen atom, it implies that they must have aligned spins even if their electrons do not interact directly.

It comes as somewhat a surprise to discover that in this time of high-temperature superconductors, the mechanism of plain old ferromagnetism is still not understood that well if the magnetic material is a conductor, such as a piece of iron.

For a conductor, the description of the exclusion effect should really be at least partly in terms of band theory, rather than electrons localized at atoms. More specifically, Aharoni [2, p. 48] notes "There is thus no doubt in anybody's mind that neither the itinerant electron theory nor the localized electron one can be considered to be a complete picture of the physical reality, and that they both should be combined into one theory."

Sproull notes that in solid iron, most of the 4s electrons move to the 4d bands. That reduces the magnetization by reducing the number of unpaired electrons.

While Sproull [42, p. 282] in 1956 describes ferromagnetism as an interaction between electrons localized at neighboring atoms, Feynman [22, p. 37-2] in 1965 notes that calculations using such a model produce the *wrong sign* for the interaction. According to Feynman, the interaction is thought to occur with [4s] conduction band electrons acting as intermediaries. More recently, Aharoni [2, p. 44] notes: "It used to be stated [...] that nobody has been able to compute a positive exchange integral for Fe, and a negative one for Cu [...]. More modern computations [...] already have the right sign, but the *magnitude* of the computed exchange still differs considerably from the experimental value. Improving the techniques [...] keeps improving the results, but not sufficiently yet."

Batista, Bonča, and Gubernatis note that "After seven decades of intense effort we still do not know what is the minimal model of itinerant ferromagnetism and, more importantly, the basic mechanism of ordering." (Phys Rev Let 88, 2002, 187203-1) and "Even though the transition metals are the most well studied itinerant ferromagnets, the ultimate reason for the stabilization of the FM phase is still unknown." (Phys Rev B 68, 2003, 214430-11)

N.23 Fundamental assumption of statistics

The assumption that all energy eigenstates with the same energy are equally likely is simply stated as an axiom in typical books, [4, p. 92], [18, p. 1], [25, p. 230], [52, p. 177]. Some of these sources quite explicitly suggest that the fact should be self-evident to the reader.

However, why could not an energy eigenstate, call it A, in which all particles have about the same energy, have a wildly different probability from some eigenstate B in which one particle has almost all the energy and the rest has very little? The two wave functions are wildly different. (Note that if the probabilities are only somewhat different, it would not affect various conclusions

much because of the vast numerical superiority of the most probable energy distribution.)

The fact that it does not take any energy to go from one state to the other [18, p. 1] does not imply that the system must spend equal time in each state, or that each state must be equally likely. It is not difficult at all to construct nonlinear systems of evolution equations that conserve energy and in which the system runs exponentially away towards specific states.

However, the coefficients of the energy eigenfunctions do not satisfy some arbitrary nonlinear system of evolution equations. They evolve according to the Schrödinger equation, and the interactions between the energy eigenstates are determined by a Hamiltonian matrix of coefficients. The Hamiltonian is a Hermitian matrix; it has to be to conserve energy. That means that the coupling constant that allows state A to increase or reduce the probability of state B is just as big as the coupling constant that allows B to increase or reduce the probability of state A. More specifically, the rate of increase of the probability of state A due to state B and vice-versa is seen to be

$$\left(\frac{\mathrm{d}|c_A|^2}{\mathrm{d}t}\right)_{\text{duetoB}} = \frac{1}{\hbar}\Im\left(c_A^* H_{AB} c_B\right) \qquad \left(\frac{\mathrm{d}|c_B|^2}{\mathrm{d}t}\right)_{\text{duetoA}} = -\frac{1}{\hbar}\Im\left(c_A^* H_{AB} c_B\right)$$

where H_{AB} is the perturbation Hamiltonian coefficient between A and B. (In the absence of perturbations, the energy eigenfunctions do not interact and $H_{AB} = 0$.) Assuming that the phase of the Hamiltonian coefficient is random compared to the phase difference between A and B, the transferred probability can go at random one way or the other regardless of which one state is initially more likely. Even if A is currently very improbable, it is just as likely to pick up probability from B as B is from A. Also note that eigenfunctions of the same energy are unusually effective in exchanging probability, since their coefficients evolve approximately in phase.

This note would argue that under such circumstances, it is simply no longer reasonable to think that the difference in probabilities between eigenstates of the same energy is enough to make a difference. How could energy eigenstates that readily and randomly exchange probability, in either direction, end up in a situation where some eigenstates have absolutely nothing, to incredible precision?

Feynman [18, p. 8] gives an argument based on time-dependent perturbation theory, chapter 11.10. However, time-dependent perturbations theory relies heavily on approximation, and worse, the measurement wild card. Until scientists, while maybe not agreeing exactly on what measurement *is*, start laying down rigorous, unambiguous, mathematical ground rules on what measurements can do and cannot do, measurement is like astrology: anything goes.

N.24 A problem if the energy is given

Examining all shelf number combinations with the given energy and then picking out the combination that has the most energy eigenfunctions seems straightforward enough, but it runs into a problem. The problem arises when it is required that the set of shelf numbers agrees with the given energy to mathematical precision. To see the problem, recall the simple model system of chapter 11.3 that had only three energy shelves. Now assume that the energy of the second shelf is not $\sqrt{9} = 3$ as assumed there, (still arbitrary units), but slightly less at $\sqrt{8}$. The difference is small, and all figures of chapter 11.3 are essentially unchanged. However, if the average energy per particle is still assumed equal to 2.5, so that the total system energy equals the number of particles I times that amount, then I_2 must be zero: it is impossible to take a nonzero multiple of an irrational number like $\sqrt{8}$ and end up with a rational number like $2.5I - I_1 - 4I_3$. What this means graphically is that the oblique energy line in the equivalent of figure 11.5 does not hit any of the centers of the squares mathematically exactly, except for the one at $I_2 = 0$. So the conclusion would be that the system must have zero particles on the middle shelf.

Of course, physically this is absolute nonsense; the energy of a large number of perturbed particles is not going to be certain to be 2.5 I to mathematical precision. There will be *some* uncertainty in energy, and the correct shelf numbers are still those of the darkest square, even if its energy is $2.499\,9\ldots I$ instead of 2.5 I exactly. Here typical textbooks will pontificate about the accuracy of your system-energy measurement device. However, this book shudders to contemplate what happens physically in your glass of ice water if you have three system-energy measurement devices, but your best one is in the shop, and you are uncertain whether to believe the unit you got for cheap at Wal-Mart or your backup unit with the sticking needle.

To avoid these conundrums, in this book it will simply be assumed that the right combination of shelf occupation numbers is still the one at the maximum in figure 11.6, i.e. the maximum when the number of energy eigenfunctions is mathematically interpolated by a continuous function. Sure, that may mean that the occupation numbers are no longer exact integers. But who is going to count 10^{20} particles to check that it is exactly right? (And note that those other books end up doing the same thing anyway in the end, since the mathematics of an integer-valued function defined on a strip is so much more impossible than that of a continuous function defined on a line.)

If fractional particles bothers you, even among 10^{20} of them, just fix things after the fact. After finding the fractional shelf numbers that have the biggest energy, select the whole shelf numbers nearest to it and then change the "given" energy to be $2.499\,999\,9\ldots$ or whatever it turns out to be at those whole shelf numbers. Then you should have perfectly correct shelf numbers with the highest number of eigenfunctions for the new given energy.

N.25 The recipe of life

Religious nuts, "creationists," "intelligent designers," or whatever they are calling themselves at the time you are reading this, call them CIDOWs for short, would like to believe that the universe was created *literally* like it says in the bible. The bible contains two creation stories, the Genesis story and the Adam and Eve story, and they conflict. At some time in the past they were put in together for simplicity, without ironing out their contradictions. CIDOWs feel that with two conflicting creation stories, surely at least one should be right? This is the *bible*, you know?

Now if you want to believe desperately enough, you are willing to accept anything that seems to reasonably support your point, without looking too hard at any opposing facts. (Critically examining facts is what a scientist would do, but you can reasonably pass yourself off as a scientist in the court system and popular press without worrying about it. You do have to pass yourself off as a scientist in the United States, since it is unconstitutional to force your religious beliefs upon the public education system unless you claim they are scientific instead of religious.) Now CIDOWs had a look at life, and it seemed to be quite nonmessy to them. So they felt its entropy was obviously low. (Actually, a human being may be a highly evolved form of life, but being largely water well above absolute zero temperature, its entropy is not particularly low.) Anyway, since the earth has been around for quite some time, they reasoned that the entropy of its surface must have been increasing for a long time, and nonmessy human beings could not possibly be true. Hence the conventional scientific explanation of the evolution of life violated the second law and could not be true. It followed that the universe just had to be created by God. The Christian God of course, don't assume now that Allah or Buddha need apply.

Hello CIDOWs! The surface of the earth is hardly an adiabatic system. See that big fireball in the sky? What do you think all that plant life is doing with all those green leaves? Baierlein [4, pp. 128-130] works out some of the rough details. Since the surface of the sun is very hot, the photons of light that reach us from the sun are high energy ones. Despite the influx of solar energy, the surface of the earth does not turn into an oven because the earth emits about the same energy back into space as it receives from the sun. But since the surface of the earth is not by far as hot as that of the sun, the photons emitted by the earth are low energy ones. Baierlein estimates that the earth emits about 20 of these low energy photons for every high energy one it receives from the sun. Each photon carries one unit of entropy on average, (11.59). So the earth *loses* 20 units of messiness for every one it receives. So, evolution towards less messy systems is exactly what you would expect for the earth surface, based on the overall entropy picture. Talk about an argument blowing up in your face!

N.26 Physics of the fundamental commutators

The fundamental commutation relations look much like a mathematical axiom. Surely, there should be some other reasons for physicists to believe that they apply to nature, beyond that they seem to produce the right answers?

Addendum {A.19} explained that the angular momentum operators correspond to small rotations of the axis system through space. So, the commutator $[\widehat{J}_x, \widehat{J}_y]$ really corresponds to the difference between a small rotation around the y-axis followed by a small rotation around the x-axis, versus a small rotation around the x-axis followed by a small rotation around the y axis. As shown below, in our normal world this difference is equivalent to the effect of a small rotation about the z-axis.

So, the fundamental commutator relations do have physical meaning; they say that this basic relationship between rotations around different axes continues to apply in the presence of spin.

This idea can be written out more precisely by using the symbols $\mathcal{R}_{x,\alpha}$, $\mathcal{R}_{y,\beta}$, and $\mathcal{R}_{z,\gamma}$ for, respectively, a rotation around the x-axis over an angle α, around the y-axis over an angle β, and the z-axis over an angle γ. Then following {A.19}, the angular momentum around the z-axis is by definition:

$$\widehat{J}_z \approx \frac{\hbar}{\mathrm{i}} \frac{\mathcal{R}_{z,\gamma} - I}{\gamma}$$

(To get this true exactly, you have to take the limit $\gamma \to 0$. But to keep things more physical, taking the mathematical limit will be delayed to the end. The above expression can be made arbitrarily accurate by just taking γ small enough.)

Of course, the x and y components of angular momentum can be written similarly. So their commutator can be written as:

$$[\widehat{J}_x, \widehat{J}_y] \equiv \widehat{J}_x \widehat{J}_y - \widehat{J}_y \widehat{J}_x \approx \frac{\hbar^2}{\mathrm{i}^2} \left(\frac{\mathcal{R}_{x,\alpha} - I}{\alpha} \frac{\mathcal{R}_{y,\beta} - I}{\beta} - \frac{\mathcal{R}_{y,\beta} - I}{\beta} \frac{\mathcal{R}_{x,\alpha} - I}{\alpha} \right)$$

or multiplying out

$$[\widehat{J}_x, \widehat{J}_y] \approx \frac{\hbar^2}{\mathrm{i}^2} \frac{\mathcal{R}_{x,\alpha}\mathcal{R}_{y,\beta} - \mathcal{R}_{y,\beta}\mathcal{R}_{x,\alpha}}{\alpha\beta}$$

The final expression is what was referred to above. Suppose you do a rotation of your axis system around the y-axis over a small angle β followed by a rotation around the x-axis around a small angle α. Then you will change the position coordinates of every point slightly. And so you will if you do the same two rotations in the opposite order. Now if you look at the difference between these two results, it is described by the numerator in the final ratio above.

All those small rotations are of course a complicated business. It turns out that in our normal world you can get the same differences in position in a much

simpler way: simply rotate the axis system around a small angle $\gamma = -\alpha\beta$ around the z-axis. The change produced by that is the numerator in the expression for the angular momentum in the z-direction given above. If the two numerators are the same for small α and β, then the fundamental commutation relation follows. At least in our normal world. So if physicists extend the fundamental commutation relations to spin, they are merely generalizing a normal property of rotations.

To show that the two numerators are the indeed the same for small angles requires a little linear algebra. You may want to take the remainder of this section for granted if you never had a course in it.

First, in linear algebra, the effects of rotations on position coordinates are described by matrices. In particular,

$$\mathcal{R}_{x,\alpha} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \qquad \mathcal{R}_{y,\beta} = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}$$

By multiplying out, the commutator is found as

$$[\widehat{J}_x, \widehat{J}_y] \approx \frac{\hbar^2}{\mathrm{i}^2\alpha\beta} \begin{pmatrix} 0 & -\sin\alpha\sin\beta & -\sin\beta(1-\cos\alpha) \\ \sin\alpha\sin\beta & 0 & -\sin\alpha(1-\cos\beta) \\ -\sin\beta(1-\cos\alpha) & -\sin\alpha(1-\cos\beta) & 0 \end{pmatrix}$$

Similarly, the angular momentum around the z-axis is

$$\widehat{J}_z \approx \frac{\hbar}{\mathrm{i}\gamma} \begin{pmatrix} \cos\gamma - 1 & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma - 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

If you take the limit that the angles become zero in both expressions, using either l'Hôpital or Taylor series expansions, you get the fundamental commutation relationship.

And of course, it does not make a difference which of your three axes you take to be the z-axis. So you get a total of three of these relationships.

N.27 Magnitude of components of vectors

You might wonder whether the fact that the square components of angular momentum must be less than total square angular momentum still applies in the quantum case. After all, those components do not exist at the same time. But it does not make a difference: just evaluate them using expectation values. Since states $|j\,m\rangle$ are eigenstates, the expectation value of total square angular momentum is the actual value, and so is the square angular momentum in the z-direction. And while the $|j\,m\rangle$ states are not eigenstates of $\widehat{J}_x$ and $\widehat{J}_y$, the

expectation values of square Hermitian operators such as $\widehat{J}_x^2$ and $\widehat{J}_y^2$ is always positive anyway (as can be seen from writing it out in terms of the eigenstates of them.)

N.28 Adding angular momentum components

The fact that net angular momentum components can be obtained by summing the single-particle angular momentum operators is clearly following the Newtonian analogy: in classical physics each particle has its own independent angular momentum, and you just add them up,

See also addendum {A.19}.

N.29 Clebsch-Gordan tables are bidirectional

The fact that you can read the tables either by rows or by columns is due to the orthonormality of the states involved. In terms of the real vectors of physics, it is simply an expression of the fact that the component of one unit vector in the direction of another unit vector is the same as the component of the second unit vector in the direction of the first.

N.30 Machine language Clebsch-Gordan tables

The usual "machine language" form of the tables leaves out the a, b, and ab identifiers, the $j_a =$ and $j_b =$ clarifications from the header, and all square root signs, the j values of particles a and b from the kets, and all ket terminator bars and brackets, but combines the two m values with missing j values together in a frame to resemble an jm ket as well as possible, and then puts it all in a font that is very easy to read with a magnifying glass or microscope.

N.31 Existence of magnetic monopoles

Actually, various advanced quantum theories really require the existence of magnetic monopoles. But having never been observed experimentally with confidence despite the big theoretical motivation for the search, they are clearly not a significant factor in real life. Classical electromagnetodynamics assumes that they do not exist at all.

N.32 More on Maxwell's third law

Since the voltage is minus the integral of the electric field, it might seem that there is a plus and minus mixed up in figure 13.5.

But actually, it is a bit more complex. The initial effect of the induced electric field is to drive the electrons towards the pole marked as negative. (Recall that the charge of electrons is negative, so the force on the electrons is in the direction opposite to the electric field.) The accumulation of electrons at the negative pole sets up a counter-acting electric field that stops further motion of the electrons. Since the leads to the load will be stranded together rather than laid out in a circle, they are not affected by the induced electric field, but only by the counter-acting one. If you want, just forget about voltages and consider that the induced electric field will force the electrons out of the negative terminal and through the load.

N.33 Setting the record straight on alignment

Some sources claim the spin is under an angle with the magnetic field; this is impossible since, as pointed out in chapter 4.2.4, the angular momentum vector does not exist. However, the angular momentum component along the magnetic field does have measurable values, and these component values, being one-dimensional, can only be aligned or anti-aligned with the magnetic field.

To intuitively grab the concept of Larmor precession, it may be useful anyway to think of the various components of angular momentum as having precise nonzero values, rather than just being uncertain. But the latter is the truth.

N.34 NuDat 2 data selection

The gamma decay data of figures 14.63 and 14.64 were retrieved from NuDat 2, ⟦12⟧, October-November 2011.

In the data selection, transitions were ignored if any ambiguity at all was indicated for any of the primary data. The primary data were the initial energy level, the released energy in the transition of interest, the half life, the initial and final spins and parities, the multipole type and order, the relative intensity of the gamma transition of interest, (see below for more), the mixing ratio for transitions of mixed multipole type, the conversion coefficient, (see below for more), and the decay rate in Weisskopf units. If there was a decay process other than gamma decay indicated, the gamma decay percentage had to be given without ambiguity. Indications of ambiguity included parentheses, square brackets, inequalities, tildes, question marks, multiple values, and more.

Note that the given data uncertainties were ignored. That is a weakness of the data, but presumably not really important in view of the very large

deviations from theory. Including uncertainties would make processing much more complicated.

As an overall check on the data, the computed transition rate was compared to the one provided in terms of Weisskopf units by NuDat 2 itself. If the difference was less than 5% and there were no other concerns, as discussed below, the transition was accepted automatically. Tests were also performed on whether the initial and final energy levels matched the energy release, and on spin and parity conservation. These tests were mainly to guard against typos in the data base and no violations were observed.

If there were any concerns, the data were printed out. A decision was then made manually on whether to accept the transition as a potential candidate for plotting. If the computed transition rate was substantially, (more than roughly 15%), above the NuDat 2 value, the transition was rejected out of hand. If the computed transition rate was below the NuDat 2 one, it was examined whether the NuDat 2 value was self-evidently missing its correction for other decay types, for the other gamma intensities, the mixing ratio, or the conversion coefficient. That was observed in a relatively small number of cases, usually for a missing conversion coefficient. In all other cases, for a substantial difference in decay rates, over about 15%, the transition was rejected.

Even if the computed decay rate matched the Weisskopf one, various decays were manually rejected. In doing so, if the gamma intensity was not given, it was assumed to be 100% only if there was only one gamma decay out of the energy level. If there were other gamma decays out of the same energy level, their intensities were, based on manual examination, allowed to be omitted (assumed to be zero), specified by an upper limit if small (assumed to be half the upper limit), or specified as approximate if small. If the conversion coefficient was not given, it was manually allowed to be zero if the incompressible ballpark value was below about 10^{-4}. Some initial energy levels with multiple gamma decays were manually rejected if the transition of interest had very low intensity and only one or two digits were given. Mixed transitions were manually examined, but it was not considered cause for rejection as long as a valid mixing ratio was given. If the multipole level was higher than needed, that was also announced, but it too was not taken to be a reason for rejection.

No, in the manual selections, the author did not select the worst nuclei to make physicists look bad.

The plot range from 30 to 3 000 keV (in the plots reduced to 2 500 keV) energy release was divided into 70 segments for which one symbol to plot each. Transitions to plot were selected by comparing them to the selected transitions in the other segments. The selection was designed to achieve a broad coverage of transitions. For plot segments for which there was only one available transition, that transition was immediately selected. Then the program iterated over the segments with more than one potential candidate for plotting. In each segment the best candidate to plot was selected according to the following criteria:

1. Candidates that were more distant in terms of Z from the currently selected candididates in the other segments received priority. Distance was here defined as the distance from the closest selected nucleus of the other segments. The intention was to cover the entire range of atomic numbers as well as possible.
2. In case of a tie, nuclei that were different from the most other selected nuclei in terms of either proton or neutron odd/evenness received priority. The intention was to include all variations of even/oddness.
3. In case of a tie, nuclei that were stable received priority. That was in the hopes that data on stable nuclei might be better quality.
4. In case of a tie, nuclei that were more different in A from the already selected nuclei received priority.
5. In case of a tie, a random choice was made between the nuclei in the tie.

Because these criteria depend on the selections in the other segments, iteration was needed. The iterations were terminated if there were no more changes in the selected candidates.

The data on the selected nuclei, including log files, are available in the web version of this document[1]. If you have suggestions on how the data could be improved, let me know.

N.35 Auger discovery

Meitner submitted the discovery of the Auger process to the Zeitschrift für Physik, a major journal, on Jan 8 1922 and it appeared in the Nov 1/Dec issue that year. The process is clearly described. Auger's first description appeared in the July 16 1923 Séance of the Comptes Rendus of the Academy of Sciences in France (in French). There is no record of Meitner having apologized to Auger for not having waited with publication even though a male physicist was clearly likely to figure it out sooner or later.

It is generally claimed that Meitner should have shared the Nobel prize with Hahn for the discovery of nuclear fission. One reason given is that it was Meitner who found the explanation of what was going on and coined the phrase "fission." Meitner also did much of the initial experimental work with Hahn that led to the discovery. Fortunately, Meitner was Jewish and had to flee Hitler's Germany in 1938. That made it much easier for Hahn to shove her out of the way and receive all the credit, rather than having to share it with some woman.

[1] http://www.dommelen.net/quansup/nudat2/

N.36 Draft: Cage-of-Faraday proposal

The gigantic errors in theoretical half-life predictions in section 14.20.5 are disconcerting to say the least. They imply that the predicted gamma decay rates, (essentially the inverse of the half-lifes), are typically either much less than theory or much larger than theory.

To explain why some of the gamma decay rates, like the E2 and high energy E3 ones, are so much faster than ballpark is relatively straightforward. Decay much faster than ballpark is only possible if not just one proton, but a lot of nucleons participate in the transition. And since the effect is systematic in E2 and high energy E3 transitions, apparently it is *normal* for a lot of nucleons to participate in gamma decay. Or at least it is for these types of gamma decay. And since the theory assumes that only one proton participates, the miserable predictions of theory can be explained.

A much bigger problem is to explain why other transitions end up so far below ballpark in a credible way. Consider in particular the E1 transitions in figures 14.63 and 14.65. How come that they are not just occasionally, but *typically* slower than theory by four orders of magnitude?

Basically, you can give two reasonable types of explanation:

1. You can assume that only one proton participates in these transition, not many as in E2 and E3 transitions. Then you must assume that in addition there is a systematic very poor overlap between the initial and final states in the relevant inner product. We do not really know the initial and final states, so, why not? Problem solved. Next question?

   This is essentially the explanation that basic nuclear textbooks that the author has seen give. Unfortunately, there are two big problems with it. First, how come that unlike the E2 and E3 transitions, suddenly only one proton participates in E1, low energy E3, M1 and most M2 transitions? If there is a very big *systematic* effect, there must be a reason. Worse, figure 14.65 seems to exclude the possibility of just one proton partipating in at the very least M1 transitions, and surely at least some very slow E1 transitions.

   The second problem is to explain why the overlap is systematically extremely bad in some types of transitions, but apparently excellent in others. Again, this is a big systematic effect, as the figures show. So there must be a reason for this too.

   Your hands are not enough to wave these problem away. If you want people to take you seriously, you should have a believable and comprehensive discussion.
2. Alternatively, you can assume that in the transitions that are much slower than theory, still many nucleons participate. One immediate advantage is, of course, that this would explain why the theory per-

> forms miserably bad on these transitions too. And you do not have to explain why in some transitions only a single proton participates while in others many nucleons do. In particular, this removes figure 14.65 as an issue. It can also explain why for some very light nuclei, E1 and M1 transitions are at ballpark or even noticeably faster than ballpark.
>
> But now of course, you face the apparently daunting problem of explaining why some types transitions can be so extremely slow, even now that there are many nucleons participating. In particular, you need to provide a reasonable explanation why for some types of transitions, the participation of many nucleons actually slows down the emission of electromagnetic radiation greatly, rather than increase it greatly. And apparently, this effect requires the presence of enough nucleons, as very light nuclei do not have the problem.

While standard nuclear textbooks give the first explanation above, the holes in the argument are worrisome. The magnitude of the effects pointed at by the textbooks just does not seem big enough to explain the data. And it is hard to think up reasons why not. There just is a lack of suspension of disbelieve for an engineer thinking in terms of ballparks.

Therefore, this book wants to argue that more serious consideration should be given to the second explanation. Its main liability is to explain why some transitions get slowed down greatly, rather than sped up, if a lot more nucleons participate.

Since nuclear wave functions are poorly understood, that would be difficult to explain from a quantum-mechanical viewpoint. So maybe it is again time to do what has been done before for nuclei; look for macroscopic models. And surely the macroscopic model that stands out in killing off electromagnetic effects is the cage of Faraday. (In this case the cage is assumed to shield the outside from the inside.)

Maybe then the nuclear surface acts as such a cage in some sense. In the liquid-drop idea, nucleons at the surface are in a state of increased energy. So it may not be such a crazy idea that nucleons at the surface might behave differently from nucleons in the interior.

Assume now at first, in this macroscopic model, that the nuclear surface is spherical and conducting. Then electric charge changes in the interior of the nucleus would not leak out. That would kill off the capability of transitions to emit radiation. So the model can provide a macroscopic explanation why some electromagnetic transitions can be greatly slowed down. Charges can still move around inside the nucleus, but because the surface nucleons move to compensate, that does not produce radiation outside it. So participation of many nucleons does indeed reduce, rather than increase, electromagnetic radiation greatly.

Do note that unlike normal cages of Faraday, a nucleus contains a net positive charge. And if a macroscopic cage of Faraday contains a net charge, there must

always ne a nonzero electric field outside. (That is due to Maxwell's first law.) But as long as the surface remains *spherical* and conducting, the outside electric field will not change if charges inside the surface are moved around. So in that case, there will be an electric field, but still no electromagnetic radiation radiated away. And the reason for that is still because many nucleons are involved, rather than a single proton.

But things change when the nuclear surface changes shape. A conducting surface makes only the electric field tangential to the surface zero. Therefore there will be variations in the electric field outside the surface if it changes shape. So now we have a situation where radiation is in fact being transmitted, and again with many nucleons involved in doing that.

That opens up the possibility of explaining why some transitions can be so far from the single proton ballpark. And why it depends on the type of transition whether the transition turns out to be much slower than ballpark or much faster than ballpark.

At least for relatively light nuclei, (but still with enough nucleons that the macroscopic picture makes sense), and small excitations, surface tension would promote a spherical surface. And "surface roughness" would not necessarily make much of a difference. That is just like small holes in a macroscopic cage do not make a difference. The field outside the nucleus is governed by the so-called Laplace equation. This equation is known to kill off short-scale perturbations quickly.

On the other hand, changes in a deformed nuclear surface shape would definitely produce nontrivial long-range electric field perturbations. Now nuclei are often modeled as spheroids or ellipsoids. Changes in such a shape would produce quadrupole and hexadecapole perturbations in the electric field outside the nuclei. So they would produce E2 radiation, but not E1 radiation. That is exactly what is needed to make some sense out of the electric transitions.

While macroscopic cages of Faraday do not block static magnetic fields, they do block changes in magnetic fields. So conceptually the model could also explain why magnetic transitions of low multipole order are often so slow. Note that there is no net magnetic "charge" inside the cage. Magnetic monopoles do not exist. So surface shape would not necessarily affect magnetic transitions much.

While this model leaves many questions unanswered, at least it suggests a reasonable way to understand how it is possible at all that the one-proton model is not just extremely miserable, but *systematically* miserable in the observed way. For one, it seems to make figure 14.65 far less unexplainable.

Web Pages

Below is a list of relevant web pages.

1. chemguide.co.uk[2]

   Jim Clarke's UK site with lots of solid info.

2. Citizendium[3]

   The Citizen's Compendium. Had a rather good write up on the quantization of the electromagnetic field.

3. Elster's lecture notes[4]

   Professor Elster gives a very helpful historical overview of the meson exchange potentials, (`fewblect_2.pdf`). She also gives the detailed potentials for scalar and vector mesons that the other references do not, (`fewblect_1.pdf`).

4. ENSDF data[5]

   The Nuclear Data Sheets are an authoritative and comprehensive data source on nuclei. The corresponding
   Nuclear Data Sheets policies[6]
   have been used repeatedly in this book to decide what conventions to take as standard.

5. Richard P. Feynman: Nobel Prize lecture[7]

   Describes the development of Feynman's path integral approach to quantum electrodynamics.

6. Hyperphysics[8]

   Gives simple explanations of almost anything in physics. An extensive source of info on chemical bonds and the periodic table.

7. ICC program[9]

   Program to compute internal conversion coefficients.

[2] `http://www.chemguide.co.uk/`
[3] `http://en.citizendium.org/`
[4] `http://www.phy.ohiou.edu/~elster/lectures/`
[5] `http://www-nds.iaea.org/relnsd/NdsEnsdf/QueryForm.html`
[6] `http://www.nndc.bnl.gov/nds/NDSPolicies.pdf`
[7] `http://nobelprize.org/nobel_prizes/physics/laureates/1965/`
[8] `http://hyperphysics.phy-astr.gsu.edu/hbase/hph.html`
[9] `http://ie.lbl.gov/programs/icc/icc.htm`

8. J. Jäckle[10]
   This web site includes a good description of the Peltier and Seebeck effects.
9. R.D. Klauber's pedagogical quantum field theory[11]
   This web site gives a fully explained description of quantum field theory.
10. Mayer, M. Goeppert: Nobel Prize lecture[12]
    An excellent introduction to the shell model of nuclear physics written for a general audience is found in the lecture.
11. NIST data[13]
    Authoritative values of physical constants from NIST.
12. NuDat 2 database[14]
    Extensive information about nuclei provided by the National Nuclear Data Center.
13. Purdue chemistry review[15]
    General chemistry help.
14. Quantum Exchange[16]
    Lots of stuff.
15. Rainwater, J.: Nobel Prize lecture[17]
    An introduction to distorted nuclei written for a general audience is found in the lecture.
16. Anthony Stone's Wigner coefficient calculators[18]
    The calculator on this site gives exact values for the Wigner 3j, 6j, and 6j symbols. The 3j symbols are readily converted to Clebsch-Gordan coefficients, {N.13}.
17. David Tong's notes on quantum field theory[19]
    Very helpful, especially in conjunction with Peskin & Schroeder, [35].
18. T. Tritt[20]
    Thermoelectric materials: principles, structure, properties, and applications. From Encyclopedia of Materials: Science and Technology. Elsevier 2002.

---

[10] http://www.uni-konstanz.de/FuF/Physik/Jaeckle/
[11] http://www.quantumfieldtheory.info/
[12] http://nobelprize.org/nobel_prizes/physics/laureates/1963/
[13] http://www.nist.gov/pml/data/
[14] http://www.nndc.bnl.gov/nudat2/
[15] http://chemed.chem.purdue.edu/genchem/
[16] http://www.compadre.org/quantum/
[17] http://nobelprize.org/nobel_prizes/physics/laureates/1975/
[18] http://www-stone.ch.cam.ac.uk/wigner.html
[19] http://www.damtp.cam.ac.uk/user/tong/qft.html
[20] http://virtual.clemson.edu/TMRL/Publications/PDFS/teoverview.pdf

19. TUNL Nuclear Data Evaluation Group[21]
    Extensive data on light nuclei from $A = 3$ to 20.
20. University of Michigan[22]
    Invaluable source on the hydrogen molecule and chemical bonds. Have a look at the animated periodic table for actual atom energy levels.
21. Wikipedia[23]
    Probably this book's primary source of information on about every loose end, though somewhat uneven. Some great, some confusing, some overly technical.

[21] http://www.tunl.duke.edu/nucldata/
[22] http://www.umich.edu/~chem461/
[23] http://wikipedia.org

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, third edition, 1965. 880, 1086, 1097, 1290, 1402

[2] A. Aharoni. *Introduction to the Theory of Ferromagnetism*. Oxford University Press, second edition, 2000. 1472

[3] G. Audi, O. Bersillon, J. Blachot, and A. H. Wapstra. The NUBASE evaluation of nuclear and decay properties. *Nuclear Physics A*, 729:3–128, 2003. xxxviii, 731

[4] R. Baierlein. *Thermal Physics*. Cambridge University Press, Cambridge, UK, 1999. xxxviii, 521, 525, 564, 1372, 1472, 1475

[5] C.A. Bertulani. *Nuclear Physics in a Nutshell*. Princeton University Press, 2007. 681, 685, 835, 1077, 1079, 1167, 1175, 1201, 1380, 1460

[6] H.A. Bethe. *Intermediate Quantum Mechanics*. W.A. Benjamin, 1964. 1024, 1026, 1031

[7] J.M. Blatt and V.F. Weisskopf. *Theoretical Nuclear Physics*. Wiley, 1952. 851, 853

[8] J.M. Blatt and V.F. Weisskopf. *Theoretical Nuclear Physics*. Springer-Verlag, 1979. 851, 853

[9] S.H. Chue. *Thermodynamics : a rigorous postulatory approach*. Wiley, 1977. 892

[10] E.U. Condon and H. Odishaw, editors. *Handbook of Physics*. McGraw-Hill, 1958. 1077, 1078

[11] E.U. Condon and H. Odishaw, editors. *Handbook of Physics*. McGraw-Hill, 2nd edition, 1967. 1053, 1069, 1075, 1077, 1078, 1460

[12] E.A. Desloge. *Thermal Physics*. Holt, Rinehart and Winston, New York, 1968. xxxvii

[13] A.R. Edmonds. *Angular Momentum in Quantum Mechanics*. Princeton, 1957. 1458

[14] J.P. Elliott. Nuclear symmetries. In D.H. Wilkinson, editor, *Isospin in Nuclear Physics*, pages 73–113. Wiley/North Holland, 1969. 800

[15] A.M. Ellis. Spectroscopic selection rules: The role of photon states. *J. Chem. Educ.*, 76:1291–1294, 1999. xxxvii

[16] L.R.B. Elton. *Introductory Nuclear Theory*. W.B. Saunders, 2nd edition, 1966. 1168, 1186, 1315, 1316, 1318

[17] Hugh Everett, III. The theory of the universal wave function. In Bryce S. DeWitt and Neill Graham, editors, *The Many-Worlds Interpretation of Quantum Mechanics*, pages 3–140. Princeton University Press, 1973. xxxvii

[18] R. P. Feynman. *Statistical Mechanics*. Westview (Perseus), 1998. xxxviii, 223, 533, 554, 885, 889, 909, 1470, 1472, 1473

[19] R. P. Feynman. *QED, the Strange Theory of Light and Matter*. Princeton, expanded edition, 2006. 634, 647, 910, 1110, 1138, 1148

[20] R.P. Feynman. *The Character of Physical Law*. BBC/Penguin, 1965. xxxvii

[21] R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics*, volume I. Addison-Wesley, 1965. 25

[22] R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics*, volume III. Addison-Wesley, 1965. xxxvii, 134, 311, 909, 1472

[23] N.B. Gove and M.J. Martin. Log f tables. *Nucl. Data Tables A*, 10:206, 1971. 1192, 1211

[24] David Griffiths. *Introduction to Elementary Particles*. Wiley-VCH, second edition, 2008. 909, 963, 971, 974, 977, 1436

[25] David J. Griffiths. *Introduction to Quantum Mechanics*. Pearson Prentice-Hall, second edition, 2005. xxxvi, xxxvii, xxxviii, 93, 415, 580, 634, 1135, 1247, 1255, 1378, 1445, 1472

[26] B.R. Holstein. The Van der Waals interaction. *Am. J. Phys.*, 69:441–449, 2001. 470

[27] K Huang. *Fundamental Forces of Nature*. World Scientific, 2007. 359, 419

[28] A. C. Hurley. *Introduction to the electron theory of small molecules*. Academic Press, 1976. 1356

[29] C. Kittel. *Introduction to Solid State Physics*. Wiley, 7th edition, 1996. xxxviii, 282, 492, 890, 909

[30] W. Koch and M. C. Holthausen. *A chemist's guide to density functional theory*. Wiley-VCH, Weinheim, second edition, 2000. 1464, 1465

[31] K.S. Krane. *Introductory Nuclear Physics*. Wiley, 1988. xxxviii, 335, 649, 654, 684, 686, 693, 711, 718, 725, 735, 741, 743, 779, 814, 818, 827, 828, 834, 835, 1077, 1079, 1156, 1175, 1201, 1380, 1403

[32] R. Machleidt. The nuclear force in the third millennium. *Nucl. Phys. A*, 689:11–22, 2001. arXiv:nucl-th/0009055v2. 685, 1177

[33] A.A. Moszkowski. Theory of multipole radiation. In K. Siegbahn, editor, *Alpha-, Beta-, and Gamma-Ray Spectroscopy*, volume 2, pages 863–886. North Holland, 1965. 1053, 1056, 1067, 1069, 1072, 1075, 1079, 1321, 1460

[34] R. G. Parr and W. Yang. *Density Functional Theory of Atoms and Molecules*. Oxford, New York, 1989. xxxviii, 1464

[35] M.E. Peskin and D.V. Schroeder. *An Introduction to Quantum Field Theory*. Westview, 1995. 933, 1003, 1020, 1486

[36] M.A. Preston and R.K. Bhaduri. *Structure of the Nucleus*. Addison-Wesley, 1975. xxxviii, 685, 711, 727, 731, 734, 735, 749, 752, 757, 761, 764, 781, 797, 1077, 1079, 1167, 1175, 1176, 1380

[37] P. Roy Chowdhury and D.N. Basu. Nuclear matter properties with the re-evaluated coefficients of liquid drop model. *Acta Physica Polonica B*, 37:1833–1846, 2006. 673, 687

[38] A. Schirrmacher. Experimenting theory: The proofs of Kirchhoff's radiation law before and after Planck. *Historical Studies in the Physical and Biological Sciences*, 33:299–335, 2003. Freely available online at http://caliber.ucpress.net/toc/hsps/33/2. 1272

[39] A. J. M. Schmets and W. Montfrooij. Teaching superfluidity at the introductory level, 2008. URL `http://arxiv.org/abs/0804.3086`. 223, 1471

[40] A. Sitenko and V. Tartakovskii. *Theory of Nucleus*. Kluwer, 1997. xxxviii, 711, 732, 741, 743, 748, 749, 780, 781, 782, 1178

[41] M.R. Spiegel and J. Liu. *Mathematical Handbook of Formulas and Tables*. Schaum's Outline Series. McGraw-Hill, second edition, 1999. 63, 64, 65, 93, 96, 227, 566, 571, 615, 1123, 1140, 1205, 1227, 1228, 1229, 1230, 1235, 1238, 1239, 1242, 1246, 1247, 1285, 1290, 1292, 1294, 1295, 1313, 1376, 1409, 1498, 1509, 1531

[42] R. L. Sproull. *Modern Physics, a textbook for engineers*. Wiley, first edition, 1956. xxxviii, 271, 1272, 1472

[43] M. Srednicki. *Quantum Field Theory*. Cambridge University Press, Cambridge, UK, 2007. xxxviii, 933, 934, 936, 937, 1003

[44] B. Stech. Die lebensdauer isomerer kerne. *Z. Naturforschg*, 7a:401–410, 1952. 1055, 1315, 1316, 1460

[45] N.J. Stone. The table of nuclear moments. *Atomic Data and Nuclear Data Tables*, 90:75–176, 2005. Also available online at bnl.gov, sciencedirect.com. xxxviii

[46] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry*. Dover, first, revised edition, 1996. xxxviii, 452, 461, 462, 463, 1356, 1464, 1465

[47] D. R. Tilley and J. Tilley. *Superfluidity and Superconductivity*. Institute of Physics Publishing, Bristol and Philadelphia, third edition, 1990. 1470

[48] E.K. Warburton and J. Weneser. The role of isospin in electromagnetic transitions. In D.H. Wilkinson, editor, *Isospin in Nuclear Physics*, pages 173–228. Wiley/North Holland, 1969. 834, 835

[49] S. Weinberg. *The Quantum Theory of Fields*, volume I: Foundations. Cambridge, 2010. 18, 952, 1010

[50] D.H. Wilkinson. Historical introduction to isospin. In D.H. Wilkinson, editor, *Isospin in Nuclear Physics*, pages 1–13. Wiley/North Holland, 1969. 791, 796, 797, 799

[51] R.B. Wiringa, V.G.J. Stoks, and R. Schiavilla. An accurate nucleon-nucleon potential with charge-independence breaking. *Phys .Rev. C*, 51:38–51, 1995. arXiv:nucl-th/9408016v1. 1158, 1160

[52] A. Yariv. *Theory and Applications of Quantum Mechanics*. Wiley & Sons, 1982. xxxvi, xxxviii, 93, 335, 362, 634, 1472

[53] A. Zee. *Quantum Field Theory in a Nutshell*. Princeton University Press, Princeton, NJ, 2003. 909, 986, 1020, 1191

Notations

The below are the simplest possible descriptions of various symbols, just to help you keep reading if you do not remember/know what they stand for. Don't cite them on a math test and then blame this book for your grade.

Watch it. There are so many ad hoc usages of symbols, some will have been overlooked here. Always use common sense first in guessing what a symbol means in a given context.

The quoted values of physical constants are usually taken from NIST CODATA in 2012 or later. The final digit of the listed value is normally doubtful. (It corresponds to the first nonzero digit of the standard deviation). Numbers ending in triple dots are exact and could be written down to more digits than listed if needed.

$\cdot$ A dot might indicate

- A dot product between vectors, if in between them.
- A time derivative of a quantity, if on top of it.

And also many more prosaic things (punctuation signs, decimal points, ...).

$\times$ Multiplication symbol. May indicate:

- An emphatic multiplication.
- Multiplication continued on the next line or from the previous line.
- A vectorial product between vectors. In index notation, the i-th component of $\vec{v} \times \vec{w}$ equals

$$(\vec{v} \times \vec{w})_i = v_{\bar{i}} w_{\bar{\bar{i}}} - v_{\bar{\bar{i}}} w_{\bar{i}}$$

where $\bar{i}$ is the index following i in the sequence 123123..., and $\bar{\bar{i}}$ the one preceding it (or second following). Alternatively, evaluate the determinant

$$\vec{v} \times \vec{w} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{vmatrix}$$

! Might be used to indicate a factorial. Example: $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$.

The function that generalizes $n!$ to noninteger values of n is called the gamma function; $n! = \Gamma(n+1)$. The gamma function generalization is due to, who else, Euler. (However, the fact that $n! = \Gamma(n+1)$ instead of $n! = \Gamma(n)$ is due to the idiocy of Legendre.) In Legendre-resistant notation,

$$n! = \int_0^\infty t^n e^{-t}\,\mathrm{d}t$$

Straightforward integration shows that $0!$ is 1 as it should, and integration by parts shows that $(n+1)! = (n+1)n!$, which ensures that the integral also produces the correct value of $n!$ for any higher integer value of n than 0. The integral, however, exists for any real value of n above -1, not just integers. The values of the integral are always positive, tending to positive infinity for both $n \downarrow -1$, (because the integral then blows up at small values of t), and for $n \uparrow \infty$, (because the integral then blows up at medium-large values of t). In particular, Stirling's formula says that for large positive n, $n!$ can be approximated as

$$n! \sim \sqrt{2\pi n} n^n e^{-n} [1 + \ldots]$$

where the value indicated by the dots becomes negligibly small for large n. The function $n!$ can be extended further to any complex value of n, except the negative integer values of n, where $n!$ is infinite, but is then no longer positive. Euler's integral can be done for $n = -\frac{1}{2}$ by making the change of variables $\sqrt{t} = u$, producing the integral $\int_0^\infty 2e^{-u^2}\,\mathrm{d}u$, or $\int_{-\infty}^\infty e^{-u^2}\,\mathrm{d}u$, which equals $\sqrt{\int_{-\infty}^\infty e^{-x^2}\,\mathrm{d}x \int_{-\infty}^\infty e^{-y^2}\,\mathrm{d}y}$ and the integral under the square root can be done analytically using polar coordinates. The result is that

$$(-\frac{1}{2})! = \int_{-\infty}^\infty e^{-u^2}\,\mathrm{d}u = \sqrt{\pi}$$

To get $\frac{1}{2}!$, multiply by $\frac{1}{2}$, since $n! = n(n-1)!$.

A double exclamation mark may mean every second item is skipped, e.g. $5!! = 1 \times 3 \times 5$. In general, $(2n+1)!! = (2n+1)!/2^n n!$. Of course, $5!!$ should logically mean $(5!)!$. Logic would indicate that $5 \times 3 \times 1$ should be indicated by something like $5!'$. But what is logic in physics?

| May indicate:

- The magnitude or absolute value of the number or vector, if enclosed between a pair of them.
- The determinant of a matrix, if enclosed between a pair of them.

- The norm of the function, if enclosed between two pairs of them.
- The end of a bra or start of a ket.
- A visual separator in inner products.

$|\ldots\rangle$ A "ket" is used to indicate some state. For example, $|l\,m\rangle$ indicates an angular momentum state with azimuthal quantum number l and magnetic quantum number m. Similarly, $|{}^1\!/_2\ {}^{-1}\!/_2\rangle$ is the spin-down state of a particle with spin $\frac{1}{2}$. Other common ones are $|\underline{x}\rangle$ for the position eigenfunction $\underline{x}$, i.e. $\delta(x-\underline{x})$, $|1\mathrm{s}\rangle$ for the 1s or ψ_{100} hydrogen state, $|2\mathrm{p}_z\rangle$ for the $2\mathrm{p}_z$ or ψ_{210} state, etcetera. In short, whatever can indicate some state can be pushed into a ket.

$\langle\ldots|$ A "bra" is like a ket $|\ldots\rangle$, but appears in the left side of inner products, instead of the right one.

$\uparrow$ Indicates the "spin up" state. Mathematically, equals the function $\uparrow(S_z)$ which is by definition equal to 1 at $S_z = \frac{1}{2}\hbar$ and equal to 0 at $S_z = -\frac{1}{2}\hbar$. A spatial wave function multiplied by $\uparrow$ is a particle in that spatial state with its spin up. For multiple particles, the spins are listed with particle 1 first.

$\downarrow$ Indicates the "spin down" state. Mathematically, equals the function $\downarrow(S_z)$ which is by definition equal to 0 at $S_z = \frac{1}{2}\hbar$ and equal to 1 at $S_z = -\frac{1}{2}\hbar$. A spatial wave function multiplied by $\downarrow$ is a particle in that spatial state with its spin down. For multiple particles, the spins are listed with particle 1 first.

$\sum$ Summation symbol. Example: if in three dimensional space a vector $\vec{f}$ has components $f_1 = 2$, $f_2 = 1$, $f_3 = 4$, then $\sum_{\text{all } i} f_i$ stands for $2+1+4=7$.

One important thing to remember: the symbol used for the summation index does not make a difference: $\sum_{\text{all } j} f_j$ is exactly the same as $\sum_{\text{all } i} f_i$. So freely rename the index, but always make sure that the new name is not already used for something else in the part that it appears in. If you use the same name for two different things, it becomes a mess.

Related to that, $\sum_{\text{all } i} f_i$ is *not* something that depends on an index i. It is just a combined simple number. Like 7 in the example above. It is commonly said that the summation index "sums away."

$\prod$ (Not to be confused with Π further down.) Multiplication symbol. Example: if in three dimensional space a vector $\vec{f}$ has components $f_1 = 2$, $f_2 = 1$, $f_3 = 4$, then $\prod_{\text{all } i} f_i$ stands for $2 \times 1 \times 4 = 6$.

One important thing to remember: the symbol used for the multiplications index does not make a difference: $\prod_{\text{all } j} f_j$ is exactly the same as $\prod_{\text{all } i} f_i$.

So freely rename the index, but always make sure that the new name is not already used for something else in the part that it appears in. If you use the same name for two different things, it becomes a mess.

Related to that, $\prod_{\text{all } i} f_i$ is *not* something that depends on an index i. It is just a combined simple number. Like 6 in the example above. It is commonly said that the multiplication index "factors away." (By who?)

$\int$ Integration symbol, the continuous version of the summation symbol. For example,

$$\int_{\text{all } x} f(x)\,\mathrm{d}x$$

is the summation of $f(x)\,\mathrm{d}x$ over all infinitesimally small fragments $\mathrm{d}x$ that make up the entire x-range. For example, $\int_{x=0}^{2}(2+x)\,\mathrm{d}x$ equals $3 \times 2 = 6$; the average value of $2+x$ between $x=0$ and $x=2$ is 3, and the sum of all the infinitesimally small segments $\mathrm{d}x$ gives the total length 2 of the range in x from 0 to 2.

One important thing to remember: the symbol used for the integration variable does not make a difference: $\int_{\text{all } y} f(y)\,\mathrm{d}y$ is exactly the same as $\int_{\text{all } x} f(x)\,\mathrm{d}x$. So freely rename the integration variable, but always make sure that the new name is not already used for something else in the part it appears in. If you use the same name for two different things, it becomes a mess.

Related to that $\int_{\text{all } x} f(x)\,\mathrm{d}x$ is *not* something that depends on a variable x. It is just a combined number. Like 6 in the example above. It is commonly said that the integration variable "integrates away."

$\rightarrow$ May indicate:

- An approaching process. $\lim_{\varepsilon\to 0}$ indicates for practical purposes the value of the expression following the lim when ε is extremely small. Similarly, $\lim_{r\to\infty}$ indicates the value of the following expression when r is extremely large.
- The fact that the left side leads to, or implies, the right-hand side.

$\vec{}$ Vector symbol. An arrow above a letter indicates it is a vector. A vector is a quantity that requires more than one number to be characterized. Typical vectors in physics include position $\vec{r}$, velocity $\vec{v}$, linear momentum $\vec{p}$, acceleration $\vec{a}$, force $\vec{F}$, angular momentum $\vec{L}$, etcetera.

$\hat{}$ A hat over a letter in this book indicates that it is the operator, turning functions into other functions.

$'$ May indicate:

- A derivative of a function. Examples: $1' = 0$, $x' = 1$, $\sin'(x) = \cos(x)$, $\cos'(x) = -\sin(x)$, $(e^x)' = e^x$.
- A small or modified quantity.
- A quantity per unit length.

∇ The spatial differentiation operator nabla. In Cartesian coordinates:

$$\nabla \equiv \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) = \hat{\imath}\frac{\partial}{\partial x} + \hat{\jmath}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

Nabla can be applied to a scalar function f in which case it gives a vector of partial derivatives called the gradient of the function:

$$\operatorname{grad} f = \nabla f = \hat{\imath}\frac{\partial f}{\partial x} + \hat{\jmath}\frac{\partial f}{\partial y} + \hat{k}\frac{\partial f}{\partial z}.$$

Nabla can be applied to a vector in a dot product multiplication, in which case it gives a scalar function called the divergence of the vector:

$$\operatorname{div} \vec{v} = \nabla \cdot \vec{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$$

or in index notation

$$\operatorname{div} \vec{v} = \nabla \cdot \vec{v} = \sum_{i=1}^{3} \frac{\partial v_i}{\partial x_i}$$

Nabla can also be applied to a vector in a vectorial product multiplication, in which case it gives a vector function called the curl or rot of the vector. In index notation, the i-th component of this vector is

$$(\operatorname{curl} \vec{v})_i = (\operatorname{rot} \vec{v})_i = (\nabla \times \vec{v})_i = \frac{\partial v_{\bar{i}}}{\partial x_{\bar{\imath}}} - \frac{\partial v_{\bar{\imath}}}{\partial x_{\bar{i}}}$$

where $\bar{\imath}$ is the index following i in the sequence 123123..., and $\bar{i}$ the one preceding it (or the second following it).

The operator ∇^2 is called the Laplacian. In Cartesian coordinates:

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

Sometimes the Laplacian is indicated as Δ. In relativistic index notation it is equal to $\partial_i \partial^i$, with maybe a minus sign depending on who you talk with.

In non Cartesian coordinates, don't guess; look these operators up in a table book, [41, pp. 124-126]: . For example, in spherical coordinates,

$$\nabla = \hat{\imath}_r \frac{\partial}{\partial r} + \hat{\imath}_\theta \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{\imath}_\phi \frac{1}{r \sin\theta} \frac{\partial}{\partial \phi} \tag{N.2}$$

That allows the gradient of a scalar function f, i.e. ∇f, to be found immediately. But if you apply ∇ on a vector, you have to be very careful because you also need to differentiate $\hat{\imath}_r$, $\hat{\imath}_\theta$, and $\hat{\imath}_\phi$. In particular, the correct divergence of a vector $\vec{v}$ is

$$\nabla \cdot \vec{v} = \frac{1}{r^2} \frac{\partial r^2 v_r}{\partial r} + \frac{1}{r \sin\theta} \frac{\partial \sin\theta\, v_\theta}{\partial \theta} + \frac{1}{r \sin\theta} \frac{\partial v_\phi}{\partial \phi} \tag{N.3}$$

The curl $\nabla \times \vec{v}$ of the vector is

$$\frac{\hat{\imath}_r}{r \sin\theta} \left(\frac{\partial \sin\theta\, v_\phi}{\partial \theta} - \frac{\partial v_\theta}{\partial \phi} \right) + \frac{\hat{\imath}_\theta}{r} \left(\frac{1}{\sin\theta} \frac{\partial v_r}{\partial \phi} - \frac{\partial r\, v_\phi}{\partial r} \right) + \frac{\hat{\imath}_\phi}{r} \left(\frac{\partial r\, v_\theta}{\partial r} - \frac{\partial v_r}{\partial \theta} \right) \tag{N.4}$$

Finally the Laplacian is:

$$\nabla^2 = \frac{1}{r^2} \left\{ \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \phi^2} \right\} \tag{N.5}$$

See also "spherical coordinates."

Cylindrical coordinates are usually indicated as r, θ and z. Here z is the Cartesian coordinate, while r is the distance from the z-axis and θ the angle around the z axis. In two dimensions, i.e. without the z terms, they are usually called polar coordinates. In cylindrical coordinates:

$$\nabla = \hat{\imath}_r \frac{\partial}{\partial r} + \hat{\imath}_\theta \frac{1}{r} \frac{\partial}{\partial \theta} + \hat{\imath}_z \frac{\partial}{\partial z} \tag{N.6}$$

$$\nabla \cdot \vec{v} = \frac{1}{r} \frac{\partial r v_r}{\partial r} + \frac{1}{r} \frac{\partial v_\theta}{\partial \theta} + \frac{\partial v_z}{\partial z} \tag{N.7}$$

$$\nabla \times \vec{v} = \hat{\imath}_r \left(\frac{1}{r} \frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z} \right) + \hat{\imath}_\theta \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r} \right) + \frac{\hat{\imath}_z}{r} \left(\frac{\partial r v_\theta}{\partial r} - \frac{\partial v_r}{\partial \theta} \right) \tag{N.8}$$

$$\nabla^2 = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} + \frac{\partial^2}{\partial z^2} \tag{N.9}$$

□ The D'Alembertian is defined as

$$\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2}$$

where c is a constant called the wave speed. In relativistic index notation, $\Box$ is equal to $-\partial_\mu \partial^\mu$.

* A superscript star normally indicates a complex conjugate. In the complex conjugate of a number, every i is changed into a −i.

$<$ Less than.

$\leqslant$ Less than or equal.

$\langle \ldots \rangle$ May indicate:

- An inner product.
- An expectation value.

$>$ Greater than.

$\geqslant$ Greater than or equal.

$[\ldots]$ May indicate:

- A grouping of terms in a formula.
- A commutator. For example, $[A, B] = AB - BA$.

$=$ Equals sign. The quantity to the left is the same as the one to the right.

$\equiv$ Emphatic equals sign. Typically means "by definition equal" or "everywhere equal."

$\approx$ Indicates approximately equal. Read it as "is approximately equal to."

$\sim$ Indicates approximately equal. Often used when the approximation applies only when something is small or large. Read it as "is approximately equal to" or as "is asymptotically equal to."

$\propto$ Proportional to. The two sides are equal except for some unknown constant factor.

α (alpha) May indicate:

- The fine structure constant, $e^2/4\pi\epsilon_0\hbar c$, equal to $7.297\,352\,570\ 10^{-3}$, or about $1/137$, in value.
- A Dirac equation matrix.
- A nuclear decay mode in which a helium-4 nucleus is emitted.
- Internal conversion rate as fraction of the gamma decay rate.
- Some constant.
- Some angle.

- An eigenfunction of a generic operator A.
- A summation index.
- Component index of a vector.

$\boldsymbol{\beta}$ (beta) May indicate:

- A nuclear decay mode in which an electron (β^-) or positron (β^+) is emitted. Sometimes β^+ is taken to also include electron capture.
- A nuclear vibrational mode that maintains the axial symmetry of the nucleus.
- Some constant.
- Some angle.
- An eigenfunction of a generic operator B.
- A summation index.

$\boldsymbol{\Gamma}$ (Gamma) May indicate:

- The Gamma function. Look under "!" for details.
- The "width" or uncertainty in energy of an approximate energy eigenstate.
- Origin in wave number space.

$\boldsymbol{\gamma}$ (gamma) May indicate:

- Gyromagnetic ratio.
- Standard symbol for a photon of electromagnetic radiation.
- A nuclear de-excitation mode in which a photon is emitted.
- A nuclear vibrational mode that messes up the axial symmetry of the nucleus.
- Summation index.
- Integral in the tunneling WKB approximation.

$\boldsymbol{\Delta}$ (capital delta) May indicate:

- An increment in the quantity following it.
- A delta particle.
- Often used to indicate the Laplacian ∇^2.

$\boldsymbol{\delta}$ (delta) May indicate:

- With two subscripts, the "Kronecker delta", which by definition is equal to one if its two subscripts are equal, and zero in all other cases.
- Without two subscripts, the "Dirac delta function", which is infinite when its argument is zero, and zero if it is not. In addition the infinity is such that the integral of the delta function over its single nonzero point is unity. The delta function is not a normal function, but a distribution. It is best to think of it as the approximate function shown in the right hand side of figure 7.10 for a very, very, small positive value of ε.

  One often important way to create a three-dimensional delta function in spherical coordinates is to take the Laplacian of the function $-1/4\pi r$. Chapter 13.3 explains why. In two dimensions, take the Laplacian of $\ln(r)/2\pi$ to get a delta function.
- Often used to indicate a small amount of the following quantity, or of a small change in the following quantity. There are nuanced differences in the usage of δ, ∂ and d that are too much to go in here.
- Often used to indicate a second small quantity in addition to ε.

$\boldsymbol{\partial}$ (partial) Indicates a vanishingly small change or interval of the following variable. For example, $\partial f/\partial x$ is the ratio of a vanishingly small change in function f divided by the vanishingly small change in variable x that causes this change in f. Such ratios define derivatives, in this case the partial derivative of f with respect to x.

Also used in relativistic index notation, chapter 1.2.5.

$\boldsymbol{\epsilon}$ (epsilon) May indicate:

- ϵ_0 is the permittivity of space. Equal to $8.854\,187\,817...\ 10^{-12}$ C^2/J m. The exact value is $1/4\pi c^2\ 10^7$ C^2/J m, because of the exact SI definitions of ampere and speed of light.
- Scaled energy.
- Orbital energy.
- Lagrangian multiplier.
- A small quantity, if symbol ε is not available.

$\boldsymbol{\varepsilon}$ (variant of epsilon) May indicate:

- A very small quantity.
- The slop in energy conservation during a decay process.

$\boldsymbol{\eta}$ (eta) May be used to indicate a y-position of a particle.

Θ (capital theta) Used in this book to indicate some function of θ to be determined.

θ (theta) May indicate:

- In spherical coordinates, the angle from the chosen z axis, with apex at the origin.
- z-position of a particle.
- A generic angle, like the one between the vectors in a cross or dot product.
- Integral acting as an angle in the classical WKB approximation.
- Integral acting as an angle in the adiabatic approximation.

ϑ (variant of theta) An alternate symbol for θ.

κ (kappa) May indicate:

- A constant that physically corresponds to some wave number.
- A summation index.
- Thermal conductivity.

Λ (Lambda) May indicate:

- Lorentz transformation matrix.

λ (lambda) May indicate:

- Wave length.
- Decay constant.
- A generic eigenvalue.
- Entry of a Lorentz transformation.
- Scaled square momentum.
- Some multiple of something.

μ (mu) May indicate:

- Magnetic dipole moment:
  Alpha particle: 0 (spin is zero).
  Deuteron: $0.433\,073\,49\ 10^{-26}$ J/T or $0.857\,438\,231\ \mu_{\mathrm{N}}$.
  Electron: $-9.284\,764\,3\ 10^{-24}$ J/T or $-1.001\,159\,652\,180\,8\ \mu_{\mathrm{B}}$.
  Helion: $-1.074\,617\,49\ 10^{-26}$ J/T or $-2.127\,625\,306\ \mu_{\mathrm{N}}$.
  Neutron: $-0.966\,236\,5\ 10^{-26}$ J/T or $-1.913\,042\,7\ \mu_{\mathrm{N}}$.
  Proton: $1.410\,606\,74\ 10^{-26}$ J/T or $2.792\,847\,36\ \mu_{\mathrm{N}}$.
  Triton: $1.504\,609\,45\ 10^{-26}$ J/T or $2.978\,962\,45\ \mu_{\mathrm{N}}$.

- $\mu_B = e\hbar/2m_e = 9.274\,009\,7\ 10^{-24}$ J/T or $5.788\,381\,807\ 10^{-5}$ eV/T is the Bohr magneton.
- $\mu_N = e\hbar/2m_p = 5.050\,783\,5\ 10^{-27}$ J/T or $3.152\,451\,261\ 10^{-8}$ eV/T is the nuclear magneton.
- A summation index.
- Chemical potential/molar Gibbs free energy.

$\boldsymbol{\nu}$ (nu) May indicate:

- Electron neutrino.
- Scaled energy eigenfunction number in solids.
- A summation index.
- Strength of a delta function potential.

$\boldsymbol{\xi}$ (xi) May indicate:

- Scaled argument of the one-dimensional harmonic oscillator eigenfunctions.
- x-position of a particle.
- A summation or integration index.

$\boldsymbol{\varPi}$ (Oblique Pi) (Not to be confused with $\prod$ described higher up.) Parity operator. Replaces $\vec{r}$ by $-\vec{r}$. That is equivalent to a mirroring in a mirror through the origin, followed by a 180° rotation around the axis normal to the mirror.

$\boldsymbol{\pi}$ (pi) May indicate:

- A constant with value 3.141 592 653 589 793 238 462. . . .
  The area of a circle of radius r is πr^2 and its perimeter is $2\pi r$.
  The volume of a sphere of radius r is $\frac{4}{3}\pi r^3$ and its surface is $4\pi r^2$.
  A 180° angle expressed in radians is π.
  Note also that $e^{\pm i\pi} = -1$ and $e^{\pm i2\pi} = 1$.
- A chemical bond that looks from the side like a p state.
- A particle involved in the forces keeping the nuclei of atoms together (π-meson or pion for short).
- Parity.

$\tilde{\boldsymbol{\pi}}$ Canonical momentum density.

$\boldsymbol{\rho}$ (rho) May indicate:

- Electric charge per unit volume.
- Scaled radial coordinate.
- Radial coordinate.
- Eigenfunction of a rotation operator $\mathcal{R}$.
- Mass-base density.
- Energy density of electromagnetic radiation.

$\boldsymbol{\sigma}$ (sigma) May indicate:

- A standard deviation of a value.
- A chemical bond that looks like an s state when seen from the side.
- Pauli spin matrix.
- Surface tension.
- Electrical conductivity.
- $\sigma_{\mathrm{B}} = 5.670\,37\ \mathrm{W/m^2\ K^4}$ is the Stefan-Boltzmann

$\boldsymbol{\tau}$ (tau) May indicate:

- A time or time interval.
- Life time or half life.
- Some coefficient.

$\boldsymbol{\Phi}$ (capital phi) May indicate:

- Some function of ϕ to be determined.
- The momentum-space wave function.
- Relativistic electromagnetic potential.

$\boldsymbol{\phi}$ (phi) May indicate:

- In spherical coordinates, the angle around the chosen z axis. Increasing ϕ by 2π encircles the z-axis exactly once.
- A phase angle.
- Something equivalent to an angle.
- Field operator $\phi(\vec{r})$ annihilates a particle at position $\vec{r}$ while $\phi^\dagger(\vec{r})$ creates one at that position.

$\boldsymbol{\varphi}$ (variant of phi) May indicate:

- A change in angle ϕ.
- An alternate symbol for ϕ.
- An electrostatic potential.
- An electrostatic quantum field.
- A hypothetical selectostatic quantum field.

$\boldsymbol{\chi}$ (chi) May indicate

- Spinor component.
- Gauge function of electromagnetic field.

$\boldsymbol{\Psi}$ (capital psi) Upper case psi is used for the wave function.

$\boldsymbol{\psi}$ (psi) Typically used to indicate an energy eigenfunction. Depending on the system, indices may be added to distinguish different ones. In some cases ψ might be used instead of Ψ to indicate a system in an energy eigenstate. Let me know and I will change it. A system in an energy eigenstate should be written as $\Psi = c\psi$, not ψ, with c a constant of magnitude 1.

$\boldsymbol{\Omega}$ (Omega) May indicate:

- Solid angle. See "angle" and "spherical coordinates."

$\boldsymbol{\omega}$ (omega) May indicate:

- Angular frequency of the classical harmonic oscillator. Equal to $\sqrt{c/m}$ where c is the spring constant and m the mass.
- Angular frequency of a system.
- Angular frequency of light waves.
- Perturbation frequency,
- Any quantity having units of frequency, 1/s.

$\boldsymbol{A}$ May indicate:

- Repeatedly used to indicate the operator for a generic physical quantity a, with eigenfunctions α.
- Electromagnetic vector potential, or four vector potential.
- Einstein A coefficient.
- Some generic matrix.
- Some constant.

- Area.

Å Ångstrom. Equal to 10^{-10} m.

a May indicate:

- The value of a generic physical quantity with operator A
- The amplitude of the spin-up state
- The amplitude of the first state in a two-state system.
- Acceleration.
- Start point of an integration interval.
- The first of a pair of particles.
- Some coefficient.
- Some constant.
- Absorptivity of electromagnetic radiation.
- Annihilation operator $\widehat{a}$ or creation operator $\widehat{a}^\dagger$.
- Bohr radius of helium ion.

***a*$_0$** May indicate:

- Bohr radius, $4\pi\epsilon_0\hbar^2/m_e e^2$ or 0.529 177 210 9 Å, with Å= 10^{-10} m. Comparable in size to atoms, and a good size to use to simplify various formulae.
- The initial value of a coefficient a.

absolute May indicate:

- The absolute value of a real number a is indicated by $|a|$. It equals a is a is positive or zero and $-a$ if a is negative.
- The absolute value of a complex number a is indicated by $|a|$. It equals the length of the number plotted as a vector in the complex plane. This simplifies to above definition if a is real.
- An absolute temperature is a temperature measured from absolute zero. At absolute zero all systems are in their ground state. Absolute zero is −273.15 °C in degrees Centrigrade (Celsius). The SI absolute temperature scale is degrees Kelvin, K. Absolute zero temperature is 0 K, while 0 °C is 273.15 K.

adiabatic An adiabatic process is a process in which there is no heat transfer with the surroundings. If the process is also reversible, it is called isentropic. Typically, these processes are fairly quick, in order not to give heat conduction enough time to do its stuff, but not so excessively quick that they become irreversible.

Adiabatic processes in quantum mechanics are defined quite differently to keep students on their toes. See chapter 7.1.5. These processes are very slow, to give the system all possible time to adjust to its surroundings. Of course, quantum physicist were not aware that the same term had already been used for a hundred years or so for relatively fast processes. They assumed they had just invented a great new term!

adjoint The adjoint A^H or $A^\dagger$ of an operator is the one you get if you take it to the other side of an inner product. (While keeping the value of the inner product the same regardless of whatever two vectors or functions may be involved.) Hermitian operators are "self-adjoint;"they do not change if you take them to the other side of an inner product. "Skew-Hermitian"operators just change sign. "Unitary operators"change into their inverse when taken to the other side of an inner product. Unitary operators generalize rotations of vectors: an inner product of vectors is the same whether you rotate the first vector one way, or the second vector the opposite way. Unitary operators preserve inner products (when applied to both vectors or functions). Fourier transforms are unitary operators on account of the Parseval equality that says that inner products are preserved.

amplitude Everything in quantum mechanics is an amplitude. However, most importantly, the "quantum amplitude" gives the coefficient of a state in a wave function. For example, the usual quantum wave function gives the quantum amplitude that the particle is at the given position.

angle Consider two semi-infinite lines extending from a common intersection point. Then the angle between these lines is defined in the following way: draw a unit circle in the plane of the lines and centered at their intersection point. The angle is then the length of the circular arc that is in between the lines. More precisely, this gives the angle in radians, rad. Sometimes an angle is expressed in degrees, where 2π rad is taken to be 360°. However, using degrees is usually a very bad idea in science.

In three dimensions, you may be interested in the so-called "solid angle" Ω inside a conical surface. This angle is defined in the following way: draw a sphere of unit radius centered at the apex of the conical surface. Then the solid angle is the area of the spherical surface that is inside the cone. Solid angles are in steradians. The cone does not need to be a circular

one, (i.e. have a circular cross section), for this to apply. In fact, the most common case is the solid angle corresponding to an infinitesimal element $\mathrm{d}\theta \times \mathrm{d}\phi$ of spherical coordinate system angles. In that case the surface of the unit sphere inside the conical surface is is approximately rectangular, with sides $\mathrm{d}\theta$ and $\sin(\theta)\mathrm{d}\phi$. That makes the enclosed solid angle equal to $\mathrm{d}\Omega = \sin(\theta)\mathrm{d}\theta\mathrm{d}\phi$.

B May indicate:

- Repeatedly used to indicate a generic second operator or matrix.
- Einstein B coefficient.
- Some constant.

$\mathcal{B}$ May indicate:

- Magnetic field strength.

b May indicate:

- Repeatedly used to indicate the amplitude of the spin-down state
- Repeatedly used to indicate the amplitude of the second state in a two-state system.
- End point of an integration interval.
- The second of a pair of particles.
- Some coefficient.
- Some constant.

basis A basis is a minimal set of vectors or functions that you can write all other vectors or functions in terms of. For example, the unit vectors $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ are a basis for normal three-dimensional space. Every three-dimensional vector can be written as a linear combination of the three.

C May indicate:

- A third matrix or operator.
- A variety of different constants.

°C Degrees Centigrade. A commonly used temperature scale that has the value -273.15 °C instead of zero when systems are in their ground state. Recommendation: use degrees Kelvin (K) instead. However, differences in temperature are the same in Centigrade as in Kelvin.

c May indicate:

- The speed of light, 299 792 458 m/s exactly (by definition of the velocity unit).
- Speed of sound.
- Spring constant.
- A variety of different constants.

Cauchy-Schwartz inequality The Cauchy-Schwartz inequality describes a limitation on the magnitude of inner products. In particular, it says that for any f and g,

$$|\langle f|g\rangle| \leq \sqrt{\langle f|f\rangle}\sqrt{\langle g|g\rangle}$$

In words, the magnitude of an inner product $\langle f|g\rangle$ is at most the magnitude (i.e. the length or norm) of f times the one of g. For example, if f and g are real vectors, the inner product is the dot product and you have $f \cdot g = |f||g|\cos\theta$, where $|f|$ is the length of vector f and $|g|$ the one of g, and θ is the angle in between the two vectors. Since a cosine is less than one in magnitude, the Cauchy-Schwartz inequality is therefore true for vectors.

But it is true even if f and g are functions. To prove it, first recognize that $\langle f|g\rangle$ may in general be a complex number, which according to (2.6) must take the form $e^{i\alpha}|\langle f|g\rangle|$ where α is some real number whose value is not important, and that $\langle g|f\rangle$ is its complex conjugate $e^{-i\alpha}|\langle f|g\rangle|$. Now, (yes, this is going to be some convoluted reasoning), look at

$$\left\langle f + \lambda e^{-i\alpha} g \middle| f + \lambda e^{-i\alpha} g \right\rangle$$

where λ is any real number. The above dot product gives the square magnitude of $f + \lambda e^{-i\alpha} g$, so it can never be negative. But if you multiply out, you get

$$\langle f|f\rangle + 2|\langle f|g\rangle|\lambda + \langle g|g\rangle\lambda^2$$

and if this quadratic form in λ is never negative, its discriminant must be less or equal to zero:

$$|\langle f|g\rangle|^2 \leq \langle f|f\rangle\langle g|g\rangle$$

and taking square roots gives the Cauchy-Schwartz inequality.

Classical Can mean any older theory. In this work, most of the time it either means "nonquantum," or "nonrelativistic."

cos The cosine function, a periodic function oscillating between 1 and -1 as shown in [41, pp. 40-]. See also "sin."

curl The curl of a vector $\vec{v}$ is defined as $\operatorname{curl}\vec{v} = \operatorname{rot}\vec{v} = \nabla \times \vec{v}$.

$\boldsymbol{D}$ May indicate:

- Difference in wave number values.

$\vec{\boldsymbol{D}}$ Primitive (translation) vector of a reciprocal lattice.

$\boldsymbol{\mathcal{D}}$ Density of states.

D Often used to indicate a state with two units of orbital angular momentum.

$\boldsymbol{d}$ May indicate:

- The distance between the protons of a hydrogen molecule.
- The distance between the atoms or lattice points in a crystal.
- A constant.

$\vec{\boldsymbol{d}}$ Primitive (translation) vector of a crystal lattice.

d Indicates a vanishingly small change or interval of the following variable. For example, $\mathrm{d}x$ can be thought of as a small segment of the x-axis.

In three dimensions, $\mathrm{d}^3\vec{r} \equiv \mathrm{d}x\mathrm{d}y\mathrm{d}z$ is an infinitesimal volume element. The symbol $\int$ means that you sum over all such infinitesimal volume elements.

derivative A derivative of a function is the ratio of a vanishingly small change in a function divided by the vanishingly small change in the independent variable that causes the change in the function. The derivative of $f(x)$ with respect to x is written as $\mathrm{d}f/\mathrm{d}x$, or also simply as f'. Note that the derivative of function $f(x)$ is again a function of x: a ratio f' can be found at every point x. The derivative of a function $f(x, y, z)$ with respect to x is written as $\partial f/\partial x$ to indicate that there are other variables, y and z, that do not vary.

determinant The determinant of a square matrix A is a single number indicated by $|A|$. If this number is nonzero, $A\vec{v}$ can be any vector $\vec{w}$ for the right choice of $\vec{v}$. Conversely, if the determinant is zero, $A\vec{v}$ can only produce a very limited set of vectors, though if it can produce a vector w, it can do so for multiple vectors $\vec{v}$.

There is a recursive algorithm that allows you to compute determinants from increasingly bigger matrices in terms of determinants of smaller matrices. For a 1×1 matrix consisting of a single number, the determinant is simply that number:

$$|a_{11}| = a_{11}$$

(This determinant should not be confused with the absolute value of the number, which is written the same way. Since you normally do not deal with 1×1 matrices, there is normally no confusion.) For 2×2 matrices, the determinant can be written in terms of 1×1 determinants:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = +a_{11} \begin{vmatrix} & a_{22} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & \end{vmatrix}$$

so the determinant is $a_{11}a_{22} - a_{12}a_{21}$ in short. For 3×3 matrices, you have

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} =$$

$$+a_{11} \begin{vmatrix} & a_{22} & a_{23} \\ & a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

and you already know how to work out those 2×2 determinants, so you now know how to do 3×3 determinants. Written out fully:

$$a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

For 4×4 determinants,

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} =$$

$$+a_{11} \begin{vmatrix} & a_{22} & a_{23} & a_{24} \\ & a_{32} & a_{33} & a_{34} \\ & a_{42} & a_{43} & a_{44} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} & a_{24} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{vmatrix}$$

$$+a_{13} \begin{vmatrix} a_{21} & a_{22} & a_{24} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44} \end{vmatrix} - a_{14} \begin{vmatrix} a_{21} & a_{22} & a_{23} & \\ a_{31} & a_{32} & a_{33} & \\ a_{41} & a_{42} & a_{43} & \end{vmatrix}$$

Etcetera. Note the alternating sign pattern of the terms.

As you might infer from the above, computing a good size determinant takes a large amount of work. Fortunately, it is possible to simplify the matrix to put zeros in suitable locations, and that can cut down the work of finding the determinant greatly. You are allowed to use the following manipulations without seriously affecting the computed determinant:

1. You can "transpose" the matrix, i.e. change its columns into its rows.
2. You can create zeros in a row by subtracting a suitable multiple of another row.
3. You can also swap rows, as long as you remember that each time that you swap two rows, it will flip over the sign of the computed determinant.
4. You can also multiply an entire row by a constant, but that will multiply the computed determinant by the same constant.

Applying these tricks in a systematic way, called "Gaussian elimination" or "reduction to upper triangular form", you can eliminate all matrix coefficients a_{ij} for which j is less than i, and that makes evaluating the determinant pretty much trivial.

div(ergence) The divergence of a vector $\vec{v}$ is defined as $\operatorname{div} \vec{v} = \nabla \cdot \vec{v}$.

E May indicate:

- The total energy. Possible values are the eigenvalues of the Hamiltonian.
- $E_n = E_1/n^2 = -m_e e^4/32\pi^2\epsilon_0^2\hbar^2 n^2 = -\hbar^2/2m_e a_0^2 n^2 = \frac{1}{2}\alpha^2 m_e c^2/n^2$ may indicate the nonrelativistic (Bohr) energy levels of the hydrogen atom. The ground state energy E_1 equals -13.605 692 5 eV. This does not include relativistic and proton motion corrections.
- Internal energy of a substance.

$\mathcal{E}$ May indicate:

- Electric field strength.

e May indicate:

- The basis for the natural logarithms. Equal to 2.718 281 828 459... This number produces the "exponential function" e^x, or $\exp(x)$, or in words "e to the power x", whose derivative with respect to x is again e^x. If a is a constant, then the derivative of e^{ax} is ae^{ax}. Also, if a is an ordinary real number, then e^{ia} is a complex number with magnitude 1.
- The magnitude of the charge of an electron or proton, equal to $1.602\,176\,57\ 10^{-19}$ C.
- Emissivity of electromagnetic radiation.
- Often used to indicate a unit vector.

- A superscript e may indicate a single-electron quantity.
- Specific internal energy of a substance.

e May indicate

- Subscript e may indicate an electron.

***e*^{iax}** Assuming that a is an ordinary real number, and x a real variable, e^{iax} is a complex function of magnitude one. The derivative of e^{iax} with respect to x is iae^{iax}

eigenvector A concept from linear algebra. A vector $\vec{v}$ is an eigenvector of a matrix A if $\vec{v}$ is nonzero and $A\vec{v} = \lambda\vec{v}$ for some number λ called the corresponding eigenvalue.

The basic quantum mechanics section of this book avoids linear algebra completely, and the advanced part almost completely. The few exceptions are almost all two-dimensional matrix eigenvalue problems. In case you did not have any linear algebra, here is the solution: the two-dimensional matrix eigenvalue problem

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \vec{v} = \lambda \vec{v}$$

has eigenvalues that are the two roots of the quadratic equation

$$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0$$

The corresponding eigenvectors are

$$\vec{v}_1 = \begin{pmatrix} a_{12} \\ \lambda_1 - a_{11} \end{pmatrix} \qquad \vec{v}_2 = \begin{pmatrix} \lambda_2 - a_{22} \\ a_{21} \end{pmatrix}$$

On occasion you may have to swap λ_1 and λ_2 to use these formulae. If λ_1 and λ_2 are equal, there might not be two eigenvectors that are not multiples of each other; then the matrix is called defective. However, Hermitian matrices are never defective.

See also "matrix" and "determinant."

eV The electron volt, a commonly used unit of energy. Its value is equal to $1.602\,176\,57\ 10^{-19}$ J.

exponential function A function of the form $e^{\cdots}$, also written as $\exp(\ldots)$. See "function" and "e."

F May indicate:

- The force in Newtonian mechanics. Equal to the negative gradient of the potential. Quantum mechanics is formulated in terms of potentials, not forces.
- The anti-derivative of some function f.
- Some function.
- Helmholtz free energy.

$\boldsymbol{\mathcal{F}}$ Fock operator.

$\boldsymbol{f}$ May indicate:

- A generic function.
- A generic vector.
- A fraction.
- The resonance factor.
- Specific Helmholtz free energy.
- Frequency.

function A mathematical object that associates values with other values. A function $f(x)$ associates every value of x with a value f. For example, the function $f(x) = x^2$ associates $x = 0$ with $f = 0$, $x = \frac{1}{2}$ with $f = \frac{1}{4}$, $x = 1$ with $f = 1$, $x = 2$ with $f = 4$, $x = 3$ with $f = 9$, and more generally, any arbitrary value of x with the square of that value x^2. Similarly, function $f(x) = x^3$ associates any arbitrary x with its cube x^3, $f(x) = \sin(x)$ associates any arbitrary x with the sine of that value, etcetera.

One way of thinking of a function is as a procedure that allows you, whenever given a number, to compute another number.

A wave function $\Psi(x, y, z)$ associates each spatial position (x, y, z) with a wave function value. Going beyond mathematics, its square magnitude associates any spatial position with the relative probability of finding the particle near there.

functional A functional associates entire functions with single numbers. For example, the expectation energy is mathematically a functional: it associates any arbitrary wave function with a number: the value of the expectation energy if physics is described by that wave function.

$\boldsymbol{G}$ May indicate:

- Gibbs free energy.
- Newton's constant of gravitation, $6.673\,8\ 10^{-11}$ m^3/kg s^2.

g May indicate:

- A second generic function or a second generic vector.
- The strength of gravity, by definition equal to 9.806 65 m/s^2 exactly under standard conditions on the surface of the earth.
- The g-factor, a nondimensional constant that indicates the gyromagnetic ratio relative to charge and mass. For the electron $g_e = -2.002\,319\,304\,361\,5$. For the proton $g_p = 5.585\,694\,71$. For the neutron, based on the mass and charge of the proton, $g_n = -3.826\,085\,5$.
- Specific Gibbs free energy/chemical potential.

Gauss' Theorem This theorem, also called divergence theorem or Gauss-Ostrogradsky theorem, says that for a continuously differentiable vector $\vec{v}$,

$$\int_V \nabla \cdot \vec{v}\,\mathrm{d}V = \int_A \vec{v} \cdot \vec{n}\,\mathrm{d}A$$

where the first integral is over the volume of an arbitrary region and the second integral is over all the surface area of that region; $\vec{n}$ is at each point found as the unit vector that is normal to the surface at that point.

grad(ient) The gradient of a scalar f is defined as $\operatorname{grad} f = \nabla f$.

H May indicate:

- The Hamiltonian, or total energy, operator. Its eigenvalues are indicated by E.
- H_n stands for the n-th order Hermite polynomial.
- Enthalpy.

ħ May indicate:

- The original Planck constant $h = 2\pi\hbar$.
- h_n is a one-dimensional harmonic oscillator eigenfunction.
- Single-electron Hamiltonian.
- Specific enthalpy.

ħ The reduced Planck constant, equal to 1.054 571 73 10^{-34} J s. A measure of the uncertainty of nature in quantum mechanics. Multiply by 2π to get the original Planck constant h. For nuclear physics, a frequently helpful value is $\hbar c = 197.326\,972$ MeV fm.

hypersphere A hypersphere is the generalization of the normal three-dimensional sphere to n-dimensional space. A sphere of radius R in three-dimensional space consists of all points satisfying

$$r_1^2 + r_2^2 + r_3^2 \leqslant R^2$$

where r_1, r_2, and r_3 are Cartesian coordinates with origin at the center of the sphere. Similarly a hypersphere in n-dimensional space is *defined* as all points satisfying

$$r_1^2 + r_2^2 + \ldots + r_n^2 \leqslant R^2$$

So a two-dimensional "hypersphere" of radius R is really just a circle of radius R. A one-dimensional "hypersphere" is really just the line segment $-R \leqslant x \leqslant -R$.

The "volume" $\mathcal{V}_n$ and surface "area" A_n of an n-dimensional hypersphere is given by

$$\mathcal{V}_n = C_n R^n \qquad A_n = nC_n R^{n-1}$$

$$C_n = \begin{cases} (2\pi)^{n/2}/2 \times 4 \times 6 \times \ldots \times n & \text{if } n \text{ is even} \\ (2\pi)^{(n-1)/2} 2/1 \times 3 \times 5 \times \ldots \times n & \text{if } n \text{ is odd} \end{cases}$$

(This is readily derived recursively. For a sphere of unit radius, note that the n-dimensional "volume" is an integration of $n-1$-dimensional volumes with respect to r_1. Then renotate r_1 as $\sin\phi$ and look up the resulting integral in a table book. The formula for the area follows because $\mathcal{V} = \int A \mathrm{d}r$ where r is the distance from the origin.) In three dimensions, $C_3 = 4\pi/3$ according to the above formula. That makes the three-dimensional "volume" $4\pi R^3/3$ equal to the actual volume of the sphere, and the three-dimensional "area" $4\pi R^2$ equal to the actual surface area. On the other hand in two dimensions, $C_2 = \pi$. That makes the two-dimensional "volume" πR^2 really the *area* of the circle. Similarly the two-dimensional surface "area" $2\pi R$ is really the perimeter of the circle. In one dimensions $C_1 = 2$ and the "volume" $2R$ is really the length of the interval, and the "area" 2 is really its number of end points.

Often the infinitesimal n-dimensional "volume" element $\mathrm{d}^n\vec{r}$ is needed. This is the infinitesimal integration element for integration over all coordinates. It is:

$$\mathrm{d}^n\vec{r} = \mathrm{d}r_1 \mathrm{d}r_2 \ldots \mathrm{d}r_n = \mathrm{d}A_n \mathrm{d}r$$

Specifically, in two dimensions:

$$\mathrm{d}^2\vec{r} = \mathrm{d}r_1 \mathrm{d}r_2 = \mathrm{d}x\mathrm{d}y = (r\,\mathrm{d}\theta)\mathrm{d}r = \mathrm{d}A_2 \mathrm{d}r$$

while in three dimensions:

$$\mathrm{d}^3\vec{r} = \mathrm{d}r_1 \mathrm{d}r_2 \mathrm{d}r_3 = \mathrm{d}x\mathrm{d}y\mathrm{d}z = (r^2 \sin\theta\,\mathrm{d}\theta\mathrm{d}\phi)\mathrm{d}r = \mathrm{d}A_3 \mathrm{d}r$$

The expressions in parentheses are $\mathrm{d}A_2$ in polar coordinates, respectively $\mathrm{d}A_3$ in spherical coordinates.

I May indicate:

- The number of electrons or particles.
- Electrical current.
- Unit matrix or operator, which does not do anything. See "matrix."
- I_A is Avogadro's number, $6.022\,141\,3\ 10^{26}$ particles per kmol. (More standard symbols are N_A or L, but they are incompatible with the general notations in this book.)

$\Im$ The imaginary part of a complex number. If $c = c_r + \mathrm{i}c_i$ with c_r and c_i real numbers, then $\Im(c) = c_i$. Note that $c - c^* = 2\mathrm{i}\Im(c)$.

I May indicate:

- $\mathcal{I}$ is radiation energy intensity.
- $\mathcal{I}_\mathrm{R}$ is moment of inertia.

i May indicate:

- The number of a particle.
- A summation index.
- A generic index or counter.

Not to be confused with i.

$\hat{\boldsymbol{\imath}}$ The unit vector in the x-direction.

i The standard square root of minus one: $\mathrm{i} = \sqrt{-1}$, $\mathrm{i}^2 = -1$, $1/\mathrm{i} = -\mathrm{i}$, $\mathrm{i}^* = -\mathrm{i}$.

index notation A more concise and powerful way of writing vector and matrix components by using a numerical index to indicate the components. For Cartesian coordinates, you might number the coordinates x as 1, y as 2, and z as 3. In that case, a sum like $v_x + v_y + v_z$ can be more concisely written as $\sum_i v_i$. And a statement like $v_x \neq 0$, $v_y \neq 0$, $v_z \neq 0$ can be more compactly written as $v_i \neq 0$. To really see how it simplifies the notations, have a look at the matrix entry. (And that one shows only 2 by 2 matrices. Just imagine 100 by 100 matrices.)

iff Emphatic "if." Should be read as "if and only if."

integer Integer numbers are the whole numbers: $\ldots, -2, -1, 0, 1, 2, 3, 4, \ldots$.

inverse (Of matrices or operators.) If an operator A converts a vector or function f into a vector or function g, then the inverse of the operator A^{-1} converts g back into f. For example, the operator 2 converts vectors or functions into two times themselves, and its inverse operator $\frac{1}{2}$ converts these back into the originals. Some operators do not have inverses. For example, the operator 0 converts all vectors or functions into zero. But given zero, there is no way to figure out what function or vector it came from; the inverse operator does not exist.

irrotational A vector $\vec{v}$ is irrotational if its curl $\nabla \times \vec{v}$ is zero.

iso Means "equal" or "constant."

- Isenthalpic: constant enthalpy.
- Isentropic: constant entropy. This is a process that is both adiabatic and reversible.
- Isobaric: constant pressure.
- Isochoric: constant (specific) volume.
- Isospin: you don't want to know.
- Isothermal: constant temperature.

isolated An isolated system is one that does not interact with its surroundings in any way. No heat is transfered with the surroundings, no work is done on or by the surroundings.

$\boldsymbol{J}$ May indicate:

- Total angular momentum.
- Number of nuclei in a quantum computation of electronic structure.

$\boldsymbol{j}$ May indicate:

- The azimuthal quantum number of total angular momentum, including both orbital and spin contributions.
- $\vec{j}$ is electric current density.
- The number of a nucleus in a quantum computation.
- A summation index.
- A generic index or counter.

$\hat{\boldsymbol{j}}$ The unit vector in the y-direction.

$\boldsymbol{K}$ May indicate:

- An exchange integral in Hartree-Fock.
- Maximum wave number value.

$\mathcal{K}$ Thomson (Kelvin) coefficient.

K May indicate:

- The atomic states or orbitals with theoretical Bohr energy E_1
- Degrees Kelvin.

k May indicate:

- A wave number. A wave number is a measure for how fast a periodic function oscillates with variations in spatial position. In quantum mechanics, k is normally defined as $\sqrt{2m(E-V)}/\hbar$. The vector $\vec{k}$ is not to be confused with the unit vector in the z-direction $\hat{k}$.
- A generic summation index.

$\hat{k}$ The unit vector in the z-direction.

k_{B} Boltzmann constant. Equal to $1.380\,649\ 10^{-23}$ J/K. Relates absolute temperature to a typical unit of heat motion energy.

kmol A kilo mole refers to $6.022\,141\,3\ 10^{26}$ atoms or molecules. The weight of this many particles is about the number of protons and neutrons in the atom nucleus/molecule nuclei. So a kmol of hydrogen atoms has a mass of about 1 kg, and a kmol of hydrogen molecules about 2 kg. A kmol of helium atoms has a mass of about 4 kg, since helium has two protons and two neutrons in its nucleus. These numbers are not very accurate, not just because the electron masses are ignored, and the free neutron and proton masses are somewhat different, but also because of relativity effects that cause actual nuclear masses to deviate from the sum of the free proton and neutron masses.

$\bar{L}$ May indicate:

- Angular momentum.
- Orbital angular momentum.

$\mathcal{L}$ Lagrangian.

L The atomic states or orbitals with theoretical Bohr energy E_2

l May indicate:

- The azimuthal quantum number of angular momentum.
- The azimuthal quantum number of orbital angular momentum. Here s is used for spin, and j for combined angular momentum.)
- A generic summation index.

$\boldsymbol{\ell}$ May indicate:

- The typical length in the harmonic oscillator problem.
- The dimensions of a solid block (with subscripts).
- A length.
- Multipole level in transitions.

$\boldsymbol{\pounds}$ Lagrangian density. This is best understood in the UK.

lim Indicates the final result of an approaching process. $\lim_{\varepsilon \to 0}$ indicates for practical purposes the value of the following expression when ε is extremely small.

linear combination A very generic concept indicating sums of objects times coefficients. For example, a position vector $\vec{r}$ in basic physics is the linear combination $x\hat{\imath} + y\hat{\jmath} + z\hat{k}$ with the objects the unit vectors $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$ and the coefficients the position coordinates x, y, and z. A linear combination of a set of functions $f_1(x), f_2(x), f_3(x), \ldots, f_n(x)$ would be the function

$$c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + \ldots c_n f_n(x)$$

where $c_1, c_2, c_3, \ldots, c_n$ are constants, i.e. independent of x.

linear dependence A set of vectors or functions is linearly dependent if at least one of the set can be expressed in terms of the others. Consider the example of a set of functions $f_1(x), f_2(x), \ldots, f_n(x)$. This set is linearly dependent if

$$c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + \ldots c_n f_n(x) = 0$$

where at least one of the constants $c_1, c_2, c_2, \ldots, c_n$ is nonzero. To see why, suppose that say c_2 is nonzero. Then you can divide by c_2 and rearrange to get

$$f_2(x) = -\frac{c_1}{c_2} f_1(x) - \frac{c_3}{c_2} f_3(x) - \ldots - \frac{c_n}{c_2} f_n(x)$$

That expresses $f_2(x)$ in terms of the other functions.

linear independence A set of vectors or functions is linearly independent if none of the set can be expressed in terms of the others. Consider the example of a set of functions $f_1(x), f_2(x), \ldots, f_n(x)$. This set is linearly independent if

$$c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) + \ldots c_n f_n(x) = 0$$

only if every one of the constants $c_1, c_2, c_3, \ldots, c_n$ is zero. To see why, assume that say $f_2(x)$ could be expressed in terms of the others,

$$f_2(x) = C_1 f_1(x) + C_3 f_3(x) + \ldots + C_n f_n(x)$$

Then taking $c_2 = 1$, $c_1 = -C_1$, $c_3 = -C_3$, $\ldots c_n = -C_n$, the condition above would be violated. So f_2 cannot be expressed in terms of the others.

M May indicate:

- Molecular mass. See "molecular mass."
- Figure of merit.

$\mathcal{M}$ Mirror operator.

M The atomic states or orbitals with theoretical Bohr energy E_3

m May indicate:

- Mass.
  - m_e: electron mass. Equal to $9.109\,382\,9\ 10^{-31}$ kg. The rest mass energy is 0.510 998 93 MeV.
  - m_p: proton mass. Equal to $1.672\,621\,78\ 10^{-27}$ kg. The rest mass energy is 938.272 013 MeV.
  - m_n: neutron mass. Equal to $1.674\,927\ 10^{-27}$ kg. The rest mass energy is 939.565 561 MeV.
  - Alpha particle: $6.644\,656\,8\ 10^{-27}$ kg or 3 727.379 24 MeV.
    Deuteron: $3.343\,583\,5\ 10^{-27}$ kg or 1 875.612 86 MeV.
    Helion: $5.006\,412\,3\ 10^{-27}$ kg or 2 808.391 482 MeV.
  - $m_u = 1.660\,538\,92\ 10^{-27}$ kg is the atomic mass constant.
  - m: generic particle mass.
- The magnetic quantum number of angular momentum. The type odf angular momentum may be indicated by a subscript l for orbital, s for spin, or j for net (orbital plus spin).
- Number of a single-electron wave function.
- Number of rows in a matrix.

- A generic summation index or generic integer.

matrix A table of numbers.

As a simple example, a two-dimensional (or 2×2) matrix A is a table of four numbers called a_{11}, a_{12}, a_{21}, and a_{22}:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

unlike a two-dimensional vector $\vec{v}$, which would consist of only two numbers v_1 and v_2 arranged in a column:

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

(Such a vector can be seen as a "rectangular matrix" of size 2×1, but let's not get into that.) (Note that in quantum mechanics, if a vector is written as a column, considered the normal case, it is called a "ket" vector. If the complex conjugates of its numbers are written as a row, it is called a "bra" vector.)

In "index notation," a matrix A is a set of numbers, or "coefficients," $\{a_{ij}\}$ indexed by two indices. The first index i is the row number at which the coefficient $\{a_{ij}\}$ is found in matrix A, and the second index j is the column number. In index notation, a matrix turns a vector $\vec{v}$ into another vector $\vec{w} = A\vec{v}$ according to the recipe

$$w_i = \sum_{\text{all } j} a_{ij} v_j \quad \text{for all } i$$

where v_j stands for "the j-th component of vector $\vec{v}$," and w_i for "the i-th component of vector $\vec{w}$."

As an example, the product of A and $\vec{v}$ above is by definition

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 \\ a_{21}v_1 + a_{22}v_2 \end{pmatrix}$$

which is just another two-dimensional ket vector.

Note that in matrix multiplications, like in the example above, in geometric terms you take dot products between the rows of the first factor and the columns of the second factor.

To multiply two matrices together, just think of the columns of the second matrix as separate vectors. For example, to multiply two 2×2 matrices A and B together:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

which is another two-dimensional matrix.

(Note that you cannot normally swap the order of matric multiplication. The matrix BA is different from matrix AB. In the special case that AB and BA are the same and A and B have complete sets of eigenvectors, then they have a common complete set of eigenvectors, {D.18}.)

In index notation, if $C = AB$, then each coefficient c_{ij} of matrix C is given in terms of the coefficients of A and B as

$$c_{ij} = \sum_k a_{ik} b_{kj}$$

Note that the index k that you sum over is the second of A but the first of B. In short, you sum over "neighboring indices." Since you sum over all k, the result does not depend on k.

The zero matrix, usually called Z, is like the number zero; it does not change a matrix it is added to. And it turns whatever it is multiplied with into zero. A zero matrix has every coefficient zero. For example, in two dimensions:

$$Z = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

A unit, or identity, matrix, usually called I, is the equivalent of the number one for matrices; it does not change the vector or matrix it is multiplied with. A unit matrix is one on its "main diagonal" $i = j$ and zero elsewhere. The 2 by 2 unit matrix is:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

More generally the coefficients, $\{\delta_{ij}\}$, of a unit matrix are one if $i = j$ and zero otherwise.

The "transpose" of a matrix A, A^{T}, is what you get if you swap the two indices. Graphically, it turns its rows into its columns and vice versa. The "adjoint" or "Hermitian adjoint" matrix $A^\dagger$ is what you get if you both swap the two indices in a matrix A and then take the complex conjugate of every coefficient. If you want to take a matrix to the other side of an inner product, you will need to change it to its Hermitian adjoint. "Hermitian matrices"are equal to their Hermitian adjoint, so this does nothing for them.

The inverse of a matrix A, A^{-1} is a matrix so that $A^{-1}A$ equals the identity matrix I. That is much like the inverse of a simple number times that number gives one. And, just like the number zero has no inverse, a matrix with zero determinant has no inverse. Otherwise, you can swap the order;

AA^{-1} equals the unit matrix too. (For numbers this is trivial, for matrices you need to look a bit closer to understand why it is true.)

See also "determinant" and "eigenvector."

metric prefixes In the metric system, the prefixes Y, Z, E, P, T, G, M, and k stand for 10^i with $i = 24$, 21, 18, 15, 12, 9, 6, and 3, respectively. Similarly, d, c, m, μ, n, p, f, a, z, y stand for 10^{-i} with $i = 1$, 2, 3, 6, 9, 12, 15, 18, 21, and 24 respectively. For example, 1 ns is 10^{-9} seconds. English letter u is often used as instead of greek μ. Names corresponding to the mentioned prefixes Y–k are yotta, zetta, exa, peta, tera, giga, mega, kilo, and corresponding to d–y are deci, centi, milli, micro, nano, pico, femto, atto, zepto, and yocto.

molecular mass Typical thermodynamics books for engineers tabulate values of the "molecular mass," as a nondimensional number. The bottom line first: these numbers should have been called the "*molar* mass" of the substance, for the naturally occurring isotope ratio on earth. And they should have been given units of kg/kmol. That is how you use these numbers in actual computations. So just ignore the fact that what these books really tabulate is officially called the "*relative* molecular mass" for the natural isotope ratio.

Don't blame these textbooks too much for making a mess of things. Physicists have historically bandied about a zillion different names for what is essentially a single number. Like "molecular mass," "relative molecular mass," "molecular weight," "atomic mass," "relative atomic mass," "atomic weight," "molar mass," "relative molar mass," etcetera are basically all the same thing.

All of these have values that equal the mass of a molecule relative to a reference value for a single nucleon. So these value are about equal to the number of nucleons (protons and neutrons) in the nuclei of a single molecule. (For an isotope ratio, that becomes the average number of nucleons. Do note that nuclei are sufficiently relativistic that a proton or neutron can be noticeably heavier in one nucleus than another, and that neutrons are a bit heavier than protons even in isolation.) The official reference nucleon weight is defined based on the most common carbon isotope carbon-12. Since carbon-12 has 6 protons plus 6 neutrons, the reference nucleon weight is taken to be one twelfth of the carbon-12 atomic weight. That is called the unified atomic mass unit (u) or Dalton (Da). The atomic mass unit (amu) is an older virtually identical unit, but physicists and chemists could never quite agree on what its value was. No kidding.

If you want to be politically correct, the deal is as follows. "Molecular mass" is just what the term says, the mass of a molecule, in mass units.

(I found zero evidence in either the IUPAC Gold Book or NIST SP811 for the claim of Wikipedia that it must always be expressed in u.) "Molar mass" is just what the words says, the mass of a mole. Official SI units are kg/mol, but you will find it in g/mol, equivalent to kg/kmol. (You cannot expect enough brains from international committees to realize that if you define the kg and not the g as unit of mass, then it would be a smart idea to also define kmol instead of mol as unit of particle count.) Simply ignore relative atomic and molecular masses, you do not care about them. (I found zero evidence in either the IUPAC Gold Book or NIST SP811 for the claims of Wikipedia that the molecular mass cannot be an average over isotopes or that the molar mass must be for a natural isotope ratio. In fact, NIST uses "molar mass of carbon-12" and specifically includes the possibility of an average in the relative molecular mass.)

See also the atomic mass constant "m_u."

N May indicate:

- Number of states.
- Number of single-particle states.
- Number of neutrons in a nucleus.

N May indicate

- The atomic states or orbitals with theoretical Bohr energy E_4.
- Subscript N indicates a nucleus.

n May indicate:

- The principal quantum number for hydrogen atom energy eigenfunctions.
- A quantum number for harmonic oscillator energy eigenfunctions.
- Number of a single-electron or single-particle wave function.
- Generic summation index over energy eigenfunctions.
- Generic summation index over other eigenfunctions.
- Integer factor in Fourier wave numbers.
- Probability density.
- Number of columns in a matrix.
- A generic summation index or generic integer.
- A natural number.
- n_s is the number of spin states.

and maybe some other stuff.

n May indicate

- A subscript n may indicate a neutron.

natural Natural numbers are the numbers: $1, 2, 3, 4, \ldots$.

normal A normal operator or matrix is one that has orthonormal eigenfunctions or eigenvectors. Since eigenvectors are not orthonormal in general, a normal operator or matrix is abnormal! Another example of a highly confusing term. Such a matrix should have been called orthodiagonalizable or something of the kind. To be fair, the author is not aware of any physicists being involved in this particular term; it may be the mathematicians that are to blame here.

For an operator or matrix A to be "normal," it must commute with its Hermitian adjoint, $[A, A^\dagger] = 0$. Hermitian matrices are normal since they are equal to their Hermitian adjoint. Skew-Hermitian matrices are normal since they are equal to the negative of their Hermitian adjoint. Unitary matrices are normal because they are the inverse of their Hermitian adjoint.

O May indicate the origin of the coordinate system.

opposite The opposite of a number a is $-a$. In other words, it is the additive inverse.

$\boldsymbol{P}$ May indicate:

- The linear momentum eigenfunction.
- A power series solution.
- Probability.
- Pressure.
- Hermitian part of an annihilation operator.

$\boldsymbol{\mathcal{P}}$ Particle exchange operator. Exchanges the positions and spins of two identical particles.

$\boldsymbol{\mathfrak{P}}$ Peltier coefficient.

P Often used to indicate a state with one unit of orbital angular momentum.

$\boldsymbol{p}$ May indicate:

- Linear momentum.

- Linear momentum in the x-direction.
- Integration variable with units of linear momentum.

p May indicate

- An energy state with orbital azimuthal quantum number $l = 1$.
- A superscript p may indicate a single-particle quantity.
- A subscript p may indicate a periodic function.
- A subscript p may indicate a proton.

perpendicular bisector For two given points P and Q, the perpendicular bisector consists of all points R that are equally far from P as they are from Q. In two dimensions, the perpendicular bisector is the line that passes through the point exactly half way in between P and Q, and that is orthogonal to the line connecting P and Q. In three dimensions, the perpendicular bisector is the plane that passes through the point exactly half way in between P and Q, and that is orthogonal to the line connecting P and Q. In vector notation, the perpendicular bisector of points P and Q is all points R whose radius vector $\vec{r}$ satisfies the equation:

$$(\vec{r} - \vec{r}_P) \cdot (\vec{r}_Q - \vec{r}_P) = \tfrac{1}{2}(\vec{r}_Q - \vec{r}_P) \cdot (\vec{r}_Q - \vec{r}_P)$$

(Note that the halfway point $\vec{r} - \vec{r}_P = \frac{1}{2}(\vec{r}_Q - \vec{r}_P)$ is included in this formula, as is the half way point plus any vector that is normal to $(\vec{r}_Q - \vec{r}_P)$.)

phase angle Any complex number can be written in "polar form" as $c = |c|e^{\mathrm{i}\alpha}$ where both the magnitude $|c|$ and the phase angle α are real numbers. Note that when the phase angle varies from zero to 2π, the complex number c varies from positive real to positive imaginary to negative real to negative imaginary and back to positive real. When the complex number is plotted in the complex plane, the phase angle is the direction of the number relative to the origin. The phase angle α is often called the argument, but so is about everything else in mathematics, so that is not very helpful.

In complex time-dependent waves of the form $e^{\mathrm{i}(\omega t - \phi)}$, and its real equivalent $\cos(\omega t - \phi)$, the phase angle ϕ gives the angular argument of the wave at time zero.

photon Unit of electromagnetic radiation (which includes light, x-rays, microwaves, etcetera). A photon has a energy $\hbar\omega$, where ω is its angular frequency, and a wave length $2\pi c/\omega$ where c is the speed of light.

potential In order to optimize confusion, pretty much everything in physics that is scalar is called potential. Potential energy is routinely concisely referred to as potential. It is the energy that a particle can pick up from a force field by changing its position. It is in Joule. But an electric potential is taken to be per unit charge, which gives it units of volts. Then there are thermodynamic potentials like the chemical potential.

$\boldsymbol{p_x}$ Linear momentum in the x-direction. (In the one-dimensional cases at the end of the unsteady evolution chapter, the x subscript is omitted.) Components in the y- and z-directions are p_y and p_z. Classical Newtonian physics has $p_x = mu$ where m is the mass and u the velocity in the x-direction. In quantum mechanics, the possible values of p_x are the eigenvalues of the operator $\widehat{p}_x$ which equals $\hbar\partial/\mathrm{i}\partial x$. (But which becomes canonical momentum in a magnetic field.)

$\boldsymbol{Q}$ May indicate

- Number of energy eigenfunctions of a system of particles.
- Anti-Hermitian part of an annihilation operator divided by i.
- Heat flow or heat.
- Charge.
- Electric quadrupole moment.
- Energy release.

$\boldsymbol{q}$ May indicate:

- Charge.
- Heat flux density.
- The number of an energy eigenfunction of a system of particles.
- Generic index.

$\boldsymbol{R}$ May indicate:

- Ideal gas constant.
- Transition rate.
- Nuclear radius.
- Reflection coefficient.
- Some radius or typical radius (like in the Yukawa potential).
- Some function of r to be determined.
- Some function of (x, y, z) to be determined.

- R_{nl} is a hydrogen radial wave function.
- $R_{\mathrm{u}} = 8.314\,462$ kJ/kmol K is the universal gas constant. It is the equivalent of Boltzmann's constant, but for a kmol instead of a single atom or molecule.

$\boldsymbol{\mathcal{R}}$ Rotation operator.

$\boldsymbol{\Re}$ The real part of a complex number. If $c = c_r + \mathrm{i}c_i$ with c_r and c_i real numbers, then $\Re(c) = c_r$. Note that $c + c^* = 2\Re(c)$.

$\boldsymbol{r}$ May indicate:

- The radial distance from the chosen origin of the coordinate system.
- r_i typically indicates the i-th Cartesian component of the radius vector $\vec{r}$.
- Some ratio.

$\vec{\boldsymbol{r}}$ The position vector. In Cartesian coordinates (x, y, z) or $x\hat{\imath} + y\hat{\jmath} + z\hat{k}$. In spherical coordinates $r\hat{\imath}_r$. Its three Cartesian components may be indicated by r_1, r_2, r_3 or by x, y, z or by x_1, x_2, x_3.

reciprocal The reciprocal of a number a is $1/a$. In other words, it is the multiplicative inverse.

relativity The special theory of relativity accounts for the experimental observation that the speed of light c is the same in all local coordinate systems. It necessarily drops the basic concepts of absolute time and length that were corner stones in Newtonian physics.

Albert Einstein should be credited with the boldness to squarely face up to the unavoidable where others wavered. However, he should also be credited for the boldness of swiping the basic ideas from Lorentz and Poincaré without giving them proper, or any, credit. The evidence is very strong he was aware of both works, and his various arguments are almost carbon copies of those of Poincaré, but in his paper it looks like it all came from Einstein, with the existence of the earlier works not mentioned. (Note that the general theory of relativity, which is of no interest to this book, is almost surely properly credited to Einstein. But he was a lot less hungry then.)

Relativity implies that a length seen by an observer moving at a speed v is shorter than the one seen by a stationary observer by a factor $\sqrt{1 - (v/c)^2}$ assuming the length is in the direction of motion. This is called Lorentz-Fitzgerald contraction. It makes galactic travel somewhat more conceivable because the size of the galaxy will contract for an astronaut in a

rocket ship moving close to the speed of light. Relativity also implies that the time that an event takes seems to be slower by a factor $1/\sqrt{1-(v/c)^2}$ if the event is seen by an observer in motion compared to the location where the event occurs. That is called time dilation. Some high-energy particles generated in space move so fast that they reach the surface of the earth though this takes much more time than the particles would last at rest in a laboratory. The decay time increases because of the motion of the particles. (Of course, as far as the particles themselves see it, the distance to travel is a lot shorter than it seems to be to earth. For them, it is a matter of length contraction.)

The following formulae give the relativistic mass, momentum, and kinetic energy of a particle in motion:

$$m = \frac{m_0}{\sqrt{1-(v/c)^2}} \qquad p = mv \qquad T = mc^2 - m_0c^2$$

where m_0 is the rest mass of the particle, i.e. the mass as measured by an observer to whom the particle seems at rest. The formula for kinetic energy reflects the fact that even if a particle is at rest, it still has an amount of "build-in" energy equal to m_0c^2 left. The total energy of a particle in empty space, being kinetic and rest mass energy, is given by

$$E = mc^2 = \sqrt{(m_0c^2)^2 + c^2p^2}$$

as can be verified by substituting in the expression for the momentum, in terms of the rest mass, and then taking both terms inside the square root under a common denominator. For small linear momentum p, this can be approximated as $\frac{1}{2}m_0v^2$.

Relativity seemed quite a dramatic departure of Newtonian physics when it developed. Then quantum mechanics started to emerge...

rot The rot of a vector $\vec{v}$ is defined as $\operatorname{curl}\vec{v} \equiv \operatorname{rot}\vec{v} \equiv \nabla \times \vec{v}$.

S May indicate:

- Number of states per unit volume.
- Number of states at a given energy level.
- Spin angular momentum (as an alternative to using L or J for generic angular momentum.)
- Entropy.
- S_{12} is a factor in the so-called tensor potential of nucleons.

S The action integral of Lagrangian mechanics, {A.1}

$\mathcal{S}$ Seebeck coefficient.

S Often used to indicate a state of zero orbital angular momentum.

s May indicate:

- Spin value of a particle. Equals $\frac{1}{2}$ for electrons, protons, and neutrons, is also half an odd natural number for other fermions, and is a nonnegative integer for bosons. It is the azimuthal quantum number l due to spin.
- Specific entropy.
- As an index, shelf number.

s May indicate:

- An energy state with orbital azimuthal quantum number $l = 0$. Such a state is spherically symmetric.

scalar A quantity that is not a vector, a quantity that is just a single number.

sin The sine function, a periodic function oscillating between 1 and -1 as shown in [41, pp. 40-]. Good to remember: $\cos^2\alpha + \sin^2\alpha = 1$ and $\sin 2\alpha = 2\sin\alpha\cos\alpha$ and $\cos 2\alpha = \cos^2\alpha - \sin^2\alpha$.

solenoidal A vector $\vec{v}$ is solenoidal if its divergence $\nabla \cdot \vec{v}$ is zero.

spectrum In this book, a spectrum normally means a plot of energy levels along the vertical axis. Often, the horizontal coordinate is used to indicate a second variable, such as the density of states or the particle velocity.

For light (photons), a spectrum can be obtained experimentally by sending the light through a prism. This separates the colors in the light, and each color means a particular energy of the photons.

The word spectrum is also often used in a more general mathematical sense, but not in this book as far as I can remember.

spherical coordinates The spherical coordinates r, θ, and ϕ of an arbitrary point P are defined as

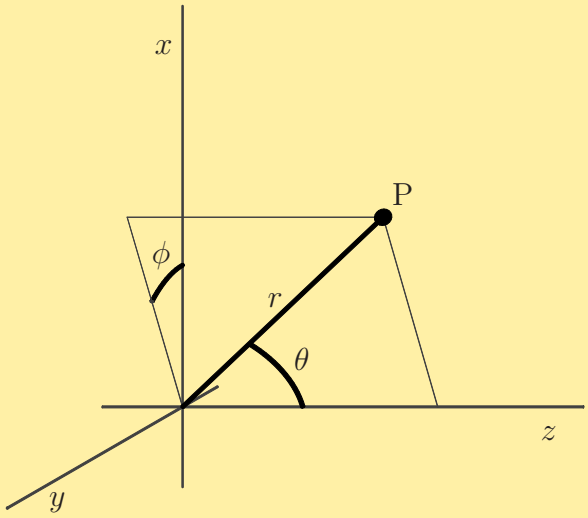


Figure N.3: Spherical coordinates of an arbitrary point P.

In Cartesian coordinates, the unit vectors in the x, y, and z directions are called $\hat{\imath}$, $\hat{\jmath}$, and $\hat{k}$. Similarly, in spherical coordinates, the unit vectors in the r, θ, and ϕ directions are called $\hat{\imath}_r$, $\hat{\imath}_\theta$, and $\hat{\imath}_\phi$. Here, say, the θ direction is defined as the direction of the change in position if you increase θ by an infinitesimally small amount while keeping r and φ the same. Note therefore in particular that the direction of $\hat{\imath}_r$ is the same as that of $\vec{r}$; radially outward.

An arbitrary vector $\vec{v}$ can be decomposed in components v_r, v_θ, and v_ϕ along these unit vectors. In particular

$$\vec{v} \equiv v_r \hat{\imath}_r + v_\theta \hat{\imath}_\theta + v_\phi \hat{\imath}_\phi$$

Recall from calculus that in spherical coordinates, a volume integral of an arbitrary function f takes the form

$$\int f \, \mathrm{d}^3\vec{r} = \int\int\int f r^2 \sin\theta \, \mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi$$

In other words, the volume element in spherical coordinates is

$$\mathrm{d}V = \mathrm{d}^3\vec{r} = r^2 \sin\theta \, \mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi$$

Often it is convenient of think of volume integrations as a two-step process: first perform an integration over the angular coordinates θ and ϕ. Physically, that integrates over spherical surfaces. Then perform an integration over r to integrate all the spherical surfaces together. The combined infinitesimal angular integration element

$$\mathrm{d}\Omega = \sin\theta \mathrm{d}\theta\mathrm{d}\phi$$

is called the infinitesimal "solid angle" $\mathrm{d}\Omega$. In two-dimensional polar coordinates r and θ, the equivalent would be the infinitesimal polar angle $\mathrm{d}\theta$. Recall that $\mathrm{d}\theta$, (in proper radians of course), equals the arclength of an infinitesimal part of the circle of integration divided by the circle radius. Similarly $\mathrm{d}\Omega$ is the surface of an infinitesimal part of the sphere of integration divided by the square sphere radius.

See the "∇" entry for the gradient operator and Laplacian in spherical coordinates.

Stokes' Theorem This theorem, first derived by Kelvin and first published by someone else I cannot recall, says that for any reasonably smoothly varying vector $\vec{v}$,

$$\int_A (\nabla \times \vec{v})\, \mathrm{d}A = \oint \vec{v} \cdot \mathrm{d}\vec{r}$$

where the first integral is over any smooth surface area A and the second integral is over the edge of that surface. How did Stokes get his name on it? He tortured his students with it, that's how!

One important consequence of the Stokes theorem is for vector fields $\vec{v}$ that are "irrotational," i.e. that have $\nabla \times \vec{v} = 0$. Such fields can be written as

$$\vec{v} = \nabla f \qquad f(\vec{r}) \equiv \int_{\underline{\vec{r}}=\vec{r}_{\rm ref}}^{\vec{r}} \vec{v}(\underline{\vec{r}}) \cdot \mathrm{d}\underline{\vec{r}}$$

Here $\vec{r}_{\rm ref}$ is the position of an arbitrarily chosen reference point, usually the origin. The reason the field $\vec{v}$ can be written this way is the Stokes theorem. Because of the theorem, it does not make a difference along which path from $\vec{r}_{\rm ref}$ to $\vec{r}$ you integrate. (Any two paths give the same answer, as long as $\vec{v}$ is irrotational everywhere in between the paths.) So the definition of f is unambiguous. And you can verify that the partial derivatives of f give the components of $\vec{v}$ by approaching the final position $\vec{r}$ in the integration from the corresponding direction.

symmetry A symmetry is an operation under which an object does not change. For example, a human face is almost, but not completely, mirror symmetric: it looks almost the same in a mirror as when seen directly. The electrical field of a single point charge is spherically symmetric; it looks the same from whatever angle you look at it, just like a sphere does. A simple smooth glass (like a glass of water) is cylindrically symmetric; it looks the same whatever way you rotate it around its vertical axis.

T May indicate:

- Absolute temperature. The absolute temperature in degrees K equals the temperature in centigrade plus 273.15. When the absolute tem-

perature is zero, (i.e. at -273.15 °C), nature is in the state of lowest possible energy.

- Kinetic energy. A hat indicates the associated operator. The operator is given by the Laplacian times $-\hbar^2/2m$.
- Isospin. A hat indicates the associated operator. A vector symbol or subscript distinguishes it from kinetic energy.
- Tesla. The unit of magnetic field strength, kg/C-s.

$\boldsymbol{\mathcal{T}}$ Translation operator that translates a wave function through space. The amount of translation is usually indicated by a subscript.

$\boldsymbol{t}$ May indicate:

- Time.
- t_t is the quantum number of square isospin.

temperature A measure of the heat motion of the particles making up macroscopic objects. At absolute zero temperature, the particles are in the "ground state" of lowest possible energy.

triple product A product of three vectors. There are two different versions:

- The scalar triple product $\vec{a} \cdot (\vec{b} \times \vec{c})$. In index notation,

$$\vec{a} \cdot (\vec{b} \times \vec{c}) = \sum_i a_i (b_{\bar{\imath}} c_{\bar{\bar{\imath}}} - b_{\bar{\bar{\imath}}} c_{\bar{\imath}})$$

  where $\bar{\imath}$ is the index following i in the sequence 123123..., and $\bar{\bar{\imath}}$ the one preceding it. This triple product equals the determinant $|\vec{a}\vec{b}\vec{c}|$ formed with the three vectors. Geometrically, it is plus or minus the volume of the parallelepiped that has vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ as edges. Either way, as long as the vectors are normal vectors and not operators,

$$\vec{a} \cdot (\vec{b} \times \vec{c}) = \vec{b} \cdot (\vec{c} \times \vec{a}) = \vec{c} \cdot (\vec{a} \times \vec{b})$$

  and you can change the two sides of the dot product without changing the triple product, and/or you can change the sides in the vectorial product with a change of sign. If any of the vectors is an operator, use the index notation expression to work it out.

- The vectorial triple product $\vec{a} \times (\vec{b} \times \vec{c})$. In index notation, component number i of this triple product is

$$a_{\bar{\imath}} (b_i c_{\bar{\imath}} - b_{\bar{\imath}} c_i) - a_{\bar{\bar{\imath}}} (b_{\bar{\bar{\imath}}} c_i - b_i c_{\bar{\bar{\imath}}})$$

which may be rewritten as

$$a_i b_i c_i + a_{\bar{i}} b_i c_{\bar{i}} + a_{\bar{i}} b_i c_{\bar{i}} - a_i b_i c_i - a_{\bar{i}} b_{\bar{i}} c_i - a_{\bar{i}} b_{\bar{i}} c_i$$

In particular, as long as the vectors are normal ones,

$$\vec{a} \times (\vec{b} \times \vec{c}) = (\vec{a} \cdot \vec{c})\vec{b} - (\vec{a} \cdot \vec{b})\vec{c}$$

U May indicate:

- A unitary operator, in other words one that does not change the magnitude of the wave function.
- Often used for energy, though not in this book.

$\mathcal{U}$ The time shift operator: $\mathcal{U}(\tau, t)$ changes the wave function $\Psi(\ldots; t)$ into $\Psi(\ldots; t + \tau)$. If the Hamiltonian is independent of time

$$\mathcal{U}(\tau, t) = \mathcal{U}_\tau = e^{-\mathrm{i}H\tau/\hbar}$$

u May indicate:

- The first velocity component in a Cartesian coordinate system.
- A complex coordinate in the derivation of spherical harmonics.
- An integration variable.

u May indicate the atomic mass constant, equivalent to $1.660\,538\,92\ 10^{-27}$ kg or $931.494\,06$ MeV/c^2.

V May indicate:

- The potential energy. V is used interchangeably for the numerical values of the potential energy and for the operator that corresponds to multiplying by V. In other words, $\widehat{V}$ is simply written as V.

$\mathcal{V}$ Volume.

v May indicate:

- The second velocity component in a Cartesian coordinate system.
- Magnitude of a velocity (speed).
- v is specific volume.
- A complex coordinate in the derivation of spherical harmonics.
- As v^{ee}, a single electron pair potential.

$\vec{v}$ May indicate:

- Velocity vector.
- Generic vector.
- Summation index of a lattice potential.

vector Simply put, a list of numbers. A vector $\vec{v}$ in index notation is a set of numbers $\{v_i\}$ indexed by an index i. In normal three-dimensional Cartesian space, i takes the values 1, 2, and 3, making the vector a list of three numbers, v_1, v_2, and v_3. These numbers are called the three components of $\vec{v}$. The list of numbers can be visualized as a column, and is then called a ket vector, or as a row, in which case it is called a bra vector. This convention indicates how multiplication should be conducted with them. A bra times a ket produces a single number, the dot product or inner product of the vectors:

$$(1,3,5)\begin{pmatrix}7\\11\\13\end{pmatrix} = 1\ 7 + 3\ 11 + 5\ 13 = 105$$

To turn a ket into a bra for purposes of taking inner products, write the complex conjugates of its components as a row.

Formal definitions of vectors vary, but real mathematicians will tell you that vectors are objects that can be manipulated in certain ways (addition and multiplication by a scalar). Some physicists define vectors as objects that transform in a certain way under coordinate transformation (one-dimensional tensors); that is *not* the same thing.

vectorial product An vectorial product, or cross product is a product of vectors that produces another vector. If

$$\vec{c} = \vec{a} \times \vec{b},$$

it means in index notation that the i-th component of vector $\vec{c}$ is

$$c_i = a_{\bar{i}} b_{\bar{\bar{i}}} - a_{\bar{\bar{i}}} b_{\bar{i}}$$

where $\bar{i}$ is the index following i in the sequence 123123..., and $\bar{\bar{i}}$ the one preceding it. For example, c_1 will equal $a_2 b_3 - a_3 b_2$.

W May indicate:

- Watt, the SI unit of power.
- The $W^{\pm}$ are the charged carriers of the weak force. See also "Z^0."

- W.u. stands for Weisskopf unit, a simple decay ballpark for gamma decay.

w May indicate:

- The third velocity component in a Cartesian coordinate system.
- Weight factor.

$\vec{\boldsymbol{w}}$ Generic vector.

X Used in this book to indicate a function of x to be determined.

x May indicate:

- First coordinate in a Cartesian coordinate system.
- A generic argument of a function.
- An unknown value.

Y Used in this book to indicate a function of y to be determined.

$\boldsymbol{Y_l^m}$ Spherical harmonic. Eigenfunction of both angular momentum in the z-direction and of total square angular momentum.

y May indicate:

- Second coordinate in a Cartesian coordinate system.
- A second generic argument of a function.
- A second unknown value.

Z May indicate:

- Atomic number (number of protons in the nucleus).
- Number of particles.
- Partition function.
- The Z^0 is the uncharged carrier of the weak force. See also "$W^{\pm}$."
- Used in this book to indicate a function of z to be determined.

z May indicate:

- Third coordinate in a Cartesian coordinate system.
- A third generic argument of a function.
- A third unknown value.

Index

$\cdot$, 1493
$\times$, 1493
!, 1494
|, 1494
$|\ldots\rangle$, 1495
$\langle\ldots|$, 1495
$\uparrow$, 1495
$\downarrow$, 1495
Σ, *37*
$\prod$, 1495
$\sum$, 1495
$\int$, 1496
$\rightarrow$, 1496
$\vec{}$, 1496
$\hat{}$, 1496
$'$, 1496
∇, 1497
$\Box$, 1498
*, 1498
$<$, 1499
$\leqslant$, 1499
$\langle\ldots\rangle$, 1499
$>$, 1499
$\geqslant$, 1499
$[\ldots]$, 1499
$=$, 1499
$\equiv$, 1499
$\approx$, 1499
$\sim$, 1499
$\propto$, 1499
α, 1499
β, 1500
Γ, 1500
γ, 1500
Δ, 1500
δ, 1500
∂, 1501
ϵ, 1501
ϵ_0, 1501
ε, 1501
η, 1501
Θ, 1501
θ, 1502
ϑ, 1502
κ, 1502
Λ, 1502
λ, 1502
μ, 1502
ν, 1503
ξ, 1503
Π, 1503
π, 1503
$\tilde{\pi}$, 1503
ρ, 1503
σ, 1504
τ, 1504
Φ, 1504
ϕ, 1504
φ, 1504
χ, 1505
Ψ, 1505
ψ, 1505
Ω, 1505
ω, 1505

21 cm line
  derivation, *1148*
  intro, *1140*

A, 1505
Å, 1506
a, 1506
a_0, 1506
absolute temperature, *212*, 1506
absolute value, 32, 1506
absolute zero
  nonzero energy, *84*
  requires ground state, *520*
absorbed dose, *671*
absorption and emission
  incoherent radiation, *379*
absorptivity, *227*
acceleration
  in quantum mechanics, *326*
acceptors
  semiconductors, *285*

actinides, 191
actinoids, 191
action, *860*
  relativistic, *27*
activation energy
  nuclear fission, *751*
  radicals, 150
active view, *948*
activity, *671*
  specific, *see* decay rate
adiabatic
  disambiguation, 1506
  quantum mechanics, *322*
  thermodynamics, *551*
adiabatic surfaces, *445*
adiabatic theorem
  derivation, *1283*
  derivation and implications, *941*
  intro, *323*
adjoint, 1507
  matrices, 1523
Aharonov-Bohm effect, 607
Airy functions
  application, *1085*
  connection formulae, *1097*
  graphs, *1096*
  software, *1091*
alkali metals, 188
alkaline metals, 188
allowed transition
  intro, *340*
allowed transitions
  beta decay, *817*
alpha, *see* α
alpha decay, *663*
  data, *690*
  definition, *663*
  Gamow/Gurney and Condon theory, *691*
  overview of data, *656*
  Q-value, *813*
  quantum mechanical tunneling, *688*
alpha particle, *688*
ammonia molecule, *151*
amplitude, 1507
  quantum, *52*
angle, 1507
angular frequency, *392*
angular momentum, *92*
  addition, *590*
    Clebsch-Gordan coefficients, *590*
  advanced treatment, *579*
  combination
    intro, *336*
  component, *93*
    eigenfunctions, *94*
    eigenvalues, *95*
  conservation in decays, *336*
  definition, *93*
  fundamental commutation relations
    as an axiom, *580*
    intro, *127*
  ladder operators, *581*
  ladders, *583*
  normalization factors, *586*
  nuclei
    data, *755*
  operator
    Cartesian, *93*
  possible values, *584*
  spin, *155*
  square angular momentum, *96*
    eigenfunctions, *96*
    eigenvalues, *98*
  symmetry and conservation, *327*
  uncertainty, *99*
anions, *265*
anomalous magnetic moment, *634*
  nucleons
    pion explanation, *682*
anti-bonding, 493
antibonding state
  intro, *154*
anticommutator, *919*
antilinear operator, *951*
antiparticles
  move backward in time, *908*
antisymmetrization requirement, *166*
  graphical depiction, *525*
  indistinguishable particles, *525*
  number of terms, *173*
  using groupings, *169*
  using occupation numbers, *913*
  using Slater determinants, *170*
antiunitary operator, *951*
astronomy
  spectral analysis, 299
asymptotic freedom
  quarks, 359
atomic mass
  conversion to nuclear mass, *673*
  versus nuclear mass, *814*
atomic mass unit, *673*

atomic matrix element, *376*
atomic number, *178*, *656*
atoms
  eigenfunctions, *178*
  eigenvalues, *178*
  ground state, *181*
  Hamiltonian, *178*
Auger effect
  Meisner, *1481*
Auger electrons, 853
Avalanche diode, *296*
average
  versus expectation value, 115
Avogadro's number, 1517
axial vector, *969*
azimuthal quantum number, *98*

B, 1508
$\mathcal{B}$, 1508
b, 1508
Balmer transitions, *107*
band gap
  and Bragg reflection, *1452*
  intro, *258*
band structure
  crossing bands, *492*
  detailed germanium structure, *277*
  nearly-free electrons, *501*
  widely spaced atoms, *478*
band theory
  electrons per primitive cell, *261*
  intro, *256*
barn, *774*
baryon, 156
baryons, 356
basis, 1508
  crystal
    intro, *279*
  diamond, *494*
  lithium (BCC), *477*
  NaCl (FCC), *474*
  spin states, *164*
  vectors or functions, *44*
  zinc blende (ZnS), *279*
battery, 247
BCC
  lithium, *477*
becquerel, *671*
Bell's theorem, *411*
  cheat, *415*
benzene molecular ring, *150*
Berry's phase, *942*
beryllium-11
  nuclear spin, *730*
Bessel functions
  spherical, *879*
beta, *see* β
beta decay, *801*
  beta-plus decay
    definition, *663*
  double
    explanation, *812*
  electron capture, *662*
  electron emission, *662*
  energetics
    data, *803*
  energy release data, *803*
  Fermi theory, *1191*
  forbidden decays, *816*
  intro, *662*
  inverse beta decay, *662*
  K or L capture, *662*
  lone neutron, *651*
  momentum conservation, *1208*
  nuclei that do, *662*
  overview of data, *656*
  positron emission, *662*
  Q-value, *813*
  superallowed decays, *825*
  von Weizsaecker predictions, *810*
beta vibration
  nuclei, 745
Bethe-von Weizsäcker formula, *687*
Big Bang, 963
binding energy
  definition, *136*
  hydrogen molecular ion, *136*
  hydrogen molecule, *149*
  lithium hydride, *153*
Biot-Savart law, *629*
  derivation, *1397*
blackbody radiation, *569*
  intro, *225*
blackbody spectrum, *226*
  extended derivation, *569*
Bloch function
  nearly-free electrons, *501*
  one-dimensional lattice, *483*
  three-dimensional lattice, *491*
Bloch wave
  explanation, *396*
  intro, *268*
Bloch's theorem, *483*

body-centered cubic, *see* BCC
Bohm
  EPR experiment, 411
Bohr energies, *106*
  relativistic corrections, *1138*
Bohr magneton, *633*
Bohr radius, *110*
Boltzmann constant, 1519
Boltzmann factor, *533*
bond
  covalent, *193*
  hydrogen, *194*
  ionic, *199*
  pi, *194*
  polar, *194*
  sigma, *193*
  Van der Waals, *469*
bond length
  definition, *135*
  hydrogen molecular ion, *137*
  hydrogen molecule, *149*
Born
  approximation, *1107*
Born series, *1108*
Born statistical interpretation, *51*
Born-Oppenheimer approximation
  and adiabatic theorem, *323*
  basic idea, *441*
  derivation, *439*
  diagonal correction, *1342*
  hydrogen molecular ion, *129*
  hydrogen molecule, *142*
  include nuclear motion, *443*
  spin degeneracy, *1337*
  vibronic coupling terms, *1340*
Borromean nucleus, *730*
Bose-Einstein condensation
  derivation, *565*
  intro, *214*
  rough explanation, *218*
  superfluidity, *1469*
Bose-Einstein distribution
  blackbody radiation, *569*
    intro, *225*
  canonical probability, *532*
  for given energy, *531*
  identify chemical potential, *562*
  intro, *223*
bosons, *156*
  ground state, *211*
  symmetrization requirement, *166*
bound states
  hydrogen
    energies, *106*
boundary conditions
  acceptable singularity, *1444*
    hydrogen atom, *1246*
  across delta function potential, *1089*
  at infinity
    harmonic oscillator, *1236*
    hydrogen atom, *1245*
  impenetrable wall, *62*
  radiation, *1084*
    accelerating potential, *1085*
    three-dimensional, *1100*
  unbounded potential, *1087*
Bq, *671*
bra, *38*, 1495, 1522
Bragg diffraction
  electrons, 518
Bragg planes
  Brillouin fragment boundaries, *492*
  energy singularities, *507*
  one-dimensional (Bragg points), *485*
  X-ray diffraction, *518*
Bragg reflection
  and band gaps, *1452*
Bragg's law, *512*
Breit-Wigner distribution, 371
Brillouin zone
  first
    FCC crystal, *277*
  intro, *273*
  one-dimensional, *484*
  three-dimensional, *491*
broadband radiation
  intro, *299*
built-in potential, *290*

C, 1508
°C, 1508
c, 1508
canonical commutation relation, *125*
canonical Hartree-Fock equations, *458*
canonical momentum
  canonical quantization, *926*
  intro, *904*
  special relativity, *27*
  with a magnetic field, *606*
canonical probability distribution, *532*
canonical quantization, *926*
  canonical momentum, *926*
carbon nanotubes

electrical properties
intro, 264
intro, 197
Carnot cycle, *543*
Cartesian tensors, 874
Casimir force, *1041*
cat, Schrödinger's, *410*
cations, *265*
Cauchy-Schwartz inequality, 1509
causality
relativity, *15*
special relativity, *16*
causality problem, 931
centrifugal stretching, *742*
chain reaction, *751*
charge
electrostatics, *616*
charge annihilation operator, *793*
charge conjugation
intro, *331*
Wu experiment, *828*
charge creation operator, *792*
charge independence
nuclear force, *646*
charge states, *790*
charge symmetry
example, 714
nuclear force, *647*
charge transfer insulators, 262
chemical bonds, *193*
covalent pi bonds, *194*
covalent sigma bonds, *193*
hybridization, *196*
ionic bonds, *199*
polar covalent bonds, *194*
promotion, *196*
sp^n hybridization, *196*
chemical equilibrium
constant pressure, 561
constant volume, 561
chemical potential, *559*
and diffusion, *246*
intro, *243*
and distributions, *562*
line up
Peltier cooler, *306*
microscopic, *562*
chi, *see* χ
Ci, *671*
circular polarization
from second quantization, *1042*
intro, *344*
photon wave function, *977*
classical, 1509
Clausius-Clapeyron equation, *560*
Clebsch-Gordan coefficients, *590*
and Wigner 3j symbols, *1456*
computing using recursion, *1378*
explicit expression, *1378*
coefficient of performance, *545*
coefficients of eigenfunctions
evaluating, *91*
give probabilities, *59*
time variation, *318*
collapse of the wave function, *57*
collision-dominated regime, *362*
collisionless regime, *362*
collisions
dual nature, *364*
color force, 647
intro, 356
commutation relation
canonical, *125*
commutator, *122*
definition, *124*
commutator eigenvalue problems, 582
commuting operators, *122*
common eigenfunctions, *122*
comparative half-life, *821*
complete set, *43*
completeness relation, *47*
complex conjugate, *32*
complex numbers, *31*
component waves, *389*
components of a vector, *34*
conduction band
intro, *258*
conductivity
effect of light, *301*
electrical, *255*
ionic, *265*
configuration mixing, *722*
confinement, *234*
quarks, 358
single particle, *74*
conjugate momentum, *see* canonical momentum
conjugate nuclei, *793*
connection formulae, *1096, 1097*
conservation laws
and symmetries, *327*
conserved vector current hypothesis, *1202*

contact potential, *247*
continuity equation
incompressible flow, *1178*
contravariant, 20
conventional cell, *489*
conversion electron, *850*
Copenhagen Interpretation, *57*
correlation energy, *465*
cos, 1509
Coulomb barrier, *692*
Coulomb condition
unconventional derivation, *1027*
Coulomb gage
instead of Lorenz gage, *1026*
Coulomb gauge, *973*
classical electromagnetics, *1125*
Coulomb integrals, *456*
Coulomb potential, *100*
Fermi derivation, *1024*
Koulomb potential
field theory derivation, *981*
Coulomb potential energy
derivation, *1299*
coupling constant, *1195*
covalent bond
hydrogen molecular ion, *129*
covalent solids, *492*
covariant, 21
creationists, *1475*
cross product, 1536
crystal
basis
diamond, *494*
NaCl (FCC), *474*
ionic conductivity, *265*
lattice, *see* lattice
lithium (BCC), *477*
one-dimensional
primitive translation vector, *483*
transparency, *300*
typical semiconductors, *277*
crystal momentum, 269, *272*
conservation, *273*
definition, *396*
light-emitting diodes, *303*
crystals
translation operator, *396*
curie, *671*
curl, 1497, 1509
cylindrical coordinates, 1498

D, 1509
$\vec{D}$, 1510
$\mathcal{D}$, 1510
D, 1510
d, 1510
$\vec{d}$, 1510
d, 1510
D'Alembertian, 907, 1498
Dalton, *673*
Darwin term, *1142*
d block
periodic table, 191
de Broglie relation, *249*
derivation, 906
Debye model, *571*
Debye temperature, *572*, 573
decay constant, *see* decay rate, *671*
decay rate, *671*
not a probability, *361*
physical mechanism, *362*
specific, *360*
deformed nuclei, *738*
degeneracy, *88*
degeneracy pressure, *232*
degenerate semiconductor, *287*
delayed neutrons, *752*
Delta, *see* Δ
delta function, *383*
three-dimensional, *383*
Delta particles
intro, *1175*
delta, *see* δ
density
mass, *537*
molar, *538*
particle, *537*
density of modes, *210*
density of states, *208*
confined, *234*
periodic box, *251*
depletion layer, *291*
derivative, 1510
determinant, 1510
deuterium, *652*
deuteron
intro, *652*
OPEP potential, *1167*
diamagnetic contribution, *635*
diamond
band gap, 492
crystal structure, *277*
intro, 197

differential cross-section, 1102
dimensional analysis, 891
dineutron
  isospin, *789*
  not bound, *653*
  OPEP potential, *1167*
diode
  semiconductor, *290*
diode laser, *303*
dipole
  classical electromagnetics, *621*
dipole moment
  electric
    nuclei, *771*
  magnetic
    classical, *633*
    nuclei, *771*
dipole strength
  molecules, *471*
dipole transition, *833*
  electric
    intro, *340*
  magnetic
    intro, *341*
dipole transitions
  magnetic
    Hamiltonian, *1304*
diproton
  isospin, *789*
  not bound, *653*
  OPEP potential, *1167*
Dirac delta function, *383*
  three-dimensional, 986
Dirac equation, *602*
  as a system, *1186*
  conserves parity, *1188*
  hydrogen atom
    low speed approximation, *1410*
  nonrelativistic limit
    no linear algebra, *1186*
  ultrarelativistic, *1187*
Dirac gamma matrices, *1122*
Dirac notation, *46*
direct gap semiconductor, *275*
discrete spectrum
  versus broadband
    intro, *299*
disintegration constant, *see* decay rate, *671*
disintegration rate, *671*
dispersion relation, *392*
distinguishable particles
  intro, *216*, *222*
div, 1497
div(ergence), 1512
divergence, 1497
divergence theorem, 1515
donors
  semiconductors, *285*
doping
  semiconductors, *282*
Doppler shift
  of light in vacuum, *10*
dose equivalent, *672*
dot product, *36*
double layer of charges
  contact surfaces, *246*
doublet states, *165*
dpm, *671*
Dulong and Petit law, *573*
dynamic phase, *942*

E, 1512
$\mathcal{E}$, 1512
e, 1512
e, 1513
effective dose, *672*
effective mass
  from equation of motion, *398*
  one-dimensional example, *271*
Ehrenfest theorem, *326*
e^{iax}, 1513
eigenfunction, *41*
eigenfunctions
  angular momentum component, *94*
  atoms, *178*
  harmonic oscillator, *85*
  hydrogen atom, *109*
  impenetrable spherical shell, *1402*
  linear momentum, *385*
  position, *382*
  square angular momentum, *96*
eigenvalue, *41*
eigenvalue problems
  commutator type, 582
  ladder operators, 582
eigenvalues
  angular momentum component, *95*
  atoms, *178*
  harmonic oscillator, *83*
  hydrogen atom, *106*
  impenetrable spherical shell, *1402*
  linear momentum, *385*
  position, *382*

square angular momentum, *98*
eigenvector, *41*, 1513
Einstein
dice, 59
summation convention, *19*
swiped special relativity, *3*
Einstein A and B coefficients, *380*
Einstein's derivation, *1314*
Einstein A coefficients
quantum derivation, *1044*
Einstein B coefficients
quantum derivation, *1310*
Einstein Podolski Rosen, *412*
Einstein summation convention
moral justification, *1010*
electric charge
electron and proton, 101
electric dipole approximation
origin of the name, *377*
electric dipole operator
intro, *377*
electric dipole transition
intro, *340*
selection rules, *347*
relativistic, *348*
electric moment
nuclei, *771*
electric multipole
photon states, *980*
electric potential
classical derivation, *608*, *1389*
quantum derivation, *605*
relativistic derivation, *27*
electrical conduction
intro, *252*
electrochemical potential
definition, *240*
electromagnetic field
Hamiltonian, *605*
Maxwell's equations, *607*
quantization, 1032
electromagnetic potentials
gauge transformation, *28*
electromagnetics
"derivation" from scratch, *981*
electron
in magnetic field, *632*
electron affinity, *472*
Hartree-Fock, 462
electron capture
definition, *662*
electron emission, *662*
electron split experiment, 52
electronegativity, *186*, 472
electrons
lack of intelligence, *233*, *296*
emission rate
spontaneous, *see* decay rate
emissivity, *227*
energy conservation, *319*
energy spectrum
harmonic oscillator, *83*
hydrogen atom, *106*
energy-time uncertainty equality
derivation, *327*
vindicated, *371*
energy-time uncertainty relation, *326*
decay of a state, *366*
Mandelshtam-Tamm version, *946*
enthalpy, *539*
enthalpy of vaporization, *561*
entropy, *548*
descriptive, 541
EPR, *412*
epsilon, *see* ϵ,ε
equipartition theorem, *573*
equivalent dose, *672*
eta, *see* η
Euler formula, *33*
eV, 1513
even-even nuclei
enhanced stability, *661*
Everett, III, *422*
every possible combination, *140*, 157
exchange force mechanism
and two-state systems, *355*
nuclear forces, *1162*
exchange integrals, *456*
exchange operator, *149*
exchange terms
twilight terms, *153*
exchanged
Las Vegas interpretation, *205*
excited determinants, *467*
exciton
intro, *300*
exclusion principle, *172*
exclusion-principle repulsion, *192*
expectation value, *114*
definition, *117*
simplified expression, *118*
versus average, 115

experimental evidence, *1200*
exponential function, 1513
exponential of an operator, *901*
exposure, *671*
extended zone scheme, *498*
  intro, *273*
extensive variable, *537*
extreme independent particle model, *719*
extreme single-particle model, *719*

F, 1513
$\mathcal{F}$, 1514
f, 1514
face centered cubic, *see* FCC
factorial, 1494
Faraday cage
  proposal for nuclei, 1482
fast ion conductors, 266
f block
  periodic table, 191
F-center
  intro, *301*
fermi, *686*
Fermi brim
  definition, *240*
Fermi decay, *817*
Fermi energy
  definition, *240*
  electrons in a box, *230*
Fermi factor, *241*
  definition, *241*
Fermi function
  intro, *1197*
  value, *1210*
Fermi integral
  intro, *821*
  value, *1211*
Fermi level
  definition, *240*
  line up
    Peltier cooler, *306*
Fermi surface
  electrons in a box, *230*
  periodic boundary conditions, *249*
  periodic zone scheme, *501*
  reduced zone scheme, *500*
Fermi temperature, *567*
Fermi theory
  comparison with data, *821*
Fermi theory of beta decay, *1191*
Fermi's golden rule, *369*, *1203*
Fermi-Dirac distribution
  canonical probability, *532*
  for given energy, *531*
  identify chemical potential, *562*
  intro, *237*
Fermi-Kurie plot, 827
fermions, *156*
  antisymmetrization requirement, *166*
  ground state, *228*
  intrinsic parity, *1188*
Feynman diagrams, 1110
Feynman slash notation, *1122*
field emission, 245
field operators, *931*
field strength tensor, *1021*
filled shells, *598*
filtering property, *383*
Fine structure, *1139*
fine structure
  hydrogen atom, *1138*
fine structure constant, *1138*
  in decay rates, *1065*
first Brillouin zone
  intro, *273*
first law of thermodynamics, 521, *540*
first-forbidden decays
  beta decay, *819*
fission
  energetics, *674*
  spontaneous
    definition, *663*
    overview of data, *656*
flopping frequency, 642
Floquet theory, *483*
fluorine-19
  nuclear spin, *729*
flux, *894*
Fock operator, *459*
Fock space kets
  beta decay, *1193*
Fock state, *914*
forbidden decays
  beta decay, *816*
forbidden transition
  intro, *340*
forbidden transitions
  alpha decay, *697*
force
  in quantum mechanics, *326*
four-vectors, *17*
Fourier analysis, *484*
Fourier coefficients, 1228

Fourier integral, 1229
Fourier series, 1228
  one-dimensional, *1079*
  three-dimensional, *1082*
Fourier transform, *393*, 1229
  one-dimensional, *1081*
  three-dimensional, *1083*
Fourier's law
  heat conduction, *895*
Fraunhofer lines, 299
free path, 254
free-electron gas
  intro, *228*
  model for crystal structure, *495*
  periodic box
    intro, *248*
  specific heat, *1377*
Frenkel defect, 265
ft-value, *821*
function, *34*, 35, 1514
functional, *865*, 1514
fundamental commutation relations
  as an axiom, *580*
  orbital angular momentum, *127*
  spin
    introduction, *160*
fundamental solution
  Poisson equation, *989*
fusion
  energetics, *674*

G, 1514
g, 1515
gage property, *1014*
Galilean transformation, *12*
gallium arsenide
  crystal structure, *277*
Galvani potential, *246*
Gamma, *see* Γ
gamma decay
  definition, *663*
gamma function, 1494
gamma matrices
  Dirac equation, *1122*
gamma rays
  intro, *829*
gamma vibration
  nuclei, 745
gamma, *see* γ
Gamow theory, *691*
Gamow-Teller decay, *817*
gauge theories
  basic ideas, *960*
gauge transformation
  electromagnetic potentials, *28*
Gauss' theorem, 1515
generalized coordinates, *858*
  intro, *903*
generalized momentum, *see* canonical
generator of rotations, *949*
geometric phase, *942*
germanium
  crystal structure, *277*
  detailed band structure, *277*
g-factor, *634*
Gibbs free energy, *556*
  microscopic, *562*
glueballs, 359
gluons, 358
grad, 1497
grad(ient), 1515
gradient, 1497
grain, *478*
grain boundaries, *478*
graphene
  electrical properties
    intro, 264
graphite
  electrical properties
    intro, 264
  intro, 197
gravitons, 359
gray, *671*
Green's function
  Laplacian, *627*
  Poisson equation, *989*
ground state
  absolute zero temperature, *212*
  atoms, *181*
  bosons, *211*
  fermions, *228*
  harmonic oscillator, *85*
  hydrogen atom, 107, *109*
  hydrogen molecular ion, *136*
  hydrogen molecule, *149*, 164, 167
  nonzero energy, *84*
group
  intro, *22*
group property
  coordinate system rotations, 1384
  Lorentz transformation, *22*
group theory, 333
group velocity, *392*

intro, *391*
Gupta-Bleuler condition, *1020*
gyromagnetic ratio, *633*

H, 1515
h, 1515
$\hbar$, 1515
half-life, *361*
halo nucleus, *730*
halogens, 188
Hamiltonian, *56*
atoms, *178*
classical, *861*
electromagnetic field, *605*
gives time variation, *317*
harmonic oscillator, *79*
partial, *81*
hydrogen atom, *100*
hydrogen molecular ion, *129*
hydrogen molecule, *142*
in matrix form, *176*
numbering of eigenfunctions, *56*
one-dimensional free space, *387*
relativistic, nonquantum, *28*
Hamiltonian dynamics
relation to Heisenberg picture, *903*
Hamiltonian perturbation coefficients, *1126*
Hankel functions
spherical, *880*
harmonic functions, 1243
harmonic oscillator, *78*
classical frequency, *79*
eigenfunctions, *85*
eigenvalues, *83*
energy spectrum, *83*
ground state, *85*
Hamiltonian, *79*
partial Hamiltonian, *81*
particle motion, 401
harmonic polynomials, *98*
Hartree product, 169, *447*
intro, 205
Hartree-Fock, *445*
Coulomb integrals, 456
exchange integrals, 456
restricted
closed shell, 450
open shell, 451
spin-adapted configuration, 453
unrestricted, 450
Hartree-Fock equations
general, *1351*
heat, *213*, *540*
heat capacity
valence electrons, *240*
heat conduction
electrons, *265*
heat flux density
including Peltier effect, *895*
omit density, *895*
heavy water, 657
Heisenberg
uncertainty principle, *53*
uncertainty relationship, *125*
helicity
definition, *1191*
photon, *977*
helion, *658*
helium
Bose-Einstein condensation, 216
helium ionization energy, *1128*
Hellmann-Feynman theorem, *1128*
Helmholtz decomposition, *1026*
Helmholtz equation, *1275*
Green's function solution, *1276*
Helmholtz free energy, *556*
microscopic, *562*
Hermitian conjugate, *46*
Hermitian conjugates
creation and annihilation operators, *918*
Hermitian matrices, 1523
Hermitian operators, *43*
hexacontatetrapole transition, *833*
hexadecapole transition, *833*
hidden variables, *59*, 412
hidden versus nonexisting, *100*
hieroglyph, 511, 1147
hole
nuclear shell model, 715
holes
in shells, *598*
light, heavy, split-off, 399
semiconductors
holes per state, *284*
holes per unit volume, *284*
intro, *263*
Hund's rules, 511
hybridization, *196*
hydrogen
metallic, 261
nonmetal, *260*
hydrogen atom, *100*
eigenfunctions, *109*

eigenvalues, *106*
energy spectrum, *106*
ground state, 107, *109*
Hamiltonian, *100*
relativistic corrections, *1138*
hydrogen bonds, *194*, 471
hydrogen molecular ion, *129*
bond length, *137*
experimental binding energy, *137*
ground state, *136*
Hamiltonian, *129*
shared states, *132*
hydrogen molecule, *142*
binding energy, *149*
bond length, *149*
ground state, *149*, 164, 167
Hamiltonian, *142*
hyperfine splitting, *1138*
hypersphere, 1515

I, 1517
$\Im$, 1517
$\mathcal{I}$, 1517
i, *35*, 1517
$\hat{\imath}$, 1517
i, *31*, 1517
reciprocal, *32*
ideal gas
quantum derivation, *568*
thermodynamic properties, *558*
ideal gas law, *568*
ideal magnetic dipole, *623*
ideality factor, 294
identical particles, *166*
identity matrix, 1352
identity operator, *47*
iff, *38*, 1517
imaginary part, *32*
impact parameter, 1103
impurities
ionic conductivity, *265*
optical effects, 301
incoherent radiation
absorption and emission, *379*
incompressibility
intro, 233
independent particle model, *719*
index notation, 1517, 1522
intro, *18*
indirect gap semiconductor, *275*
indistinguishable
definition, *912*
indistinguishable particles, *525*
(anti) symmetrization requirement, *525*
intro, *216*, *222*
inner product
multiple variables, *47*
inner product of functions, *38*
inner product of vectors, *37*
insulated system, *551*
insulators
examples, *259*
integer, 1517
integral Schrödinger equation, *904*
intelligent designers, *1475*
intensive variable, *537*
intermediate vector bosons, *356*
internal conversion, *850*
definition, *664*
intro, *830*
internal conversion coefficient, *852*
internal energy, *538*
internal pair production
intro, *830*
internal transition
definition, *664*
interpretation
interpretations, *58*
many worlds, *422*
orthodox, *57*
relative state, *422*
statistical, *57*
interstitials
ionic conductivity, 265
interval
special relativity, *see* space-time interval
intrinsic semiconductor, *282*
intrinsic state
nuclei, *739*
inverse, 1517
inverse beta decay
definition, *662*
inversion
parity operator, *330*
ionic bonds, *199*
ionic conductivity, *265*
ionic molecules, *472*
ionic solids, *472*
ionization, *107*
ionization energy, 472
Hartree-Fock, 461
helium, *1128*
hydrogen atom, *107*

irrotational, 1518
gradient of a scalar, 1533
vector potential, *1025*
irrotational flow, *1179*
islands of isomerism, *836*
iso, 1518
isobar
nuclei, *658*
isobaric analog states, *790*
isobaric multiplets, *790*
isobaric spin, *787*
isolated, 1518
isolated system, *551*
isomer, *835*
isomeric transition
definition, *664*
isospin, *787*
beta decay, *1194*
isothermal atmosphere, *242*
isotones, *657*
isotope, *657*
isotopic spin, *787*
i-spin, *787*

J, 1518
j, 1518
$\hat{\jmath}$, 1518

K, 1518
$\mathcal{K}$, 1519
K, 1519
k, 1519
$\hat{k}$, 1519
k_{B}, 1519
kappa, *see* κ
K-capture
definition, *662*
Kelvin coefficient, *314*
Kelvin heat, *313*
Kelvin relationships
thermoelectrics, *897*
intro, *313*
ket, *38*, 1495, 1522
ket notation
spherical harmonics, *97*
spin states, *156*
kinetic energy
nuclear decay, *813*
operator, *55*
kinetic energy operator
in spherical coordinates, *101*
Klein-Gordon equation, 603, *906*
kmol, 1519
Koopman's theorem, *461*
Kramers relation, *1421*
Kronecker delta, 1352

L, 1519
$\mathcal{L}$, 1519
L, 1519
l, 1519
ℓ, 1520
$\pounds$, 1520
ladder operators
angular momentum, *581*
Lagrangian
for classical fields, *991*
relativistic, *27*
simplest case, *857*
Lagrangian density, *862*
example, *995*
Lagrangian dynamics
for classical fields, *991*
Lagrangian mechanics, *857*
Lagrangian multipliers
derivations, *1334*
for variational statements, *437*
Lamb shift, 1138, 1147
Lambda, *see* Λ
lambda, *see* λ
Landé g-factor, *1147*
lanthanides, 191
lanthanoids, 191
Laplace equation, *1390*
solution in spherical coordinates, *1179*
solutions in spherical coordinates, *1243*
Laplacian, 1497
Larmor frequency
definition, *639*
Larmor precession, *641*
laser
operating principle, *373*
laser diode, *303*
latent heat of vaporization, *561*
lattice
diamond, *493*
FCC, *475*
primitive vectors, *278*
intro, *475*
lithium (BCC), *477*
NaCl, *475*
one-dimensional, *478*
primitive translation vector, *483*
primitive translation vectors

diamond, *494*
reciprocal, *see* reciprocal lattice
translation operator, *396*
unit cell, *475*
zinc blende (FCC), *278*
law of mass action
semiconductors, *288*
L-capture
definition, *662*
Lebesgue integration, 1083
LED, *303*
length of a vector, *37*
Lennard-Jones potential, *469*
Casimir-Polder, *470*
lepton number
conservation, *331*, *801*
lifetime, *361*, *667*
light wave
plane
terminology, *374*
light waves
classical, *614*
light-cone
special relativity, *16*
light-emitting diode, *303*
light-emitting diodes
crystal momentum, *303*
lim, 1520
linear combination, 1520
linear dependence, 1520
linear independence, 1520
linear momentum
classical, *53*
eigenfunctions, *385*
eigenvalues, *385*
operator, *55*
symmetry and conservation, *327*
linear polarization
from Maxwell's equations, *614*
from second quantization, *1041*
intro, *344*
photon wave function, *977*
liquid drop model
nuclear binding energy, *687*
nuclear radius, *686*
nuclei
intro, *686*
locality
quantum field theories, 1003
localization
absence of, *389*
London forces, *469*
Casimir-Polder, *470*
Lorentz factor, *10*
Lorentz force
derivation, *1227*
special relativity, *27*
Lorentz invariant
field theories, *1008*
Lorentz transform
improper, 871
nonorthochronous, 872
Lorentz transformation, *11*
derivation, *1224*
group property, *22*
group property derivation, *1225*
index notation, *18*
parity transformation, *871*
time-reversal, *872*
Lorentz-Fitzgerald contraction, *9*
Lorentz[ian] profile, 371
Lorenz condition, *972*
classical electromagnetics, *1125*
not Lorentz, *1017*
unconventional derivation, *1017*
Lorenz gauge, *972*
classical electromagnetics, *1125*
lowering indices, *873*
luminosity
particle beam, 1102
Lyman transitions, *107*

M, 1521
$\mathcal{M}$, 1521
M, 1521
m, 1521
m_{e}, 1521
m_{n}, 1521
m_{p}, 1521
Madelung constant, *474*
magic numbers
40?, *711*
and beta decay, *809*
intro, *666*
shell model, *701*
magnetic dipole
idealized, *623*
magnetic dipole moment
classical, *633*
magnetic dipole transition
intro, *341*
selection rules, *349*
relativistic, *349*

magnetic dipole transitions
Hamiltonian, *1304*
magnetic moment
nuclei, *771*
magnetic multipole
photon states, *980*
magnetic quantum number, *94*
magnetic spin anomaly, *634*
magnetic vector potential
classical derivation, *1395*
in the Dirac equation, *1399*
quantum derivation, *605*
relativistic derivation, *27*
magnitude, *32*
main group
periodic table, 188
majority carriers, 287
maser
ammonia, *154*
operating principle, *373*
mass number, *657*
mass-energy relation
derivation, *24*
Dirac equation, *603*
fine-structure, *1140*
for nuclei, *672*
Lagrangian derivation, *28*
need for quantum field theory, *908*
matching regions, 1097
mathematicians, 20, 1353
matrix, *40*, 1522
matrix element, *363*
maximum principle
Laplace equation, *1390*
Maxwell relations, *557*
Maxwell's equations, *607*
"derivation" from scratch, *981*
Maxwell-Boltzmann distribution
canonical probability, *532*
for given energy, *531*
intro, *241*
mean lifetime, *667*
mean value property
Laplace equation, *1390*
measurable values, *57*
measurement, *58*
Meisner
credit, *1481*
mesic charge, 1170
meson, 156
mesons, 356
metalloids, 188
compared to semimetals, *264*
metals, *476*
examples, *259*
method of stationary phase, *1326*
metric prefixes, 1524
Minkowski metric, 871
minority carriers, 287
mirror nuclei, 686, *793*
beta decay, *1202*
mass difference data, *809*
mirror operator, *886*
molar mass, *538*
versus molecular mass etc., 1524
mole, *538*
molecular mass, *538*
versus molar mass etc., 1524
molecular solids, *469*
molecules
ionic, *472*
moment
electromagnetic
nuclei, *771*
momentum conservation
beta decay, *1208*
momentum space wave function, *385*
integral transform
one-dimensional, *1081*
three-dimensional, *1083*
Moszkowski estimate, *1068*
Moszkowski unit, *1068*
derivation, *1077*
Mott insulators, 262
moving mass, *4*
derivation, *23*
Lagrangian derivation, *26*
mu, *see* μ
multipole expansion, 626
multipole transition
intro, *340*

N, 1525
N, 1525
n, 1525
n, 1526
nabla, 1497
nanoionics, 266
natural, 1526
natural width, *335*
nearly-free electron model, *502*
negaton, *663*
negatron, *663*

neon-19
nuclear spin, *729*
Neumann functions
spherical, *879*
neutrino
needed in beta decay, *814*
neutrinos
do not conserve parity, *1188*
helicity, *1191*
no intrinsic parity, *1188*
relativistic theory, *1186*
states like screws, *1190*
neutron
intro, *651*
mixed beta decay, *817*
neutron emission
definition, *663*
neutron excess, *659*
neutron stars, 233, 663
Newton's second law
in quantum mechanics, *326*
Newtonian analogy, *55*
Newtonian mechanics, *50*
in quantum mechanics, *324*
nitrogen-11
nuclear spin, *730*
NMR
spin one-half, *777*
noble gas, *182*
noble gases, 188
non canonical Hartree-Fock equations, *1351*
nonequilibrium thermodynamics, *893*
nonexisting versus hidden, *100*
nonholonomic, *942*
Nordheim rules, *760*
norm of a function, *38*
normal operators
are abnormal, 1526
normalized, *38*
normalized wave functions, *52*
n-p-n transistor, *295*
n-type semiconductor, *285*
nu, *see* ν
nuclear decay
overview of data, *656*
nuclear force, *646*
nuclear forces
pion exchange mechanism, *1162*
nuclear magnetic resonance, *637*
nuclear magneton, *635*, *775*
nuclear parity
intro, *649*
nuclear radius, *686*
nuclear reactions
antiparticles, 801, *803*
nuclear spin
intro, *648*
nuclei
beta vibration, 745
do not contain electrons, *1192*
gamma vibration, 745
internal conversion, *850*
intro, *656*
liquid drop model
intro, *686*
pairing energy
evidence, *676*
parity
data, *764*
intro, *707*
perturbed shell model, *717*
rotational bands, *738*
spin one-half, *743*
spin zero, *745*
shell model, *701*
nonspherical nuclei, 742
Rainwater-type justification, *706*
shells
evidence, *676*
spin
Nordheim rules, *760*
stable odd-odd ones, *810*
unperturbed shell model, *717*
vibrating drop model
derivations, *1177*
stability, *732*
vibrational states, *734*
nucleon number, *657*
nucleons, 646

O, 1526
OBEP, 1176
oblate spheroid, *774*
observable values, *57*
occupation numbers
beta decay, *1193*
intro, *211*
single-state, *913*
octupole transition, *833*
octupole vibration
nuclei, *737*
odd-odd nuclei
reduced stability, *661*

odd-particle shell model, *718*
Omega, *see* Ω
omega, *see* ω
one-boson exchange potential, 1176
one-dimensional free space
  Hamiltonian, *387*
one-particle shell model, *719*
one-pion exchange potential, *1166*
Onsager reciprocal relations, *897*
OPEP, *1166*
  intro, *684*
OPEP potential
  introduction, *1166*
  loose derivation, *1167*
operator
  exponential of an operator, *901*
operators, *40*
  angular momentum component, *93*
  Hamiltonian, *56*
  kinetic energy, *55*
    in spherical coordinates, *101*
  linear momentum, *55*
  position, *55*
  positive (semi)definite, *1282*
  potential energy, *56*
  quantum mechanics, *54*
  square angular momentum, *96*
  total energy, *56*
opposite, 1526
orbital, *446*
orbital angular momentum
  relativistic coupling with spin, *1189*
orthodox interpretation, *57*
orthogonal, *38*
orthonormal, *38*
orthonormal matrix
  in coordinate rotations, *870*

P, 1526
$\mathcal{P}$, 1526
$\mathcal{P}$, 1526
P, 1526
p, 1526
p, 1527
parity
  alpha decay, *697*
  combination
    intro, *338*
  conservation in decays, *336*
  intro, *328*
  nuclei
    data, *764*
    intro, *707*
  orbital
    derivation, *1401*
  spherical harmonics
    derivation, *1242*
  symmetry and conservation, *327*
  violation of conservation, *330*
parity operator
  spatial inversion, *330*
parity transformation, *330*
  as a Lorentz transformation, *871*
parity violation
  Wu experiment, *827*
Parseval identity
  Fourier series
    one-dimensional, *1080*
    three-dimensional, *1083*
  Fourier transform
    one-dimensional, *1082*
    three-dimensional, *1083*
partial wave amplitude, *1106*
partial wave analysis, *1103*
  phase shifts, *1106*
particle
  tensor, *1200*
particle exchange
  symmetry, *888*
partition function, *533*
Paschen transitions, *107*
passive view, *948*
Pasternack relation, *1421*
Pauli exclusion principle, *172*, 447
  atoms, *183*
  common phrasing, *184*
Pauli repulsion, *192*
Pauli spin matrices, *598*
  generalized, *601*
p block
  periodic table, 190
Peltier coefficient, *306*
Peltier effect, *304*
periodic box, *248*
  a tricky version, *1029*
  beta decay, *1196*
periodic table, *183*
  full, *188*
periodic zone scheme, *501*
  intro, *276*
permanents, *171*
permittivity of space, *101*
perpendicular bisector, 1527

perturbation theory
helium ionization energy, *1128*
second order, *1127*
time dependent, *366*
time-independent, *1126*
weak lattice potential, *503*
perturbed shell model, *715*
phase angle, 1527
phase equilibrium, 560
phase shift
partial waves, *1106*
phase speed, *389*
phenomenological nuclear potentials, *685*
Phi, *see* Φ
phi, *see* ϕ,φ
phonons, 573
nuclei, *734*
photoconductivity
intro, *301*
photon, *107*, 1527
energy, *108*
spin value, *156*
wave function, *971*
photons
density of modes, *210*
photovoltaic cell, *302*
physicists, 15, 17–19, 58, 59, 94, 98, 110, 190, 191, 229, 232, 240, 247, 268, 276, 280, 287, 310, 322, 327, 330, 331, 338, 340, 342, 347, 356, 357, 360, 371, *377*, 380, 444, 465, 468, 483, 492, 511, 615, 616, 633, 634, 648, 649, 656–658, 662–664, 667, 671, 673, 691, 692, 704, 719, *774*, 797, 828, 830, 835, 847, 852, 853, 874, 875, 896, 951, 960, 1009, 1010, 1014, 1020, 1033, 1077, 1078, 1102, 1103, 1106, 1160, 1194, 1200, 1202, 1456, 1460, 1464, 1466, 1481, 1524, 1527
hypothetical shortcomings, 16, 982
more or less redeemed, 359, 363, 667, 830, 1242, 1248
more or less trusted, 1167
redeemed, 258, 275, 1139, 1206
unverified shortcomings, 1107
pi, *see* π
pi bonds, *194*
pion exchange
multiple, *1175*
pions
intro, *681*
Plancherel theorem, *1082*
Planck's blackbody spectrum, *226*
Planck's constant, 55
Planck-Einstein relation, *108*
derivation, *906*
p-n junction, *289*
p-n-p transistor, *295*
point charge
static, *616*
pointer states, *112*
Poisson bracket, *903*
Poisson equation, *627*
fundamental solution, *989*
Green's function solution
derivation, *1214*
screened
Green's function solution, *1217*
variational derivation, *988*
polar bonds, *194*
polar coordinates, 1498
polar vector, *969*
polariton
Bose-Einstein condensation, 217
polarization, *see* linear polarization, circular polarization
poly-crystalline, *478*
population inversion, *373*
position
eigenfunctions, *382*
eigenvalues, *382*
operator, *55*
positive (semi)definite operators, *1282*
positon, *663*
positron emission, *662*
possible values, *57*
potassium-40
decay modes, *810*
potential, 1527
existence, *1389*
potential energy
operator, *56*
potential energy surfaces, *445*
Poynting vector, *615*
prefixes
YZEPTGMkmunpfazy, 1524
pressure, *539*
primitive cell, *489*
in band theory, *261*
versus unit cell, *279*
primitive translation vector
one-dimensional, *483*

primitive translation vectors
FCC
intro, *278*
lithium (BCC), *488*
reciprocal lattice, *491*
primitive vectors, *see* above
principal quantum number, *104*
principle of relativity, *6*
probabilities
evaluating, *91*
from coefficients, *59*
probability current, *1111*
probability density, *144*
probability to find the particle, *51*
projection operator, *47*
prolate spheroid, *774*
promotion, *196*
nuclei, *730*
prompt neutrons, *751*
proper distance, *14*
as dot product, *17*
proper time, *14*
causality, *16*
proton
intro, *650*
proton emission
definition, *663*
pseudoscalar particle, *1176*
pseudovector, *969*
pseudovector particle, *1176*
Psi, *see* Ψ
psi, *see* ψ
p-type semiconductor, *285*
pure substance, *519*
p_x, 1528
Pythagorean theorem, 14

Q, 1528
q, 1528
quadrupole moment
electric
intro, *653*
nuclei, *771*
intrinsic
nuclei, *780*
spin one-half, *777*
quadrupole transition, *833*
intro, *340*
quadrupole transitions
electric
Hamiltonian, *1305*
selection rules, *349*
quadrupole vibration
nuclei, *737*
quality factor, *672*
quantum chromodynamics, 647
intro, 356
quantum confinement, *234*
single particle, *74*
quantum dot, *76*
density of states, *236*
quantum electrodynamics
electron g factor, 634
Feynman's book, 910
intro, 355
quantum field
definition, *936*
quantum field theory, *908*
Coulomb potential derivation, *1024*
Koulomb potential derivation, *981*
quantum interference, 52
quantum mechanics
acceleration, *326*
force, *326*
Newton's second law, *326*
Newtonian mechanics, *324*
velocity, *325*
quantum well, *76*
density of states, *235*
quantum wire, *76*
density of states, *236*
quark
spin, 156
quarks, 356, 647
Dirac equation, 603
proton and neutron, 634
Q-value
alpha and beta decay, *813*
nuclei, *692*

R, 1528
$\mathcal{R}$, 1529
$\Re$, 1529
r, 1529
$\vec{r}$, 1529
Rabi flopping frequency, 642
rad, *671*
radiation
emission and absorption, *372*
quantization, 1032
radiation probability, *see* decay rate
radiation weighting factor, *672*
radioactivity
intro, *656*

radium emanation, *664*
radium X, 664
raising indices, *873*
Ramsauer effect, 329
random number generator, *58*
rare earths, 191
RaX, *664*
Rayleigh formula
  partial waves, *881*
  spherical Bessel functions, *879*
RE, *664*
real part, *32*
reciprocal, 1529
reciprocal lattice
  lithium, *491*
  NaCl, *491*
  one-dimensional, *484*
  primitive vectors, *491*
  three-dimensional, *491*
recombination
  semiconductors, *288*
recombination centers, 291
reduced mass
  hydrogen atom electron, *101*
reduced zone scheme, *498*
  intro, *273*
reflection coefficient, *405, 406, 1113*
relative state formulation, 425
relative state interpretation, *422*
relativistic corrections
  hydrogen atom, *1138*
Relativistic effects
  Dirac equation, *602*
relativistic mass, *see* moving mass
relativistic quantum mechanics
  beta decay, *1193*
relativity, *see* special relativity, 1529
rem, *672*
residual strong force, *647*
resistivity
  electrical, *254, 255*
resonance factor, *642*
rest mass, *4*
rest mass energy, *5*
  derivation, *24*
restricted Hartree-Fock, *450*
reversibility, *543*
RHF, *450*
rho, *see* ρ
roentgen, *671*
röntgen, *671*
rot, 1497, 1530
rotational band
  nuclei, *741*
rotational bands
  seenuclei, 738

S, 1530
$\mathcal{S}$, 1530
$\mathcal{S}$, 1530
S, 1531
s, 1531
s, 1531
saturated, 560
s block
  periodic table, 190
scalar, 1531
scalar particle, *1176*
scattering, *402*
  one-dimensional coefficients, *405*
  three-dimensional, *1100*
scattering amplitude, *1101*
Schmidt lines, *778*
Schottky defect, 266
Schottky effect, *245*
Schrödinger equation, *317*
  failure?, *420*
  integral version, *904*
Schrödinger's cat, *410*
second law of thermodynamics, *541*
second quantization, *926*, 1033
Seebeck coefficient, *310*
Seebeck effect, *309*
seething cauldron, 1041
selection rules
  derivation, *1302*
  electric dipole transitions, *347*
    relativistic, *348*
  electric quadrupole transitions, *349*
  intro, *345*
  magnetic dipole transitions, *349*
    relativistic, *349*
self-adjoint, 1507
self-conjugate nuclei, *793*
self-consistent field method, *460*
semi-conductors
  band gap, 492
semi-empirical mass formula, *687*
semiconductor
  degenerate, *287*
  direct gap, *275*
  intrinsic, *282*
  intro, *264*

n and p-type, *285*
semiconductor laser, *303*
semiconductors
compensation, *288*
conduction electrons per state, *282*
conduction electrons per volume, *284*
crystal structure, *277*
doping, *282*
holes
intro, *263*
holes per state, *284*
holes per unit volume, *284*
semimetal
intro, *264*
separation of variables, *80*
for atoms, *178*
linear momentum, *385*
position, *382*
shell model
with pairing, *715*
with perturbations, *715*
shell model of nuclei, *701*
shielding approximation, *179*
Shockley diode equation, *294*
SI prefixes, 1524
sievert, *672*
sigma, *see* σ
sigma bonds, *193*
silicon
crystal structure, *277*
simple cubic lattice, *497*
sin, 1531
singlet color state, 358
singlet state, *164*
derivation, 587
skew-Hermitian, 1507
Slater determinants, *170*
small perturbation theory, *503*
solar cell, *302*
solar spectrum, 299
solenoidal, 1531
vector potential, *1026*
solid angle, *1101*, 1507
infinitesimal
spherical coordinates, 1533
solid electrolytes, 266
solids, *469*
covalent, *492*
ionic, *472*
molecular, *469*
spectra
intro, *299*
sp^n hybridization, *196*
space charge region, *291*
space-like
special relativity, *15*
space-time
special relativity, *17*
space-time interval
ambiguous definition, *15*
causality, *16*
spatial inversion
parity operator, *330*
special relativity, *3*
canonical momentum, *27*
causality, *16*
four-vectors, *17*
dot product, *17*
in terms of momentum, *5*
index notation, *18*
light-cone, *16*
Lorentz force, *27*
Lorentz transformation, *11*
Lorentz-Fitzgerald contraction, *9*
mass-energy relation, *4*
mechanics
intro, *22*
Lagrangian, *25*
momentum four-vector, *23*
proper distance, *14*
as dot product, *17*
proper time, *14*
rest mass energy, *5*
space-like, *15*
space-time, *17*
space-time interval, *15*
superluminal interaction, *15*
time-dilation, *9*
time-like, *15*
velocity transformation, *13*
specific activity, *671*
specific decay rate, *670*
specific heat
constant pressure, *540*
constant volume, *540*
values, *573*
specific volume, *537*
molar, *538*
spectral analysis
intro, *298*
spectral line broadening, *335*
spectrum, 1531

hydrogen, *108*
spherical Bessel functions, *879*
spherical coordinates, *93*, 1531
unit vectors, 1532
volume integral, 1532
spherical Hankel functions, 880
spherical harmonics
derivation, 1377
derivation from the ODE, *1240*
derivation using ladders, *1377*
generic expression, *1242*
intro, *96*
Laplace equation derivation, *1244*
parity, *1242*
spherical Neumann functions, *879*
spheroid, *774*
spin, *155*
fundamental commutation relations
introduction, *160*
nuclei
data, *755*
value, *156*
x- and y-eigenstates, 601
spin down, *156*
spin orbital, *446*
spin states
ambiguity in sign, 1384
axis rotation, 1383
spin up, *156*
spin-adapted configuration, 453
spin-orbit interaction
nucleons, *707*
spinor, *158*
spontaneous emission
multiple initial or final states, *370*
quantum derivation, *1044*
spontaneous fission, *751*
s state, *110, 111*
standard deviation, *114*
definition, *116*
simplified expression, *118*
standard model, 332
Stark effect, 1134
stationary states, *321*
statistical interpretation, *57*
Stefan-Boltzmann formula, *570*
Stefan-Boltzmann law, *227*
steradians, *1102*
Stern-Gerlach apparatus, *636*
stoichiometric coefficient, *561*
Stokes' theorem, 1533
string theory, 936
strong force, 647
intro, 356
superallowed beta decays, *1202*
superallowed decay
beta decay, *825*
superconductivity, 256
Cooper pairs, 217
superfluidity
Feynman argument, 222
superionic conductors, 266
superluminal interaction
Bell's theorem, *411*
hidden variables, *412*
many worlds interpretation, *425*
quantum, *52*
do not allow communication, *414*
produce paradoxes, *414*
relativistic paradoxes, *15*
surface tension, *732*
symmetrization requirement
fermions, *see* antisymmetrization
graphical depiction, *524*
identical bosons, *166*
indistinguishable particles, *525*
using groupings, *171*
using occupation numbers, *913*
using permanents, *171*
symmetry, 1533

T, 1533
$\mathcal{T}$, 1534
t, 1534
tantalum-180m, *832*
tau, *see* τ
temperature, *520*, 1534
definition, *532*
Carnot, *548*
definition using entropy, *558*
intro, *212*
tensor particle, *1200*
tensor potential
deuteron, 1160
tensors
compared to linear algebra, *870*
intro, *18*
thermal de Broglie wavelength, *564*
thermal efficiency, *546*
thermal equilibrium, *520*
thermionic emission, *244*
thermocouple, *309*
thermodynamics

first law, *540*
second law, *541*
third law, *553*
thermoelectric generator, *310*
thermoelectrics
figure of merit, *891*
macroscopic equations, *893*
thermogenerator, *310*
Theta, *see* Θ
theta, *see* θ,ϑ
third law of thermodynamics, *553*
Thomson coefficient, *314*
Thomson effect, *313*
Thomson relationships
thermoelectrics, *897*
intro, *313*
throw the dice, 59
TID, *671*
time
directionality, *428*
time symmetry
reservations, *963*
time variation
Hamiltonian, *317*
time-dependent perturbation theory, *366*
time-dilation, *9*
time-like
special relativity, *15*
time-reversal
as a Lorentz transformation, *872*
tin
white and grey, *262*
tissue weighting factor, *672*
T-multiplets, *790*
total cross-section, *1103*
total energy
operator, *56*
total ionizing dose, *671*
transistor, *295*
transition
multipole
selection rules, *833*
multipole names, *833*
quadrupole, *see* quadrupole transition
transition elements, 191
transition metals, 191
transition probability, *see* decay rate
transition rate
spontaneous, *see* decay rate
transitions
hydrogen atom, *107*
translation operator
crystals, *396*
transmission coefficient, *405, 406*, *1114*
transparent crystals, *300*
transpose
matrices, 1523
transpose of a matrix, 1512
transverse gauge
classical electromagnetics, *1125*
traveling waves, *see* linear polarization
triakontadipole transition, *833*
triangle inequality, *337*
triple alpha process, 674
triple product, 1534
triplet states, *164*
derivation, 587
tritium, *658*
triton, 658
tunneling, *403*
field emission, 245
Stark effect, 1137
WKB approximation, *405*
Zener diodes, 297
turning point, *401*
turning points
WKB approximation, 1094
twilight terms, *152*
exchange terms, *153*
Lennard-Jones/London force, *1117*
lithium hydride, *153*
spontaneous emission, *1049*
two state systems
ground state energy, *150*
time variation, *354*
two-state systems
atom-photon model, *1044*

U, 1535
$\mathcal{U}$, 1535
u, 1535
u, 1535
UHF, *450*
uncertainty principle
angular momentum, *99*
energy, *87*, *321*
Heisenberg, *53*
position and linear momentum, *53*
uncertainty relationship
generalized, *124*
Heisenberg, *125*
unified atomic mass unit, *673*
unit cell

FCC, *475*
intro, *475*
lithium (BCC), *477*
versus primitive cell, *279*
zinc blende, *278*
unit matrix, 1352, 1523
unit vectors
in spherical coordinates, 1532
unitary
Fourier series, *1080*
matrix, *1353*
time advance operator, *902*
unitary matrix
in coordinate rotations, *870*
unitary operator, *951*
unitary operators, 1507
universal gas constant, *558*, *573*
universal mass unit, *673*
unperturbed shell model, *715*
unrestricted Hartree-Fock, *450*

V, 1535
$\mathcal{V}$, 1535
v, 1535
$\vec{v}$, 1536
vacancies
ionic conductivity, 265
optical effects, 301
vacuum energy, 429, *915*, 1040
seething cauldron, 1041
vacuum state, *916*
valence band
intro, *258*
values
observable, *57*
Van der Waals forces, *469*
Casimir-Polder, *470*
variational calculus
worked out example, *864*
variational method, *135*
helium ionization energy, *1130*
hydrogen molecular ion, *135*
hydrogen molecule, *148*
variational principle, *433*
basic statement, *433*
differential form, *435*
Lagrangian multipliers, *436*
vector, *34*, 1536
vector bosons, *964*
vector particle, *1176*
vectorial product, 1536
velocity
in quantum mechanics, *325*
vibrational states
seenuclei, 734
vibronic coupling terms, *1340*
virial theorem, *324*
virtual work, *860*
viscosity, *544*
Volta potential, *247*
volume integral
in spherical coordinates, 1532
von Weizsäcker formula, *687*

W, 1536
w, 1537
$\vec{w}$, 1537
warp factor, *15*
wave function, *50*
multiple particles, *140*
multiple particles with spin, *161*
with spin, *157*
wave number, 42, 65, *392*
Floquet, 484
Fourier versus Floquet, 484
one-dimensional Fourier series, *1079*
one-dimensional Fourier transform, *1081*
wave number vector
and linear momentum, *249*
Bloch function, *491*
Fourier series, *1082*
Fourier transform, *1083*
wave numbers, 206
wave packet
accelerated motion, 400
definition, *390*
free space, *387*, *399*
harmonic oscillator, 401
partial reflection, *403*
physical interpretation, *390*
reflection, 401
wave vector
conservation, *276*
weak force
intro, 356
Weisskopf estimates, *1068*
comparison with data, *844*
figures, *836*
Weisskopf unit
derivation, *1077*
Weisskopf units, *1068*
well
deuteron, *1154*
Weyl neutrinos, *1186*

width
    particle decay, 647
width of a state, *335*
Wigner 3j,6j and 9j coefficients, 1456
Wigner-Eckart theorem, 1457
Wigner-Seitz cell, *489*
WKB approximation
    connection formulae, *1096*
WKB connection formulae, *1097*
WKB theory, *1092*
Woods-Saxon potential, *707*
work function, *245*
Wronskian, *1112*
W.u., *1077*

X, 1537
x, 1537
xi, *see* ξ
X-ray diffraction, *512*

Y, 1537
Y_l^m, 1537
y, 1537
yrast line, 748
YSZ, 266
yttria-stabilized zirconia, 266
Yukawa potential, *1165*
    loose derivation, *1162*

Z, 1537
z, 1537
Zeeman effect, 1133
    intermediate, 1146
    weak, *1146*
Zener diode, *296*
zero matrix, 1523
zero point energy, *443*
zeroth law of thermodynamics, *520*
zinc blende
    crystal structure, *277*
ZnS, *see* zinc blende