

Quantum Mechanics for Engineers

Leon van Dommelen

08/09/10 Version 5.08 alpha

Copyright

Copyright 2004, 2007, 2008, 2010 and on, Leon van Dommelen. You are allowed to copy or print out this work for your personal use. You are allowed to attach additional notes, corrections, and additions, as long as they are clearly identified as not being part of the original document nor written by its author.

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.

Dedication

To my parents, Piet and Rietje van Dommelen.

Contents

Preface	xxxiii
To the Student	xxxiii
Acknowledgments	xxxiv
Comments and Feedback	xxxvi
I Basic Quantum Mechanics	1
1 Mathematical Prerequisites	3
1.1 Complex Numbers	3
1.2 Functions as Vectors	6
1.3 The Dot, oops, INNER Product	8
1.4 Operators	12
1.5 Eigenvalue Problems	13
1.6 Hermitian Operators	15
1.7 Additional Points	17
1.7.1 Dirac notation	18
1.7.2 Additional independent variables	18
2 Basic Ideas of Quantum Mechanics	19
2.1 The Revised Picture of Nature	20
2.2 The Heisenberg Uncertainty Principle	23
2.3 The Operators of Quantum Mechanics	25
2.4 The Orthodox Statistical Interpretation	27
2.4.1 Only eigenvalues	27
2.4.2 Statistical selection	29
2.5 A Particle Confined Inside a Pipe	30
2.5.1 The physical system	30
2.5.2 Mathematical notations	31
2.5.3 The Hamiltonian	32
2.5.4 The Hamiltonian eigenvalue problem	33
2.5.5 All solutions of the eigenvalue problem	33

2.5.6	Discussion of the energy values	37
2.5.7	Discussion of the eigenfunctions	39
2.5.8	Three-dimensional solution	41
2.5.9	Quantum confinement	45
2.6	The Harmonic Oscillator	47
2.6.1	The Hamiltonian	48
2.6.2	Solution using separation of variables	49
2.6.3	Discussion of the eigenvalues	52
2.6.4	Discussion of the eigenfunctions	54
2.6.5	Degeneracy	58
2.6.6	Non-eigenstates	60
3	Single-Particle Systems	63
3.1	Angular Momentum	64
3.1.1	Definition of angular momentum	64
3.1.2	Angular momentum in an arbitrary direction	65
3.1.3	Square angular momentum	67
3.1.4	Angular momentum uncertainty	71
3.2	The Hydrogen Atom	72
3.2.1	The Hamiltonian	72
3.2.2	Solution using separation of variables	73
3.2.3	Discussion of the eigenvalues	78
3.2.4	Discussion of the eigenfunctions	81
3.3	Expectation Value and Standard Deviation	86
3.3.1	Statistics of a die	87
3.3.2	Statistics of quantum operators	88
3.3.3	Simplified expressions	90
3.3.4	Some examples	91
3.4	The Commutator	93
3.4.1	Commuting operators	94
3.4.2	Noncommuting operators and their commutator	95
3.4.3	The Heisenberg uncertainty relationship	96
3.4.4	Commutator reference [Reference]	97
3.5	The Hydrogen Molecular Ion	100
3.5.1	The Hamiltonian	101
3.5.2	Energy when fully dissociated	101
3.5.3	Energy when closer together	102
3.5.4	States that share the electron	103
3.5.5	Comparative energies of the states	106
3.5.6	Variational approximation of the ground state	106
3.5.7	Comparison with the exact ground state	108

4 Multiple-Particle Systems	111
4.1 Wave Function for Multiple Particles	112
4.2 The Hydrogen Molecule	114
4.2.1 The Hamiltonian	114
4.2.2 Initial approximation to the lowest energy state	115
4.2.3 The probability density	117
4.2.4 States that share the electrons	118
4.2.5 Variational approximation of the ground state	120
4.2.6 Comparison with the exact ground state	121
4.3 Two-State Systems	122
4.4 Spin	127
4.5 Multiple-Particle Systems Including Spin	129
4.5.1 Wave function for a single particle with spin	129
4.5.2 Inner products including spin	131
4.5.3 Commutators including spin	132
4.5.4 Wave function for multiple particles with spin	133
4.5.5 Example: the hydrogen molecule	136
4.5.6 Triplet and singlet states	136
4.6 Identical Particles	138
4.7 Ways to Symmetrize the Wave Function	140
4.8 Matrix Formulation	146
4.9 Heavier Atoms [Descriptive]	150
4.9.1 The Hamiltonian eigenvalue problem	150
4.9.2 Approximate solution using separation of variables	151
4.9.3 Hydrogen and helium	153
4.9.4 Lithium to neon	155
4.9.5 Sodium to argon	159
4.9.6 Potassium to krypton	160
4.9.7 Full periodic table	161
4.10 Pauli Repulsion [Descriptive]	164
4.11 Chemical Bonds [Descriptive]	165
4.11.1 Covalent sigma bonds	165
4.11.2 Covalent pi bonds	166
4.11.3 Polar covalent bonds and hydrogen bonds	167
4.11.4 Promotion and hybridization	169
4.11.5 Ionic bonds	172
4.11.6 Limitations of valence bond theory	173
5 Macroscopic Systems	175
5.1 Intro to Particles in a Box	176
5.2 The Single-Particle States	178
5.3 Density of States	180

5.4	Ground State of a System of Bosons	183
5.5	About Temperature	184
5.6	Bose-Einstein Condensation	186
5.6.1	Rough explanation of the condensation	189
5.7	Bose-Einstein Distribution	195
5.8	Blackbody Radiation	196
5.9	Ground State of a System of Electrons	200
5.10	Fermi Energy of the Free-Electron Gas	202
5.11	Degeneracy Pressure	204
5.12	Confinement and the DOS	206
5.13	Fermi-Dirac Distribution	210
5.14	Maxwell-Boltzmann Distribution	214
5.15	Thermionic Emission	217
5.16	Chemical Potential and Diffusion	218
5.17	Intro to the Periodic Box	220
5.18	Periodic Single-Particle States	221
5.19	DOS for a Periodic Box	224
5.20	Intro to Electrical Conduction	224
5.21	Intro to Band Structure	228
5.21.1	Metals and insulators	229
5.21.2	Typical metals and insulators	232
5.21.3	Semiconductors	234
5.21.4	Semimetals	235
5.21.5	Electronic heat conduction	236
5.21.6	Ionic conductivity	237
5.22	Electrons in crystals	237
5.22.1	Bloch waves	238
5.22.2	Example spectra	239
5.22.3	Effective mass	241
5.22.4	Crystal momentum	243
5.23	Semiconductors	247
5.24	The <i>p-n</i> junction	253
5.25	The transistor	259
5.26	Zener and avalanche effects	261
5.27	Optical applications	262
5.27.1	Atomic spectra	262
5.27.2	Spectra of solids	263
5.27.3	Band gap effects	263
5.27.4	Effects of crystal imperfections	264
5.27.5	Photoconductivity	264
5.27.6	Photovoltaic cells	265
5.27.7	Light-emitting diodes	265

5.28	Thermoelectric applications	267
5.28.1	Peltier effect	268
5.28.2	Seebeck effect	272
5.28.3	Thomson effect	277
6	Time Evolution	279
6.1	The Schrödinger Equation	280
6.1.1	Intro to the equation	281
6.1.2	Some examples	282
6.1.3	Energy conservation [Descriptive]	285
6.1.4	Stationary states [Descriptive]	286
6.1.5	Particle exchange [Descriptive]	287
6.1.6	Energy-time uncertainty relation [Descriptive]	289
6.1.7	Time variation of expectation values [Descriptive]	290
6.1.8	Newtonian motion [Descriptive]	291
6.1.9	The adiabatic approximation [Descriptive]	292
6.1.10	Heisenberg picture [Descriptive]	293
6.2	Conservation Laws and Symmetries	296
6.3	Unsteady Perturbations of Systems	300
6.3.1	Schrödinger equation for a two-state system	301
6.3.2	Spontaneous and stimulated emission	303
6.3.3	Effect of a single wave	304
6.3.4	Forbidden transitions	307
6.3.5	Selection rules	308
6.3.6	Angular momentum conservation	309
6.3.7	Parity	312
6.3.8	Absorption of a single weak wave	314
6.3.9	Absorption of incoherent radiation	317
6.3.10	Spontaneous emission of radiation	319
6.4	Position and Linear Momentum	322
6.4.1	The position eigenfunction	322
6.4.2	The linear momentum eigenfunction	325
6.5	Wave Packets	327
6.5.1	Solution of the Schrödinger equation.	328
6.5.2	Component wave solutions	329
6.5.3	Wave packets	330
6.5.4	Group velocity	332
6.5.5	Electron motion through crystals	336
6.6	Almost Classical Motion [Descriptive]	340
6.6.1	Motion through free space	340
6.6.2	Accelerated motion	340
6.6.3	Decelerated motion	341

6.6.4	The harmonic oscillator	342
6.7	WKB Theory of Nearly Classical Motion	343
6.8	Scattering	347
6.8.1	Partial reflection	348
6.8.2	Tunneling	348
6.9	Reflection and Transmission Coefficients	350
II	Gateway Topics	353
7	Numerical Procedures	355
7.1	The Variational Method	355
7.1.1	Basic variational statement	355
7.1.2	Differential form of the statement	356
7.1.3	Example application using Lagrangian multipliers	357
7.2	The Born-Oppenheimer Approximation	359
7.2.1	The Hamiltonian	360
7.2.2	The basic Born-Oppenheimer approximation	361
7.2.3	Going one better	363
7.3	The Hartree-Fock Approximation	366
7.3.1	Wave function approximation	366
7.3.2	The Hamiltonian	372
7.3.3	The expectation value of energy	374
7.3.4	The canonical Hartree-Fock equations	376
7.3.5	Additional points	378
8	Solids	387
8.1	Molecular Solids [Descriptive]	387
8.2	Ionic Solids [Descriptive]	390
8.3	Metals [Descriptive]	394
8.3.1	Lithium	394
8.3.2	One-dimensional crystals	396
8.3.3	Wave functions of one-dimensional crystals	397
8.3.4	Analysis of the wave functions	400
8.3.5	Floquet (Bloch) theory	401
8.3.6	Fourier analysis	402
8.3.7	The reciprocal lattice	403
8.3.8	The energy levels	404
8.3.9	Merging and splitting bands	405
8.3.10	Three-dimensional metals	407
8.4	Covalent Materials [Descriptive]	411
8.5	Free-Electron Gas	414

8.5.1	Lattice for the free electrons	415
8.5.2	Occupied states and Brillouin zones	417
8.6	Nearly-Free Electrons	421
8.6.1	Energy changes due to a weak lattice potential	422
8.6.2	Discussion of the energy changes	424
8.7	Additional Points [Descriptive]	429
8.7.1	About ferromagnetism	429
8.7.2	X-ray diffraction	432
9	Basic and Quantum Thermodynamics	439
9.1	Temperature	440
9.2	Single-Particle and System Eigenfunctions	441
9.3	How Many System Eigenfunctions?	446
9.4	Particle-Energy Distribution Functions	451
9.5	The Canonical Probability Distribution	453
9.6	Low Temperature Behavior	455
9.7	The Basic Thermodynamic Variables	458
9.8	Intro to the Second Law	462
9.9	The Reversible Ideal	463
9.10	Entropy	469
9.11	The Big Lie of Distinguishable Particles	476
9.12	The New Variables	476
9.13	Microscopic Meaning of the Variables	483
9.14	Application to Particles in a Box	484
9.14.1	Bose-Einstein condensation	486
9.14.2	Fermions at low temperatures	487
9.14.3	A generalized ideal gas law	489
9.14.4	The ideal gas	489
9.14.5	Blackbody radiation	491
9.14.6	The Debye model	493
9.15	Specific Heats	494
10	Electromagnetism	501
10.1	All About Angular Momentum	501
10.1.1	The fundamental commutation relations	502
10.1.2	Ladders	503
10.1.3	Possible values of angular momentum	506
10.1.4	A warning about angular momentum	508
10.1.5	Triplet and singlet states	509
10.1.6	Clebsch-Gordan coefficients	511
10.1.7	Some important results	515
10.1.8	Momentum of partially filled shells	517

10.1.9	Pauli spin matrices	520
10.1.10	General spin matrices	523
10.2	The Relativistic Dirac Equation	524
10.3	The Electromagnetic Hamiltonian	526
10.4	Maxwell's Equations [Descriptive]	529
10.5	Example Static Electromagnetic Fields	536
10.5.1	Point charge at the origin	537
10.5.2	Dipoles	542
10.5.3	Arbitrary charge distributions	546
10.5.4	Solution of the Poisson equation	548
10.5.5	Currents	549
10.5.6	Principle of the electric motor	551
10.6	Particles in Magnetic Fields	554
10.7	Stern-Gerlach Apparatus [Descriptive]	557
10.8	Nuclear Magnetic Resonance	558
10.8.1	Description of the method	558
10.8.2	The Hamiltonian	559
10.8.3	The unperturbed system	561
10.8.4	Effect of the perturbation	563
11	Nuclei [Unfinished Draft]	567
11.1	Fundamental Concepts	568
11.2	The Simplest Nuclei	568
11.3	Overview of Nuclei	570
11.4	Magic numbers	576
11.5	Radioactivity	577
11.5.1	Decay rate	577
11.5.2	Other definitions	578
11.6	Mass and energy	579
11.7	Binding energy	581
11.8	Nucleon separation energies	583
11.9	Nuclear Forces	588
11.10	Liquid drop model	595
11.10.1	Nuclear radius	595
11.10.2	von Weizsäcker formula	596
11.10.3	Explanation of the formula	596
11.10.4	Accuracy of the formula	597
11.11	Alpha Decay	599
11.11.1	Decay mechanism	599
11.11.2	Comparison with data	602
11.11.3	Forbidden decays	604
11.11.4	Why alpha decay?	608

11.12	Shell model	610
11.12.1	Average potential	611
11.12.2	Spin-orbit interaction	617
11.12.3	Example occupation levels	621
11.12.4	Shell model with pairing	625
11.12.5	Configuration mixing	632
11.12.6	Shell model failures	638
11.13	Collective Structure	641
11.13.1	Classical liquid drop	642
11.13.2	Nuclear vibrations	644
11.13.3	Nonspherical nuclei	646
11.13.4	Rotational bands	648
11.14	Fission	661
11.14.1	Basic concepts	661
11.14.2	Some basic features	662
11.15	Spin Data	665
11.15.1	Even-even nuclei	665
11.15.2	Odd mass number nuclei	667
11.15.3	Odd-odd nuclei	670
11.16	Parity Data	674
11.16.1	Even-even nuclei	674
11.16.2	Odd mass number nuclei	674
11.16.3	Odd-odd nuclei	679
11.16.4	Parity Summary	679
11.17	Electromagnetic Moments	679
11.17.1	Classical description	682
11.17.2	Quantum description	684
11.17.3	Magnetic moment data	691
11.17.4	Quadrupole moment data	695
11.18	Isospin	699
11.19	Beta decay	704
11.19.1	Energetics Data	704
11.19.2	Von Weizsäcker approximation	711
11.19.3	Kinetic Energies	714
11.19.4	Forbidden decays	718
11.19.5	Data and Fermi theory	723
11.19.6	Parity violation	729
11.20	Gamma Decay	730
11.20.1	Energetics	731
11.20.2	Forbidden decays	732
11.20.3	Isomers	735
11.20.4	Weisskopf estimates	736

11.20.5 Internal conversion	744
12 Some Additional Topics	747
12.1 Perturbation Theory	747
12.1.1 Basic perturbation theory	747
12.1.2 Ionization energy of helium	749
12.1.3 Degenerate perturbation theory	753
12.1.4 The Zeeman effect	755
12.1.5 The Stark effect	756
12.1.6 The hydrogen atom fine structure	759
12.2 Quantum Field Theory in a Nanoshell	772
12.2.1 Occupation numbers	773
12.2.2 Annihilation and creation operators	779
12.2.3 Quantization of radiation	787
12.2.4 Spontaneous emission	794
12.2.5 Field operators	797
12.2.6 An example using field operators	798
13 The Interpretation of Quantum Mechanics	803
13.1 Schrödinger's Cat	804
13.2 Instantaneous Interactions	805
13.3 Global Symmetrization	810
13.4 Failure of the Schrödinger Equation?	810
13.5 The Many-Worlds Interpretation	813
13.6 The Arrow of Time	819
A Notes	823
A.1 Why another book on quantum mechanics?	823
A.2 History and wish list	827
A.3 Lagrangian mechanics	832
A.3.1 Introduction	832
A.3.2 Generalized coordinates	833
A.3.3 Lagrangian equations of motion	834
A.3.4 Hamiltonian dynamics	837
A.4 Special relativity	839
A.4.1 History	839
A.4.2 Overview of relativity	840
A.4.3 Lorentz transformation	843
A.4.4 Proper time and distance	845
A.4.5 Subluminal and superluminal effects	847
A.4.6 Four-vectors	848
A.4.7 Index notation	849

A.4.8	Group property	851
A.4.9	Intro to relativistic mechanics	852
A.4.10	Lagrangian mechanics	856
A.5	Completeness of Fourier modes	859
A.6	Derivation of the Euler formula	863
A.7	Nature and real eigenvalues	863
A.8	Are Hermitian operators really like that?	864
A.9	Are linear momentum operators Hermitian?	864
A.10	Why boundary conditions are tricky	864
A.11	Extension to three-dimensional solutions	865
A.12	Derivation of the harmonic oscillator solution	867
A.13	More on the harmonic oscillator and uncertainty	870
A.14	Derivation of a vector identity	871
A.15	Derivation of the spherical harmonics	871
A.16	The reduced mass	874
A.17	The hydrogen radial wave functions	877
A.18	Inner product for the expectation value	880
A.19	Why commuting operators have common eigenvectors	880
A.20	The generalized uncertainty relationship	881
A.21	Derivation of the commutator rules	882
A.22	Is the variational approximation best?	884
A.23	Solution of the hydrogen molecular ion	885
A.24	Accuracy of the variational method	886
A.25	Positive molecular ion wave function	887
A.26	Molecular ion wave function symmetries	888
A.27	Solution of the hydrogen molecule	889
A.28	Hydrogen molecule ground state and spin	890
A.29	Number of boson states	891
A.30	Shielding approximation limitations	892
A.31	Why the s states have the least energy	893
A.32	Density of states	893
A.33	Radiation from a hole	896
A.34	Kirchhoff's law	897
A.35	The thermionic emission equation	898
A.36	Explanation of the band gaps	900
A.37	Number of conduction band electrons	906
A.38	Thermoelectric effects	906
A.38.1	Peltier and Seebeck coefficient ballpark	906
A.38.2	Figure of merit	908
A.38.3	Physical Seebeck mechanism	909
A.38.4	Full thermoelectric equations	910
A.38.5	Charge locations in thermoelectrics	913

A.38.6 Kelvin relationships	914
A.39 Why energy eigenstates are stationary	918
A.40 Better description of two-state systems	919
A.41 The evolution of expectation values	919
A.42 The virial theorem	919
A.43 The energy-time uncertainty relationship	920
A.44 The adiabatic theorem	921
A.44.1 Derivation of the theorem	921
A.44.2 Some implications	924
A.45 Symmetry eigenvalue conservation	925
A.46 The two-state approximation of radiation	925
A.47 Selection rules	926
A.48 About spectral broadening	930
A.49 Derivation of the Einstein B coefficients	931
A.50 Parseval and the Fourier inversion theorem	934
A.51 Derivation of group velocity	935
A.52 Motion through crystals	938
A.52.1 Propagation speed	938
A.52.2 Motion under an external force	938
A.52.3 Free-electron gas with constant electric field	940
A.53 Details of the animations	941
A.54 Derivation of the WKB approximation	949
A.55 WKB solution near the turning points	951
A.56 Three-dimensional scattering	955
A.56.1 Partial wave analysis	957
A.56.2 The Born approximation	961
A.56.3 The Born series	964
A.57 The evolution of probability	965
A.58 A basic description of Lagrangian multipliers	969
A.59 The generalized variational principle	970
A.60 Spin degeneracy	972
A.61 Derivation of the approximation	972
A.62 Why a single Slater determinant is not exact	977
A.63 Simplification of the Hartree-Fock energy	978
A.64 Integral constraints	983
A.65 Generalized orbitals	984
A.66 Derivation of the Hartree-Fock equations	985
A.67 Why the Fock operator is Hermitian	992
A.68 “Correlation energy”	993
A.69 Explanation of the London forces	996
A.70 Ambiguities in the definition of electron affinity	1000
A.71 Why Floquet theory should be called so	1002

A.72	Superfluidity versus BEC	1002
A.73	Explanation of Hund's first rule	1004
A.74	The mechanism of ferromagnetism	1006
A.75	Number of system eigenfunctions	1007
A.76	The fundamental assumption of quantum statistics	1011
A.77	A problem if the energy is given	1012
A.78	Derivation of the particle energy distributions	1013
A.79	The canonical probability distribution	1019
A.80	Analysis of the ideal gas Carnot cycle	1021
A.81	The recipe of life	1022
A.82	The third law	1023
A.83	Checks on the expression for entropy	1025
A.84	Chemical potential and distribution functions	1028
A.85	Fermi-Dirac integrals at low temperature	1032
A.86	Physics of the fundamental commutation relations	1034
A.87	Multiple angular momentum components	1035
A.88	Components of vectors are less than the total vector	1035
A.89	The spherical harmonics with ladder operators	1036
A.90	Why angular momenta components can be added	1036
A.91	Why the Clebsch-Gordan tables are bidirectional	1037
A.92	How to make Clebsch-Gordan tables	1037
A.93	Machine language version of the Clebsch-Gordan tables	1037
A.94	The triangle inequality	1038
A.95	Momentum of shells	1039
A.96	Awkward questions about spin	1041
A.97	More awkwardness about spin	1043
A.98	Emergence of spin from relativity	1043
A.99	Electromagnetic evolution of expectation values	1046
A.100	Existence of magnetic monopoles	1048
A.101	More on Maxwell's third law	1048
A.102	Various electrostatic derivations.	1049
A.102.1	Existence of a potential	1049
A.102.2	The Laplace equation	1050
A.102.3	Egg-shaped dipole field lines	1051
A.102.4	Ideal charge dipole delta function	1051
A.102.5	Integrals of the current density	1052
A.102.6	Lorentz forces on a current distribution	1053
A.102.7	Field of a current dipole	1054
A.102.8	Biot-Savart law	1056
A.103	Energy due to orbital motion in a magnetic field	1057
A.104	Energy due to electron spin in a magnetic field	1058
A.105	Setting the record straight on alignment	1059

A.106 Solving the NMR equations	1060
A.107 Harmonic oscillator revisited	1060
A.108 Impenetrable spherical shell	1062
A.109 Classical vibrating drop	1062
A.109.1 Basic definitions	1062
A.109.2 Kinetic energy	1063
A.109.3 Energy due to surface tension	1066
A.109.4 Energy due to Coulomb repulsion	1069
A.109.5 Frequency of vibration	1071
A.110 Shell model quadrupole moment	1071
A.111 Fermi theory	1072
A.111.1 Form of the wave function	1073
A.111.2 Source of the decay	1075
A.111.3 Allowed or forbidden	1079
A.111.4 The nuclear operator	1081
A.111.5 Fermi's golden rule	1084
A.111.6 Mopping up	1088
A.111.7 Electron capture	1093
A.112 Weisskopf estimates	1094
A.112.1 Very loose derivation	1094
A.112.2 Official loose derivation	1099
A.113 Auger discovery	1100
A.114 Derivation of perturbation theory	1100
A.115 Hydrogen ground state Stark effect	1105
A.116 Dirac fine structure Hamiltonian	1107
A.117 Classical spin-orbit derivation	1114
A.118 Expectation powers of r for hydrogen	1117
A.119 A tenth of a googol in universes	1121
Web Pages	1127
Notations	1131

List of Figures

1.1	The classical picture of a vector.	6
1.2	Spike diagram of a vector.	7
1.3	More dimensions.	7
1.4	Infinite dimensions.	7
1.5	The classical picture of a function.	8
1.6	Forming the dot product of two vectors.	9
1.7	Forming the inner product of two functions.	10
1.8	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 <i>not</i> 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	14
2.1	A visualization of an arbitrary wave function.	21
2.2	Combined plot of position and momentum components.	24
2.3	The uncertainty principle illustrated.	24
2.4	Classical picture of a particle in a closed pipe.	31
2.5	Quantum mechanics picture of a particle in a closed pipe.	31
2.6	Definitions for one-dimensional motion in a pipe.	32
2.7	One-dimensional energy spectrum for a particle in a pipe.	38
2.8	One-dimensional ground state of a particle in a pipe.	40
2.9	Second and third lowest one-dimensional energy states.	41
2.10	Definition of all variables for motion in a pipe.	42
2.11	True ground state of a particle in a pipe.	43
2.12	True second and third lowest energy states.	44
2.13	A combination of ψ_{111} and ψ_{211} seen at some typical times.	46
2.14	The harmonic oscillator.	48
2.15	The energy spectrum of the harmonic oscillator.	53
2.16	Ground state of the harmonic oscillator	55
2.17	Wave functions ψ_{100} and ψ_{010}	56

2.18 Energy eigenfunction ψ_{213} .	57
2.19 Arbitrary wave function (not an energy eigenfunction).	60
3.1 Spherical coordinates of an arbitrary point P.	65
3.2 Spectrum of the hydrogen atom.	78
3.3 Ground state wave function of the hydrogen atom.	81
3.4 Eigenfunction ψ_{200} .	82
3.5 Eigenfunction ψ_{210} , or $2p_z$.	83
3.6 Eigenfunction ψ_{211} (and ψ_{21-1}).	83
3.7 Eigenfunctions $2p_x$, left, and $2p_y$, right.	84
3.8 Hydrogen atom plus free proton far apart.	102
3.9 Hydrogen atom plus free proton closer together.	102
3.10 The electron being anti-symmetrically shared.	104
3.11 The electron being symmetrically shared.	105
4.1 State with two neutral atoms.	117
4.2 Symmetric sharing of the electrons.	119
4.3 Antisymmetric sharing of the electrons.	119
4.4 Approximate solutions for hydrogen (left) and helium (right) atoms.	154
4.5 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.	156
4.6 Approximate solutions for lithium (left) and beryllium (right).	157
4.7 Example approximate solution for boron.	158
4.8 Periodic table of the elements.	162
4.9 Covalent sigma bond consisting of two $2p_z$ states.	166
4.10 Covalent pi bond consisting of two $2p_x$ states.	167
4.11 Covalent sigma bond consisting of a $2p_z$ and a $1s$ state.	168
4.12 Shape of an sp^3 hybrid state.	170
4.13 Shapes of the sp^2 (left) and sp (right) hybrids.	171
5.1 Allowed wave number vectors, left, and energy spectrum, right.	179
5.2 Ground state of a system of noninteracting bosons in a box.	183
5.3 The system of bosons at a very low temperature.	187
5.4 The system of bosons at a relatively low temperature.	187

5.5	Ground state system energy eigenfunction for a simple model system with only 3 single-particle energy levels, 6 single-particle states, and 3 distinguishable spinless particles. Left: mathematical form. Right: graphical representation. All three particles are in the single-particle ground state.	190
5.6	Example system energy eigenfunction with five times the single-particle ground state energy.	190
5.7	For distinguishable particles, there are 9 system energy eigenfunctions that have energy distribution A.	191
5.8	For distinguishable particles, there are 12 system energy eigenfunctions that have energy distribution B.	192
5.9	For identical bosons, there are only 3 system energy eigenfunctions that have energy distribution A.	193
5.10	For identical bosons, there are also only 3 system energy eigenfunctions that have energy distribution B.	193
5.11	Ground state of a system of noninteracting electrons, or other fermions, in a box.	201
5.12	Severe confinement in the y -direction, as in a quantum well. . . .	207
5.13	Severe confinement in both the y - and z -directions, as in a quantum wire.	208
5.14	Severe confinement in all three directions, as in a quantum dot or artificial atom.	209
5.15	A system of fermions at a nonzero temperature.	211
5.16	Particles at high-enough temperature and low-enough particle density.	215
5.17	Ground state of a system of noninteracting electrons, or other fermions, in a periodic box.	223
5.18	Conduction in the free-electron gas model.	225
5.19	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.	230
5.20	Sketch of electron energy spectra in solids at a nonzero temperature.	235
5.21	Potential energy seen by an electron along a line of nuclei. The potential energy is in green, the nuclei are in red.	237
5.22	Potential energy seen by an electron in the one-dimensional simplified model of Kronig & Penney.	238

5.23 Example Kronig & Penney spectra.	240
5.24 Spectrum against wave number in the extended zone scheme.	243
5.25 Spectrum against wave number in the reduced zone scheme.	244
5.26 Spectrum against wave number in the periodic zone scheme.	246
5.27 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.	247
5.28 Relationship between conduction electron density and hole density. Intrinsic semiconductors have neither much conduction electrons nor holes.	252
5.29 The <i>p-n</i> 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 as this greatly simplified schematic indicates).	254
5.30 Schematic of the operation of an <i>p-n</i> junction.	257
5.31 Schematic of the operation of an <i>n-p-n</i> transistor.	259
5.32 Vicinity of the band gap in the 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.	263
5.33 Peltier cooling. Top: physical device. Bottom: Electron energy spectra of the semiconductor materials. Quantum states filled with electrons are shown in red.	269
5.34 Seebeck voltage generator.	273
5.35 The Galvani potential jump over the contact surface does not produce a usable voltage.	275
5.36 The Seebeck effect is not directly measurable.	276
 6.1 The ground state wave function looks the same at all times.	282
6.2 The first excited state at all times.	283
6.3 A combination of ψ_{111} and ψ_{211} seen at some typical times.	283
6.4 Emission and absorption of radiation by an atom.	303
6.5 Triangle inequality.	310
6.6 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}	323

6.7	The real part (red) and envelope (black) of an example wave	329
6.8	The wave moves with the phase speed.	330
6.9	The real part (red) and magnitude or envelope (black) of a wave packet. (Schematic).	331
6.10	The velocities of wave and envelope are not equal.	332
6.11	A particle in free space.	341
6.12	An accelerating particle.	341
6.13	An decelerating particle.	342
6.14	Unsteady solution for the harmonic oscillator. The third picture shows the maximum distance from the nominal position that the wave packet reaches.	343
6.15	Harmonic oscillator potential energy V , eigenfunction h_{50} , and its energy E_{50}	343
6.16	A partial reflection.	348
6.17	An tunneling particle.	349
6.18	Penetration of an infinitely high potential energy barrier.	349
6.19	Schematic of a scattering potential and the asymptotic behavior of an example energy eigenfunction for a wave packet coming in from the far left.	351
8.1	Billiard-ball model of the salt molecule.	390
8.2	Billiard-ball model of a salt crystal.	392
8.3	The salt crystal disassembled to show its structure.	393
8.4	The lithium atom, scaled more correctly than in chapter 4.9	395
8.5	Body-centered-cubic (bcc) structure of lithium.	396
8.6	Fully periodic wave function of a two-atom lithium “crystal.”	397
8.7	Flip-flop wave function of a two-atom lithium “crystal.”	398
8.8	Wave functions of a four-atom lithium “crystal.” The actual picture is that of the fully periodic mode.	399
8.9	Reciprocal lattice of a one-dimensional crystal.	403
8.10	Schematic of energy bands.	404
8.11	Schematic of merging bands.	406
8.12	A primitive cell and primitive translation vectors of lithium.	407
8.13	Wigner-Seitz cell of the bcc lattice.	408
8.14	Schematic of crossing bands.	412
8.15	Ball and stick schematic of the diamond crystal.	413
8.16	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.	416
8.17	Occupied states for one, two, and three free electrons per physical lattice cell.	418

8.18 Redefinition of the occupied wave number vectors into Brillouin zones.	419
8.19 Second, third, and fourth Brillouin zones seen in the periodic zone scheme.	420
8.20 The red dot shows the wavenumber vector of a sample free electron wave function. It is to be corrected for the lattice potential.	422
8.21 The grid of nonzero Hamiltonian perturbation coefficients and the problem sphere in wave number space.	423
8.22 Tearing apart of the wave number space energies.	425
8.23 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.	426
8.24 Bragg planes seen in wave number space cross section.	426
8.25 Occupied states for the energies of figure 8.23 if there are two valence electrons per lattice cell. Left: energy. Right: wave numbers.	427
8.26 Smaller lattice potential. From top to bottom shows one, two and three valence electrons per lattice cell. Left: energy. Right: wave numbers.	428
8.27 Depiction of an electromagnetic ray.	433
8.28 Law of reflection in elastic scattering from a plane.	434
8.29 Scattering from multiple “planes of atoms”.	435
8.30 Difference in travel distance when scattered from P rather than O.	436
9.1 Graphical depiction of an arbitrary system energy eigenfunction for 36 distinguishable particles.	443
9.2 Graphical depiction of an arbitrary system energy eigenfunction for 36 identical bosons.	445
9.3 Graphical depiction of an arbitrary system energy eigenfunction for 33 identical fermions.	445
9.4 Illustrative small model system having 4 distinguishable particles. The particular eigenfunction shown is arbitrary.	448
9.5 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.	448

9.6	Number of energy eigenfunctions on the oblique energy line in 9.5. (The curves are mathematically interpolated to allow a continuously varying fraction of particles on shelf 2.) Left: 4 particles, middle: 64, right: 1,024.	450
9.7	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 figure 9.5.	455
9.8	Probabilities of shelf-number sets for the simple 64 particle model system if shelf 1 is a non-degenerate ground state. Left: identical bosons, middle: distinguishable particles, right: identical fermions. The temperature is the same as in figure 9.7.	456
9.9	Like figure 9.8, but at a lower temperature.	456
9.10	Like figure 9.8, but at a still lower temperature.	457
9.11	Schematic of the Carnot refrigeration cycle.	464
9.12	Schematic of the Carnot heat engine.	467
9.13	A generic heat pump next to a reversed Carnot one with the same heat delivery.	468
9.14	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.	470
9.15	Specific heat at constant volume of gases. Temperatures from absolute zero to 1,200 K. Data from NIST-JANAF and AIP.	495
9.16	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.	497
10.1	Example bosonic ladders.	505
10.2	Example fermionic ladders.	505
10.3	Triplet and singlet states in terms of ladders	511
10.4	Clebsch-Gordan coefficients of two spin one half particles.	512
10.5	Clebsch-Gordan coefficients for a spin-one-half second particle.	514
10.6	Clebsch-Gordan coefficients for a spin-one second particle.	516
10.7	Relationship of Maxwell's first equation to Coulomb's law.	530
10.8	Maxwell's first equation for a more arbitrary region. The figure to the right includes the field lines through the selected points.	531
10.9	The net number of field lines leaving a region is a measure for the net charge inside that region.	532
10.10	Since magnetic monopoles do not exist, the net number of magnetic field lines leaving a region is always zero.	533

10.11 Electric power generation.	534
10.12 Two ways to generate a magnetic field: using a current (left) or using a varying electric field (right).	535
10.13 Electric field and potential of a charge that is distributed uni- formly within a small sphere. The dotted lines indicate the values for a point charge.	540
10.14 Electric field of a two-dimensional line charge.	541
10.15 Field lines of a vertical electric dipole.	542
10.16 Electric field of a two-dimensional dipole.	543
10.17 Field of an ideal magnetic dipole.	544
10.18 Electric field of an almost ideal two-dimensional dipole.	545
10.19 Magnetic field lines around an infinite straight electric wire. . .	550
10.20 An electromagnet consisting of a single wire loop. The generated magnetic field lines are in blue.	550
10.21 A current dipole.	551
10.22 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.	552
10.23 Variables for the computation of the moment on a wire loop in a magnetic field.	553
10.24 Larmor precession of the expectation spin (or magnetic moment) vector around the magnetic field.	563
10.25 Probability of being able to find the nuclei at elevated energy versus time for a given perturbation frequency ω	564
10.26 Maximum probability of finding the nuclei at elevated energy. .	564
10.27 A perturbing magnetic field, rotating at precisely the Larmor frequency, causes the expectation spin vector to come cascading down out of the ground state.	565
11.1 Nuclear decay modes.	572
11.2 Binding energy per nucleon.	582
11.3 Proton separation energy.	584
11.4 Neutron separation energy.	585
11.5 Proton pair separation energy.	586
11.6 Neutron pair separation energy.	587
11.7 Error in the von Weizsäcker formula.	598

11.8 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. Inset: really cherry-picking, only a few even mass numbers for thorium and uranium! Bottom: all the nuclei except one.	600
11.9 Schematic potential for an alpha particle that tunnels out of a nucleus.	601
11.10 Half-life predicted by the Gamow / Gurney & Condon theory versus the true value. Top: even-even nuclei only. Bottom: all the nuclei except one.	605
11.11 Example average nuclear potentials: (a) harmonic oscillator, (b) impenetrable surface, (c) Woods-Saxon, (d) Woods-Saxon for protons.	612
11.12 Nuclear energy levels for various assumptions about the average nuclear potential. The signs indicate the parity of the states.	615
11.13 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).	620
11.14 Energy levels for doubly-magic oxygen-16 and neighbors.	622
11.15 Nucleon pairing effect.	626
11.16 Energy levels for neighbors of doubly-magic calcium-40.	631
11.17 2^+ excitation energy of even-even nuclei.	634
11.18 Collective motion effects.	636
11.19 Failures of the shell model.	639
11.20 An excitation energy ratio for even-even nuclei.	647
11.21 Textbook vibrating nucleus tellurium-120.	648
11.22 Rotational bands of hafnium-177.	650
11.23 Ground state rotational band of tungsten-183.	655
11.24 Rotational bands of aluminum-25.	656
11.25 Rotational bands of erbium-164.	657
11.26 Ground state rotational band of magnesium-24.	658
11.27 Rotational bands of osmium-190.	660
11.28 Simplified energetics for fission of fermium-256.	664
11.29 Spin of even-even nuclei.	666
11.30 Spin of even-odd nuclei.	668
11.31 Spin of odd-even nuclei.	669
11.32 Spin of odd-odd nuclei.	671
11.33 Selected odd-odd spins predicted using the neighbors.	673
11.34 Selected odd-odd spins predicted from theory.	675
11.35 Parity of even-even nuclei.	676

11.36 Parity of even-odd nuclei.	677
11.37 Parity of odd-even nuclei.	678
11.38 Parity of odd-odd nuclei.	680
11.39 Parity versus the shell model.	681
11.40 Magnetic dipole moments of the ground-state nuclei.	693
11.41 2^+ magnetic moment of even-even nuclei.	694
11.42 Electric quadrupole moment.	696
11.43 Electric quadrupole moment corrected for spin.	698
11.44 Isobaric analog states.	702
11.45 Energy release in beta decay of even-odd nuclei.	706
11.46 Energy release in beta decay of odd-even nuclei.	707
11.47 Energy release in beta decay of odd-odd nuclei.	708
11.48 Energy release in beta decay of even-even nuclei.	709
11.49 Examples of beta decay.	712
11.50 The Fermi integral. It shows the effects of energy release and nuclear charge on the beta decay rate of allowed transitions. Other effects exists.	722
11.51 Beta decay rates.	724
11.52 Beta decay rates as fraction of a ballparked value.	725
11.53 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.	729
11.54 Energy levels of tantalum-180.	734
11.55 Half-life of the longest-lived even-odd isomers.	737
11.56 Half-life of the longest-lived odd-even isomers.	738
11.57 Half-life of the longest-lived odd-odd isomers.	739
11.58 Half-life of the longest-lived even-even isomers.	740
11.59 Weisskopf ballpark half-lifes for electromagnetic transitions versus energy release. Broken lines include ballparked internal conversion.	742
11.60 Moszkowski ballpark half-lifes for magnetic transitions versus energy release. Broken lines include ballparked internal conversion.	743
12.1 Graphical depiction of an arbitrary system energy eigenfunction for 36 distinguishable particles.	774
12.2 Graphical depiction of an arbitrary system energy eigenfunction for 36 identical bosons.	774
12.3 Graphical depiction of an arbitrary system energy eigenfunction for 33 identical fermions.	775
12.4 Example wave functions for a system with just one type of single particle state. Left: identical bosons; right: identical fermions.	777

12.5 Annihilation and creation operators for a system with just one type of single particle state. Left: identical bosons; right: identical fermions.	780
13.1 Separating the hydrogen ion.	805
13.2 The Bohm experiment before the Venus measurement (left), and immediately after it (right).	806
13.3 Spin measurement directions.	807
13.4 Earth's view of events (left), and that of a moving observer (right).	809
13.5 Bohm's version of the Einstein, Podolski, Rosen Paradox.	813
13.6 Non entangled positron and electron spins; up and down.	814
13.7 Non entangled positron and electron spins; down and up.	814
13.8 The wave functions of two universes combined	814
13.9 The Bohm experiment repeated.	817
13.10 Repeated experiments on the same electron.	818
A.1 Coordinate systems for the Lorentz transformation.	843
A.2 Example elastic collision seen by different observers.	853
A.3 A completely inelastic collision.	855
A.4 Bosons in single-particle-state boxes.	892
A.5 Spectrum for a weak potential.	901
A.6 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.	902
A.7 Analysis of conduction.	914
A.8 Example energy eigenfunction for the particle in free space.	942
A.9 Example energy eigenfunction for a particle entering a constant accelerating force field.	942
A.10 Example energy eigenfunction for a particle entering a constant decelerating force field.	944
A.11 Example energy eigenfunction for the harmonic oscillator.	945
A.12 Example energy eigenfunction for a particle encountering a brief accelerating force.	945
A.13 Example energy eigenfunction for a particle tunneling through a barrier.	946
A.14 Example energy eigenfunction for tunneling through a delta function barrier.	946
A.15 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.	951

A.16 Connection formulae for a turning point from classical to tunneling.	953
A.17 Connection formulae for a turning point from tunneling to classical.	953
A.18 WKB approximation of tunneling.	954
A.19 Scattering of a beam off a target.	955
A.20 Graphical interpretation of the Born series.	964
A.21 Possible polarizations of a pair of hydrogen atoms.	997
A.22 Schematic of an example boson distribution on a shelf.	1010
A.23 Schematic of the Carnot refrigeration cycle.	1022
A.24 Energy slop diagram.	1086
A.25 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.	1089

List of Tables

2.1	First few one-dimensional eigenfunctions of the harmonic oscillator.	51
3.1	The first few spherical harmonics.	68
3.2	The first few spherical harmonics rewritten.	70
3.3	The first few radial wave functions for hydrogen.	76
5.1	Energy of the lowest single-particle state in a cube with 1 cm sides.	180
10.1	Possible combined angular momentum of identical fermions in shells of single-particle states that differ in magnetic quantum number.	518
10.2	Possible combined angular momentum of identical bosons. . . .	521
10.3	Elecromagnetics I: Fundamental equations and basic solutions. .	538
10.4	Elecromagnetics II: Electromagnetostatic solutions.	539
11.1	Alternate names for nuclei.	575
11.2	Candidates for nuclei ejected by uranium-238, radium-223, and fermium-256.	609
A.1	Additional combined angular momentum values.	1040

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.

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.

Some sections are marked [descriptive]. These sections do not provide new analytical techniques. Instead they describe important ideas and conclusions that follow from the quantum mechanics. Read through these sections more than once, so that you have a good idea of what they are all about. Do not worry too much about the details of any illustrative analysis that might be there.

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 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, {A.1} A version history is in note {A.2}.

Acknowledgments

This book is for a large part based on my reading of the excellent book by Griffiths, [17]. 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 numerical methods, thermodynamics, solid mechanics, and electromagnetism.

Somewhat to my surprise, I find that my coverage actually tends to be closer to Yariv's book, [34]. I still think Griffiths is more readable for an engineer, though Yariv has some very good items that Griffiths does not.

The discussions on two-state systems are mainly based on Feynman's notes, [15, 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.

The discussion of the Onsager theorem comes from Desloge, [9], an emeritus professor of physics at the Florida State University.

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 Johann Joss for pointing out a mistake in the formula for the averaged energy of two-state systems. Harald Kirsch reported various problems in the sections on conservation laws and on position eigenfunctions.

The note on the derivation of the selection rules is from [17] and lecture notes from a University of Tennessee quantum course taught by Marianne Breinig. The subsection on conservation laws and selection rules is mainly from Ellis, [10].

The section on the Born-Oppenheimer approximation comes from Wikipedia, [[18]], with modifications including the inclusion of spin.

The section on the Hartree-Fock method is mainly based on Szabo and Ostlund [32], a well-written book, with some Parr and Yang [22] thrown in.

The section on solids is mainly based on Sproull, [29], 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 [19]. The discussion of ionic solids really comes straight from hyperphysics [[5]]. 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, [17], which follows Yariv's, [34]. However, it expanded greatly during writing. It now comes mostly from Baierlein [4], with some help from Feynman, [13], and some of the books I use in undergraduate thermo.

The derivation of the classical energy of a spinning particle in a magnetic field is from Yariv, [34].

The initial inspiration for the chapter on nuclear physics was the Nobel Prize

acceptance lecture of Goeppert Mayer [[8]]. This is an excellent introduction to nuclear physics for a non-specialist audience. It is freely available on the web. As the chapter expanded, the main references became the popular book by Krane [21], as well as [23] and [27]. 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 [31]. Nu-Dat 2 [[11]] provided the the excited energy levels and additional reference data to validate various data in [31].

The brief description of quantum field theory is mostly from Wikipedia, with a bit of fill-in from Feynman [13] and Kittel [19]. The example on field operators is an exercise from Srednicki [30], whose solution was posted online by a TA of Joe Polchinski from UCSB.

The many-worlds discussion is based on Everett's exposition, [11]. It is brilliant but quite impenetrable.

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

Acknowledgements 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, and Mark Vanderlaan. 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, ME 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 among the half million or so 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

Basic Quantum Mechanics

Chapter 1

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 $i = \sqrt{-1}$ 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.

1.1 Complex Numbers

Quantum mechanics is full of complex numbers, numbers involving

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

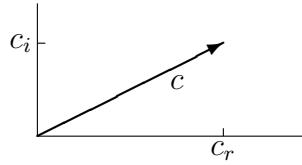
Note that $\sqrt{-1}$ is not an ordinary, “real”, number, since there is no real number whose square is -1 ; the square of a real number is always positive. This section summarizes the most important properties of complex numbers.

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

$$c = c_r + i c_i \quad (1.1)$$

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:

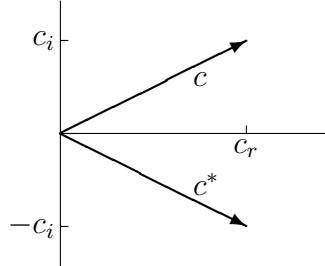
$$\frac{1}{i} = -i \quad (1.2)$$

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

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

$$c^* = c_r - i c_i \quad (1.3)$$

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:

$$|c| = \sqrt{c^*c} \quad (1.4)$$

If $c = c_r + ic_i$, 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

$$c = |c| (\cos \alpha + i \sin \alpha)$$

The critically important Euler formula says that:

$$\cos \alpha + i \sin \alpha = e^{i\alpha} \quad (1.5)$$

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

$$c = |c|e^{i\alpha} \quad (1.6)$$

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 $e^{i\alpha}$. 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 $-i$ for $\alpha = -\pi/2$.

(See note {A.6} 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.

- The Euler formula relates exponentials to sines and cosines.
-
-

1.1 Review Questions

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

1.2 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:

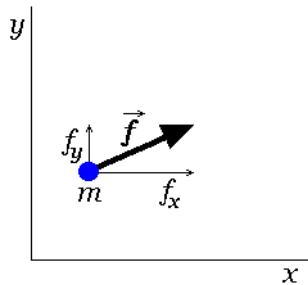


Figure 1.1: The classical picture of a vector.

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

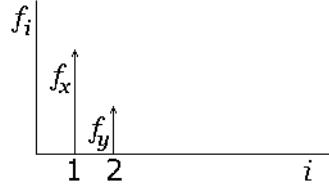


Figure 1.2: Spike diagram of a vector.

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

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:

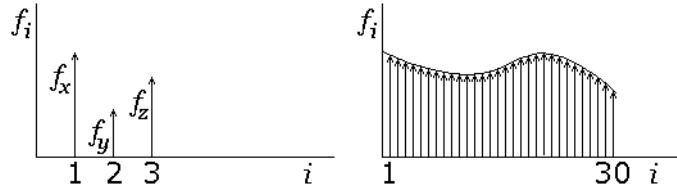


Figure 1.3: More dimensions.

For a large number of dimensions, and in particular in the limit of infinitely many dimensions, 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 $f(x)$:



Figure 1.4: Infinite dimensions.

The spikes are usually not shown:

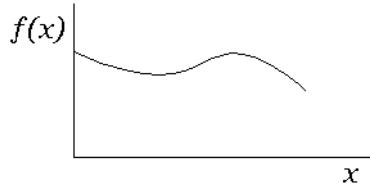


Figure 1.5: The classical picture of a function.

In this way, a function is just a vector in infinitely many dimensions.

Key Points

- □ Functions can be thought of as vectors with infinitely many components.
- □ This allows quantum mechanics do the same things with functions as you can do with vectors.

1.2 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 $f(x) = 0.5x$.
- 2 Graphically compare the spike diagram of the 10-dimensional unit vector \hat{i}_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.)

1.3 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 1.6 shows multiplied components using equal colors.

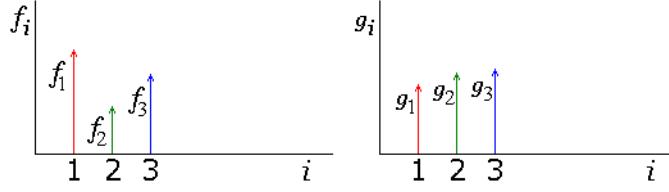


Figure 1.6: Forming the dot product of two vectors.

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, $i^2 = -1$.

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

$$\langle \vec{f} | \vec{g} \rangle \equiv \sum_{\text{all } i} f_i^* g_i \quad (1.7)$$

If the vector \vec{f} is real, the complex conjugate does nothing, and the inner product $\langle \vec{f} | \vec{g} \rangle$ 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:

$$|\vec{f}| = \sqrt{\langle \vec{f} | \vec{f} \rangle} = \sqrt{\sum_{\text{all } i} |f_i|^2} \quad (1.8)$$

Physicists take the inner product “bracket” verbally apart as

$$\begin{array}{c} \langle \vec{f} | \quad | \vec{g} \rangle \\ \text{bra} \quad \not\in \text{ket} \end{array}$$

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:

$$\langle f|g \rangle = \int_{\text{all } x} f^*(x)g(x) dx \quad (1.9)$$

as illustrated in figure 1.7.

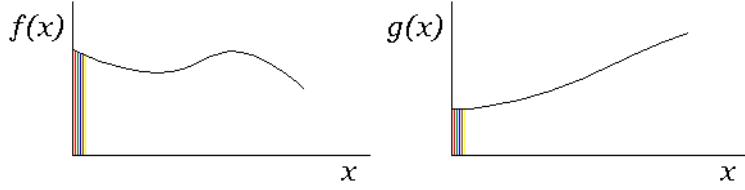


Figure 1.7: Forming the inner product of two functions.

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

$$\|f\| \equiv \sqrt{\langle f|f \rangle} = \sqrt{\int_{\text{all } x} |f(x)|^2 dx} \quad (1.10)$$

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:

$$\langle f|f \rangle = 1 \text{ iff } f \text{ is normalized.} \quad (1.11)$$

(“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:

$$\langle f|g \rangle = 0 \text{ iff } f \text{ and } g \text{ are orthogonal.} \quad (1.12)$$

- Sets of vectors or functions that are all mutually orthogonal, and 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, \dots is orthonormal if

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

and

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

Key Points

- For complex vectors and functions, the normal dot product becomes the inner product.
- To take an inner product of vectors, (1) take complex conjugates of the components of the first vector; (2) multiply corresponding components of the two vectors together; and (3) sum these products.
- To take an inner product of functions, (1) take the complex conjugate of the first function; (2) multiply the two functions; and (3) integrate the product function. The real difference from vectors is integration instead of summation.
- 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 functions, or a pair of vectors, is orthogonal if their inner product is zero.
- A set of functions, or a set of vectors, forms an orthonormal set if every one is orthogonal to all the rest, and every one is of unit norm or length.

1.3 Review Questions

1 Find the following inner product of the two vectors:

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

2 Find the length of the vector

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

3 Find the inner product of the functions $\sin(x)$ and $\cos(x)$ on the interval $0 \leq x \leq 1$.

4 Show that the functions $\sin(x)$ and $\cos(x)$ are orthogonal on the interval $0 \leq x \leq 2\pi$.

5 Verify that $\sin(x)$ is not a normalized function on the interval $0 \leq x \leq 2\pi$, and normalize it by dividing by its norm.

- 6** Verify that the most general multiple of $\sin(x)$ that is normalized on the interval $0 \leq x \leq 2\pi$ is $e^{i\alpha} \sin(x)/\sqrt{\pi}$ 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 $i\sin(x)/\sqrt{\pi}$, (for $\alpha = \pi/2$).
- 7** Show that the functions $e^{4i\pi x}$ and $e^{6i\pi x}$ are an orthonormal set on the interval $0 \leq x \leq 1$.
-

1.4 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:

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

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

Note that a hat is often used to indicate operators; for example, \hat{x} is the symbol for the operator that corresponds to multiplying by x . If it is clear that something is an operator, such as d/dx , 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 f by a number, Af increases by that same number; also, if you sum f and g , $A(f + g)$ will be Af plus Ag .

Key Points

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

1.4 Review Questions

- 1 So what is the result if the operator d/dx is applied to the function $\sin(x)$?
 - 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 ?
 - 3 A less self-evident operator than the above examples is a translation operator like $T_{\pi/2}$ that translates the graph of a function towards the left by an amount $\pi/2$: $T_{\pi/2}f(x) = f\left(x + \frac{1}{2}\pi\right)$. (Curiously enough, translation operators turn out to be responsible for the law of conservation of momentum.) Show that $T\pi/2$ turns $\sin(x)$ into $\cos(x)$.
 - 4 The inversion operator Inv 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 Inv leaves $\cos(x)$ unchanged, but turns $\sin(x)$ into $-\sin(x)$.
-

1.5 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:

$$A\vec{v} = a\vec{v} \text{ iff } \vec{v} \text{ is an eigenvector of } A \quad (1.13)$$

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:

$$Af = af \text{ iff } f \text{ is an eigenfunction of } A. \quad (1.14)$$

For example, e^x is an eigenfunction of the operator d/dx with eigenvalue 1, since $de^x/dx = 1e^x$. Another simple example is illustrated in figure 1.8; the function $\sin(2x)$ is *not* an eigenfunction of the first derivative operator d/dx . However it *is* an eigenfunction of the second derivative operator d^2/dx^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 d/dx that do

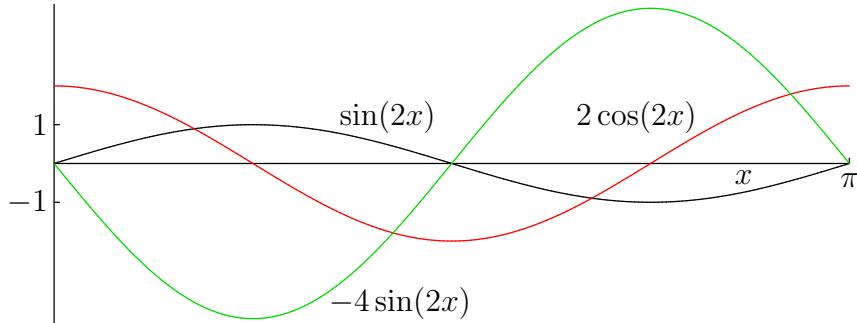


Figure 1.8: Illustration of the eigenfunction concept. Function $\sin(2x)$ is shown in black. Its first derivative $2\cos(2x)$, shown in red, is not just a multiple of $\sin(2x)$. Therefore $\sin(2x)$ is *not* an eigenfunction of the first derivative operator. However, the second derivative of $\sin(2x)$ 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 .

appear a lot are of the form e^{ikx} , where $i = \sqrt{-1}$ and k is an arbitrary real number. The eigenvalue is ik :

$$\frac{d}{dx} e^{ikx} = ike^{ikx}$$

Function e^{ikx} does not blow up at large x ; in particular, the Euler formula (1.5) says:

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

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.

1.5 Review Questions

- 1 Show that e^{ikx} , above, is also an eigenfunction of d^2/dx^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.

- 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 d^2/dx^2 , though they are not eigenfunctions of d/dx
- 3** Show that $\sin(kx)$ and $\cos(kx)$, with k a constant, are eigenfunctions of the inversion operator Inv , which turns any function $f(x)$ into $f(-x)$, and find the eigenvalues.
-

1.6 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:

$$\langle f | Ag \rangle = \langle Af | g \rangle \text{ always iff } A \text{ is Hermitian.} \quad (1.15)$$

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

- They always have real eigenvalues, not involving $i = \sqrt{-1}$. (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 {A.7}.
- 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 note {A.5} for an important example. But see also {A.8}.) 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{i} , \hat{j} , 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:

$$\text{If } A \text{ is Hermitian: } \langle g | Af \rangle = \langle f | Ag \rangle^*, \quad \langle f | Af \rangle \text{ is real.} \quad (1.16)$$

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).

1.6 Review Questions

- 1 A matrix A is defined to convert any vector $\vec{r} = x\hat{i} + y\hat{j}$ into $\vec{r}_2 = 2x\hat{i} + 4y\hat{j}$. Verify that \hat{i} and \hat{j} are orthonormal eigenvectors of this matrix, with eigenvalues 2, respectively 4.
- 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 $(\cos 45^\circ, \sin 45^\circ)$ and $(\cos 45^\circ, -\sin 45^\circ)$ are orthonormal eigenvectors of this matrix, with eigenvalues 2 respectively 0. Note: $\cos 45^\circ = \sin 45^\circ = \frac{1}{2}\sqrt{2}$
- 3 Show that the operator $\hat{2}$ is a Hermitian operator, but \hat{i} is not.
- 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;

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

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.

- 5 Show that an operator such as \hat{x}^2 , corresponding to multiplying by a real function, is an Hermitian operator.
- 6 Show that the operator d/dx is *not* a Hermitian operator, but id/dx is, assuming that the functions on which they act vanish at the ends of the interval $a \leq x \leq b$ 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 $x = a$.)
- 7 Show that if A is a Hermitian operator, then so is A^2 . As a result, under the conditions of the previous question, $-d^2/dx^2$ is a Hermitian operator too. (And so is just d^2/dx^2 , of course, but $-d^2/dx^2$ is the one with the positive eigenvalues, the squares of the eigenvalues of id/dx .)
- 8 A complete set of orthonormal eigenfunctions of $-d^2/dx^2$ on the interval $0 \leq x \leq \pi$ that are zero at the end points is the infinite set of functions

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

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

$$1, 4, 9, 16, \dots$$

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

- 9 A complete set of orthonormal eigenfunctions of the operator id/dx that are periodic on the interval $0 \leq x \leq 2\pi$ are the infinite set of functions

$$\dots, \frac{e^{-3ix}}{\sqrt{2\pi}}, \frac{e^{-2ix}}{\sqrt{2\pi}}, \frac{e^{-ix}}{\sqrt{2\pi}}, \frac{1}{\sqrt{2\pi}}, \frac{e^{ix}}{\sqrt{2\pi}}, \frac{e^{2ix}}{\sqrt{2\pi}}, \frac{e^{3ix}}{\sqrt{2\pi}}, \dots$$

Check that these functions are indeed periodic, orthonormal, and that they are eigenfunctions of id/dx with the real eigenvalues

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

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

1.7 Additional Points

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

1.7.1 Dirac notation

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

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

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

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

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

In any case, $\langle f|Ag\rangle$ and $\langle f|A|g\rangle$ mean the same thing:

$$\int_{\text{all } x} f^*(x) (Ag(x)) \, dx$$

If A is a Hermitian operator, this book will on occasion use the additional bar to indicate that the operator has been brought to the other side to act on f instead of g .

1.7.2 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:

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

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

Chapter 2

Basic Ideas of Quantum Mechanics

Abstract

In this chapter the basic ideas of quantum mechanics are described and then two examples are 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]*

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.

First of all, in this chapter the classical picture of particles with positions and velocities will be thrown out. 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, or “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 turn the position of a particle into certainty, its linear momentum explodes. If the position is certain, the linear momentum has infinite uncertainty and vice-versa. In reality neither position nor linear momentum can have a definite 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,” and is associated with the total energy of the particle. The chapter will conclude by finding the eigenvalues of the Hamiltonian for two basic cases. These eigenvalues describe the only possible values that the total energy of the particle can have for those systems.

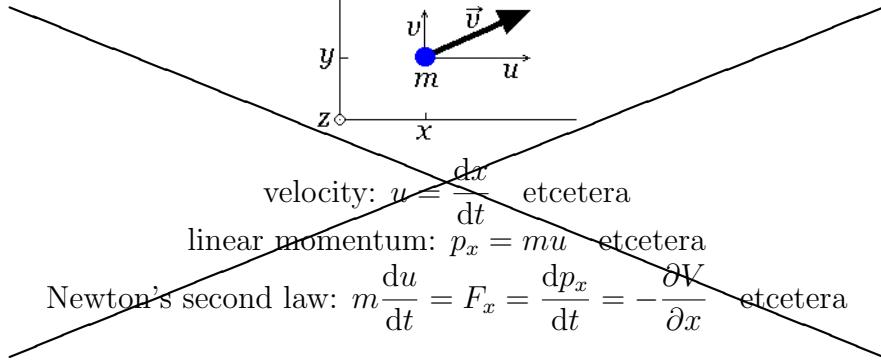
Because the two systems are so basic, much quantum mechanical analysis starts with one or the other. The first system is a particle stuck in a pipe of square cross section. While relatively simple, this case describes some of the quantum effects encountered in nano technology. In later chapters, it will be found that this case also provides a basic model for such systems as valence electrons in metals and ideal gases.

The second system is the quantum version of the simple spring mass system, the harmonic oscillator. It provides a model not just for vibrations of atoms in crystals, but also for the creation of the photons of electromagnetic radiation.

2.1 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:



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):

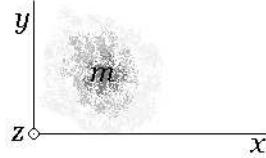


Figure 2.1: A visualization of an arbitrary wave function.

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

$$|\Psi(\vec{r}; t)|^2 d^3\vec{r} \quad (2.1)$$

is the probability of finding the particle within a small volume, of size $d^3\vec{r} = dx dy dz$, 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):

$$\int_{\text{all } \vec{r}} |\Psi(\vec{r}; t)|^2 d^3\vec{r} = 1 \quad (2.2)$$

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, the so-called 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 13.2.

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 this quantum amplitude, 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 passing through the second hole may arrive with the opposite sign of the wave function passing through the first hole. If that happens, the net wave function becomes zero, 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 definite 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.
-

2.2 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 2.2 is a combination plot of the position x of a particle and the corresponding linear momentum $p_x = mu$, (with m the mass and u the velocity in the x -direction):

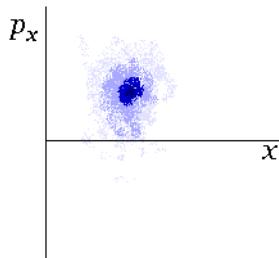


Figure 2.2: Combined plot of position and momentum components.

Figure 2.3 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:

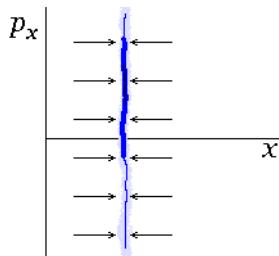


Figure 2.3: The uncertainty principle illustrated.

Heisenberg showed that according to quantum mechanics, the area of the blue “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.

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.

- o It also implies that a particle cannot have a mathematically precise linear momentum (velocity), since that would imply an infinite uncertainty in position.
-

2.3 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* \hat{x} . It turns the wave function Ψ into $x\Psi$:

$$\Psi(x, y, z, t) \xrightarrow{\hat{x}} x\Psi(x, y, z, t) \quad (2.3)$$

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

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

$$\boxed{\hat{p}_x = \frac{\hbar}{i} \frac{\partial}{\partial x}} \quad (2.4)$$

that turns Ψ into its *x*-derivative:

$$\Psi(x, y, z, t) \xrightarrow{\hat{p}_x = \frac{\hbar}{i} \frac{\partial}{\partial x}} \frac{\hbar}{i} \Psi_x(x, y, z, t) \quad (2.5)$$

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 m²/s.

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

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

$$\boxed{\hat{p}_y = \frac{\hbar}{i} \frac{\partial}{\partial y} \quad \hat{p}_z = \frac{\hbar}{i} \frac{\partial}{\partial z}} \quad (2.6)$$

The *kinetic energy operator* \hat{T} is:

$$\hat{T} = \frac{\hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2}{2m} \quad (2.7)$$

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

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

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:

$$\hat{T} = -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right). \quad (2.8)$$

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

$$\boxed{\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}} \quad (2.9)$$

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

$$\boxed{\hat{T} = -\frac{\hbar^2}{2m} \nabla^2} \quad (2.10)$$

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

$$\boxed{H = -\frac{\hbar^2}{2m} \nabla^2 + V} \quad (2.11)$$

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, \dots with:

$$H\psi_n = E_n\psi_n \quad \text{for } n = 1, 2, 3, \dots \quad (2.12)$$

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.
-

2.4 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.

2.4.1 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:

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

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 6.1), that says how the wave function evolves. This equation contradicts the collapse, (chapter 13.4.)

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.

- 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.
-

2.4.2 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,

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

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 13.5, 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 13.1.

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.

2.5 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 end. 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 2.5.9 shows, the particle in a pipe is really quite interesting despite its simplicity.

2.5.1 The physical system

The system to be analyzed is shown in figure 2.4 as it would appear in classical non-quantum physics. A particle is bouncing around between the two ends of a 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 2.4 shows.

The actual quantum system to be analyzed is shown in figure 2.5. A particle



Figure 2.4: Classical picture of a particle in a closed pipe.



Figure 2.5: 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

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

2.5.2 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 2.6. Also, the length of the pipe will be called ℓ_x .

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

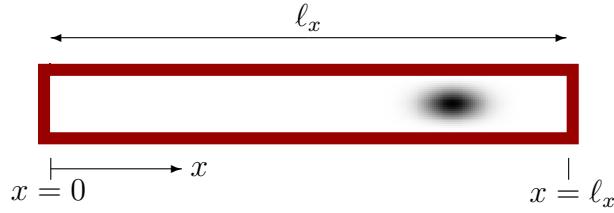


Figure 2.6: Definitions for one-dimensional motion in a pipe.

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

2.5.3 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 \hat{T} is needed. You can just look up its precise form in section 2.3 and find it is:

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

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:

$$H = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \tag{2.13}$$

Key Points

- The one-dimensional Hamiltonian (2.13) has been written down.
-

2.5.4 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:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} = E\psi \quad (2.14)$$

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 {A.10}:

$$\psi = 0 \text{ at } x = 0 \quad \text{and} \quad \psi = 0 \text{ at } x = \ell_x \quad (2.15)$$

Key Points

- The Hamiltonian eigenvalue problem (2.14) has been found.
 - It also includes the boundary conditions (2.15).
-

2.5.5 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. [28, 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

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

This solution may easily be 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 $C_2 = -C_1$,

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

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 [28, 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 [28, item 19.7]:

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

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 $\pi, 2\pi, 3\pi, \dots$. So there is a nonzero solution for each of the following values of the positive constant k :

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

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

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

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

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

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:

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

or generically:

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

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:

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

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:

$$\begin{aligned}\psi_1 &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{\pi}{\ell_x}x\right) & E_1 &= \frac{\hbar^2\pi^2}{2m\ell_x^2} \\ \psi_2 &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{2\pi}{\ell_x}x\right) & E_2 &= \frac{2^2\hbar^2\pi^2}{2m\ell_x^2} \\ \psi_3 &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{3\pi}{\ell_x}x\right) & E_3 &= \frac{3^2\hbar^2\pi^2}{2m\ell_x^2} \\ &\vdots & &\vdots\end{aligned}\tag{2.16}$$

or in generic form:

$$\psi_n = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{n\pi}{\ell_x}x\right) \quad E_n = \frac{n^2\hbar^2\pi^2}{2m\ell_x^2} \quad \text{for } n = 1, 2, 3, 4, 5, \dots\tag{2.17}$$

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

Key Points

- o After a lot of grinding mathematics armed with table books, the energy eigenfunctions and eigenvalues have finally been found
- o There are infinitely many of them.
- o They are as listed in (2.17) above. The first few are also written out explicitly in (2.16).

2.5.5 Review Questions

- 1 Write down eigenfunction number 6.
- 2 Write down eigenvalue number 6.

2.5.6 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

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

derived in the previous subsection.

Energy values are typically shown graphically in the form of an “energy spectrum”, as in figure 2.7. Energy is plotted upwards, so the vertical height

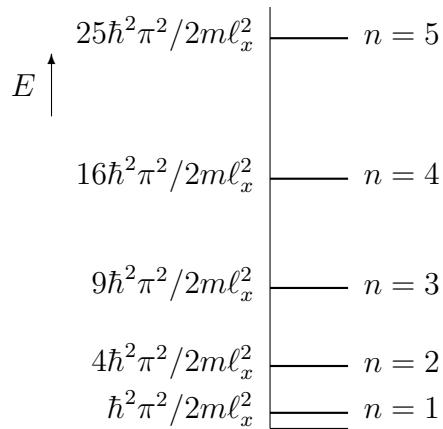


Figure 2.7: One-dimensional energy spectrum for a particle in a pipe.

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 2.7. 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 2.7. 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.

- o The ground state of lowest energy has nonzero kinetic energy.
-

2.5.6 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.
 - 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.
 - 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?
-

2.5.7 Discussion of the eigenfunctions

This subsection discusses the one-dimensional energy eigenfunctions of the particle in the pipe. The solution of subsection 2.5.5 found them to be:

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

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 2.8. As noted in section 2.1, it is the *square* magnitude of a wave function that gives the probability of finding the particle. So, the second graph in figure 2.8 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 2.8. It is seen 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 2.9 shows the two next lowest energy states

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

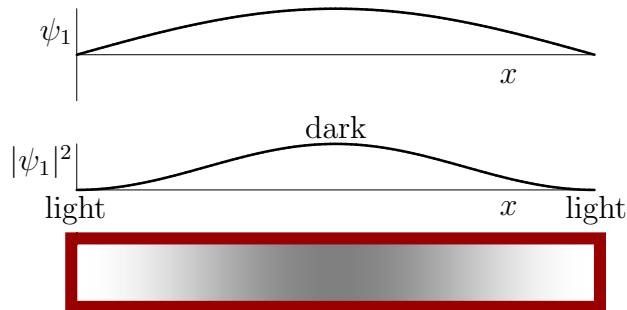


Figure 2.8: One-dimensional ground state of a particle in a pipe.

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 2.5. Moreover, it turns out that energy eigenstates are stationary states: the probabilities shown in figures 2.8 and 2.9 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 6.6.4 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.

2.5.7 Review Questions

- 1 So how does, say, the one-dimensional eigenstate ψ_6 look?
- 2 Generalizing the results above, for eigenfunction ψ_n , any n , how many distinct regions are there where the particle may be found?

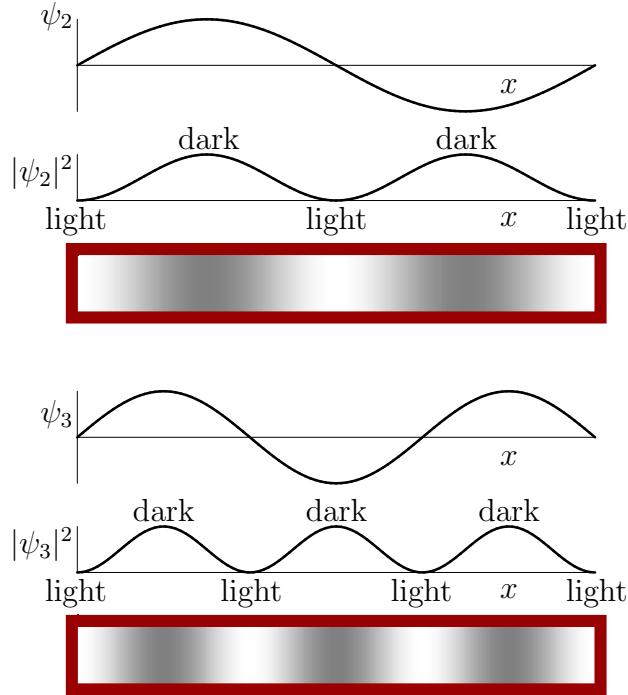


Figure 2.9: Second and third lowest one-dimensional energy states.

- 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?
-

2.5.8 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 2.10: in addi-

tion 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.

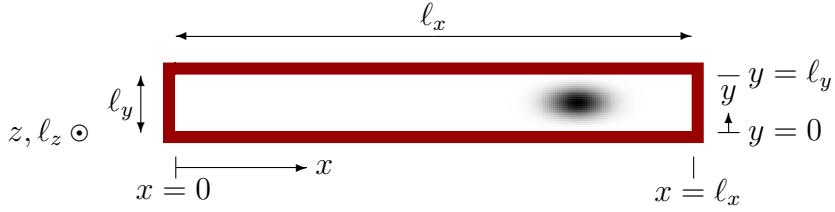


Figure 2.10: Definition of all variables for motion in a pipe.

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 :

$$\begin{aligned} \psi_{x1} &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{\pi}{\ell_x}x\right) & E_{x1} &= \frac{\hbar^2\pi^2}{2m\ell_x^2} \\ \psi_{x2} &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{2\pi}{\ell_x}x\right) & E_{x2} &= \frac{2^2\hbar^2\pi^2}{2m\ell_x^2} \\ \psi_{x3} &= \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{3\pi}{\ell_x}x\right) & E_{x3} &= \frac{3^2\hbar^2\pi^2}{2m\ell_x^2} \\ &\vdots & &\vdots \end{aligned} \tag{2.18}$$

or in generic form:

$$\psi_{xn_x} = \sqrt{\frac{2}{\ell_x}} \sin\left(\frac{n_x\pi}{\ell_x}x\right) \quad E_{xn_x} = \frac{n_x^2\hbar^2\pi^2}{2m\ell_x^2} \quad \text{for } n_x = 1, 2, 3, \dots \tag{2.19}$$

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:

$$\psi_{yn_y} = \sqrt{\frac{2}{\ell_y}} \sin\left(\frac{n_y\pi}{\ell_y}y\right) \quad E_{yn_y} = \frac{n_y^2\hbar^2\pi^2}{2m\ell_y^2} \quad \text{for } n_y = 1, 2, 3, \dots \tag{2.20}$$

and

$$\psi_{zn_z} = \sqrt{\frac{2}{\ell_z}} \sin\left(\frac{n_z\pi}{\ell_z}z\right) \quad E_{zn_z} = \frac{n_z^2\hbar^2\pi^2}{2m\ell_z^2} \quad \text{for } n_z = 1, 2, 3, \dots \tag{2.21}$$

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, {A.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:

$$\boxed{\psi_{n_x n_y n_z} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{n_x \pi}{\ell_x} x\right) \sin\left(\frac{n_y \pi}{\ell_y} y\right) \sin\left(\frac{n_z \pi}{\ell_z} z\right)} \quad (2.22)$$

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:

$$\boxed{E_{n_x n_y n_z} = \frac{n_x^2 \hbar^2 \pi^2}{2m\ell_x^2} + \frac{n_y^2 \hbar^2 \pi^2}{2m\ell_y^2} + \frac{n_z^2 \hbar^2 \pi^2}{2m\ell_z^2}} \quad (2.23)$$

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:

$$\boxed{\psi_{111} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{\pi}{\ell_x} x\right) \sin\left(\frac{\pi}{\ell_y} y\right) \sin\left(\frac{\pi}{\ell_z} z\right)} \quad (2.24)$$

This ground state is shown in figure 2.11. The y - and z -factors ensure that the wave function is now zero at all the surfaces of the pipe.

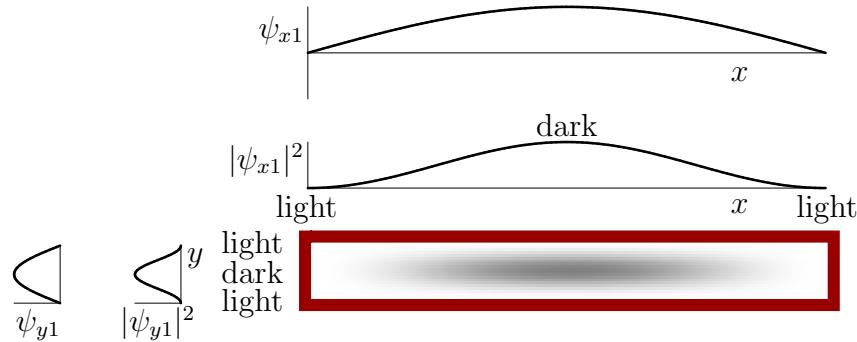


Figure 2.11: True ground state of a particle in a pipe.

The ground state energy is:

$$E_{111} = \frac{\hbar^2 \pi^2}{2m\ell_x^2} + \frac{\hbar^2 \pi^2}{2m\ell_y^2} + \frac{\hbar^2 \pi^2}{2m\ell_z^2} \quad (2.25)$$

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

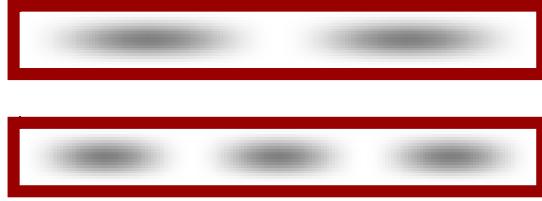


Figure 2.12: True second and third lowest energy states.

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

$$\psi_{211} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{2\pi}{\ell_x}x\right) \sin\left(\frac{\pi}{\ell_y}y\right) \sin\left(\frac{\pi}{\ell_z}z\right) \quad (2.26)$$

$$\psi_{311} = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin\left(\frac{3\pi}{\ell_x}x\right) \sin\left(\frac{\pi}{\ell_y}y\right) \sin\left(\frac{\pi}{\ell_z}z\right) \quad (2.27)$$

are shown in figure 2.12. They have energy levels:

$$E_{211} = \frac{4\hbar^2\pi^2}{2m\ell_x^2} + \frac{\hbar^2\pi^2}{2m\ell_y^2} + \frac{\hbar^2\pi^2}{2m\ell_z^2} \quad E_{311} = \frac{9\hbar^2\pi^2}{2m\ell_x^2} + \frac{\hbar^2\pi^2}{2m\ell_y^2} + \frac{\hbar^2\pi^2}{2m\ell_z^2} \quad (2.28)$$

Key Points

- o Three-dimensional energy eigenfunctions can be found as products of one-dimensional ones.
- o Three-dimensional energies can be found as sums of one-dimensional ones.
- o Example three-dimensional eigenstates have been shown.

2.5.8 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?

- 2 At what ratio of ℓ_y/ℓ_x does the energy E_{121} become higher than the energy E_{311} ?
 - 3 Shade the regions where the particle is likely to be found in the ψ_{322} energy eigenstate.
-

2.5.9 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 2.11. 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 2.12, 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:

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

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.

In chapter 6.1 it will be found that for nonzero times, the wave function of this particle is given by

$$\Psi = \sqrt{\frac{4}{5}}e^{-iE_{111}t/\hbar}\psi_{111} + \sqrt{\frac{1}{5}}e^{-iE_{211}t/\hbar}\psi_{211}.$$

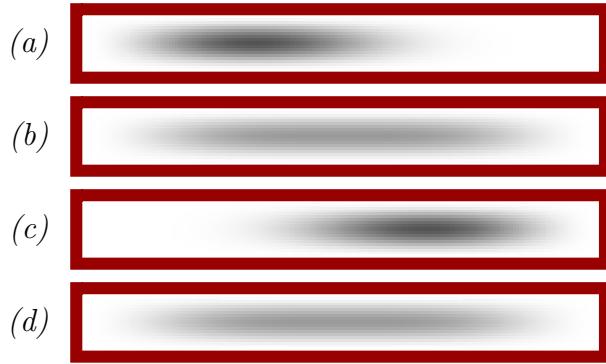


Figure 2.13: A combination of ψ_{111} and ψ_{211} seen at some typical times.

Using this expression, the probability of finding the particle, $|\Psi|^2$, can be plotted for various times. That is done in figure 2.13 for four typical times. It shows that with uncertainty in energy, the wave function blob does move. It performs a periodic oscillation: after figure 2.13(d), the wave function returns to state 2.13(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 $\psi_{311}, \psi_{411}, \dots$, you can get closer and closer to a localized wave function blob bouncing around.

But if you look closer at figure 2.13, 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

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

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.

2.6 The Harmonic Oscillator

This section provides an in-depth discussion of a basic quantum system. The case to be analyzed is a particle constrained by 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. This is usually also a good approximation for other constrained systems 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 2.14. The stiffness of the springs is characterized by the

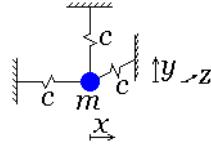


Figure 2.14: The harmonic oscillator.

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.

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

$$\omega = \sqrt{\frac{c}{m}} \quad (2.29)$$

in radians per second. This frequency 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.

2.6.1 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 \hat{T} from section 2.3 to get the Hamiltonian:

$$H = -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + \frac{1}{2}c(x^2 + y^2 + z^2) \quad (2.30)$$

Key Points

- The Hamiltonian (2.30) has been found.
-

2.6.2 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”):

$$\left[-\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + \frac{1}{2}c(x^2 + y^2 + z^2) \right] \psi = E\psi \quad (2.31)$$

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:

$$\psi = \psi_x(x)\psi_y(y)\psi_z(z) \quad (2.32)$$

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 :

$$\begin{aligned} E &= E_x + E_y + E_z \\ E_x &= -\frac{\hbar^2}{2m} \frac{\psi''_x(x)}{\psi_x(x)} + \frac{1}{2}cx^2 \\ E_y &= -\frac{\hbar^2}{2m} \frac{\psi''_y(y)}{\psi_y(y)} + \frac{1}{2}cy^2 \\ E_z &= -\frac{\hbar^2}{2m} \frac{\psi''_z(z)}{\psi_z(z)} + \frac{1}{2}cz^2 \end{aligned} \quad (2.33)$$

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:

$$\left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2}cx^2 \right] \psi_x = E_x \psi_x \quad (2.34)$$

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 note {A.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 :

$$\begin{aligned} E_{x0} &= \frac{1}{2}\hbar\omega & \psi_{x0}(x) &= h_0(x) \\ E_{x1} &= \frac{3}{2}\hbar\omega & \psi_{x1}(x) &= h_1(x) \\ E_{x2} &= \frac{5}{2}\hbar\omega & \psi_{x2}(x) &= h_2(x) \\ &\vdots & &\vdots \end{aligned} \quad (2.35)$$

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 2.1 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, (A.27), in note {A.12}.

But it are the eigenvalues that you may 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 -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:

$$E_{xn_x} = \frac{2n_x + 1}{2}\hbar\omega \quad \text{for } n_x = 0, 1, 2, 3, \dots \quad (2.36)$$

$h_0(x) = \frac{1}{(\pi\ell^2)^{1/4}} e^{-\xi^2/2}$	
$h_1(x) = \frac{2\xi}{(4\pi\ell^2)^{1/4}} e^{-\xi^2/2}$	$\omega = \sqrt{\frac{c}{m}}$
$h_2(x) = \frac{2\xi^2 - 1}{(4\pi\ell^2)^{1/4}} e^{-\xi^2/2}$	$\ell = \sqrt{\frac{\hbar}{m\omega}}$
$h_3(x) = \frac{2\xi^3 - 3\xi}{(9\pi\ell^2)^{1/4}} e^{-\xi^2/2}$	$\xi = \frac{x}{\ell}$
$h_4(x) = \frac{4\xi^4 - 12\xi^2 + 3}{(576\pi\ell^2)^{1/4}} e^{-\xi^2/2}$	

Table 2.1: First few one-dimensional eigenfunctions of the harmonic oscillator.

The eigenvalue problem has now been solved, because the equations for Y and Z are mathematically the same and must therefore have corresponding solutions:

$$E_{yn_y} = \frac{2n_y + 1}{2}\hbar\omega \quad \text{for } n_y = 0, 1, 2, 3, \dots \quad (2.37)$$

$$E_{zn_z} = \frac{2n_z + 1}{2}\hbar\omega \quad \text{for } n_z = 0, 1, 2, 3, \dots \quad (2.38)$$

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:

$$E_{n_x n_y n_z} = \frac{2n_x + 2n_y + 2n_z + 3}{2}\hbar\omega \quad (2.39)$$

where the “quantum numbers” n_x , n_y , and n_z may each have any value in the range $0, 1, 2, 3, \dots$

The corresponding eigenfunction of the complete system is:

$$\psi_{n_x n_y n_z} = h_{n_x}(x)h_{n_y}(y)h_{n_z}(z) \quad (2.40)$$

where the functions h_0, h_1, \dots are in table 2.1 or in (A.27) 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.
- Generic expressions for the eigenvalues are above in (2.39) and for the eigenfunctions in (2.40).

2.6.2 Review Questions

- 1 Write out the ground state energy.
- 2 Write out the ground state wave function fully.
- 3 Write out the energy E_{100} .
- 4 Write out the eigenstate ψ_{100} fully.

2.6.3 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, (2.40) and a corresponding energy eigenvalue (2.39). 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 2.15. 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 m²/s, macroscopically the

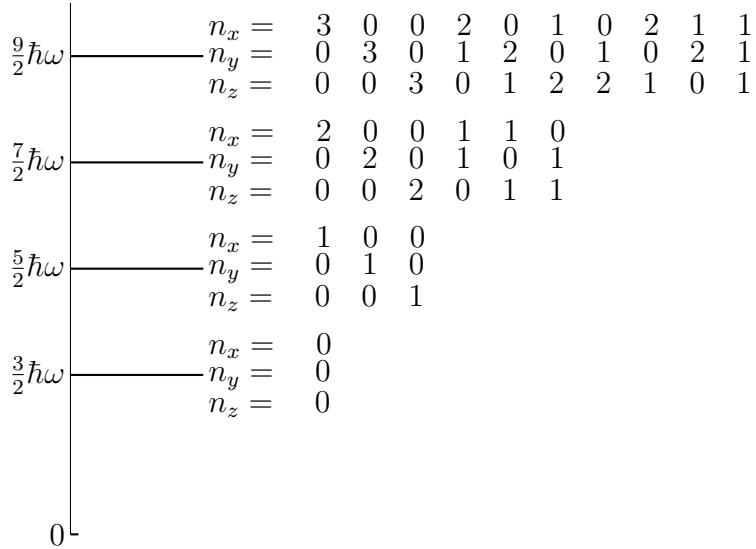


Figure 2.15: The energy spectrum of the harmonic oscillator.

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 has a value:

$$E_{000} = \frac{3}{2}\hbar\omega \quad (2.41)$$

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 certain position. Also, to get the kinetic energy to be zero, the linear momentum would have to be zero for certain, and the uncertainty relationship 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 un-

certainty in position to keep the momentum low, minimizing kinetic energy. In fact, the compromise results in potential and kinetic energies that are exactly equal, {A.13}.

For energy levels above the ground state, figure 2.15 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

- o Energy values can be graphically represented as an energy spectrum.
- o The energy values of the harmonic oscillator are equally spaced, with a constant energy difference of $\hbar\omega$ between successive levels.
- o The ground state of lowest energy has nonzero kinetic and potential energy.
- o For any energy level above the ground state, there is more than one eigenstate that produces that energy.

2.6.3 Review Questions

- 1 Verify that the sets of quantum numbers shown in the spectrum figure 2.15 do indeed produce the indicated energy levels.
- 2 Verify that there are no sets of quantum numbers missing in the spectrum figure 2.15; the listed ones are the only ones that produce those energy levels.

2.6.4 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 (2.40) of subsection 2.6.2 equal to

$$\psi_{000} = h_0(x)h_0(y)h_0(z) \quad (2.42)$$

where function h_0 is in table 2.1. The wave function in the ground state must be equal to the eigenfunction to within a constant:

$$\Psi_{\text{gs}} = c_{000}h_0(x)h_0(y)h_0(z) \quad (2.43)$$

where the magnitude of the constant c_{000} must be one. Using the expression for function h_0 from table 2.1, the properties of the ground state can be explored.

As noted earlier in section 2.1, 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 2.16. 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.

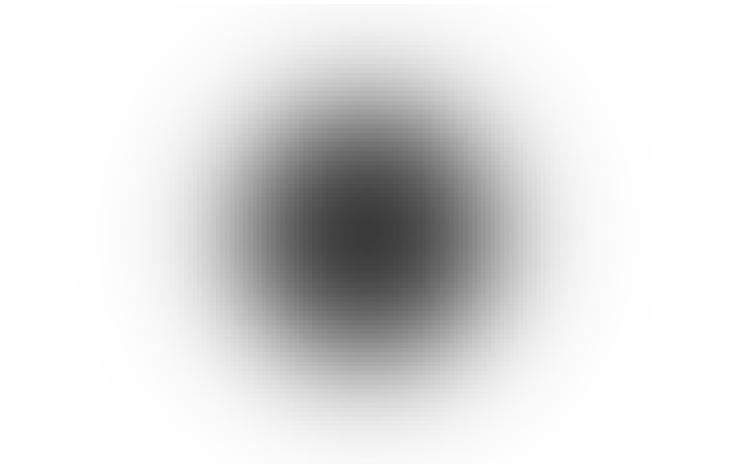


Figure 2.16: Ground state of the harmonic oscillator

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 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 2.16 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 2.17 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.

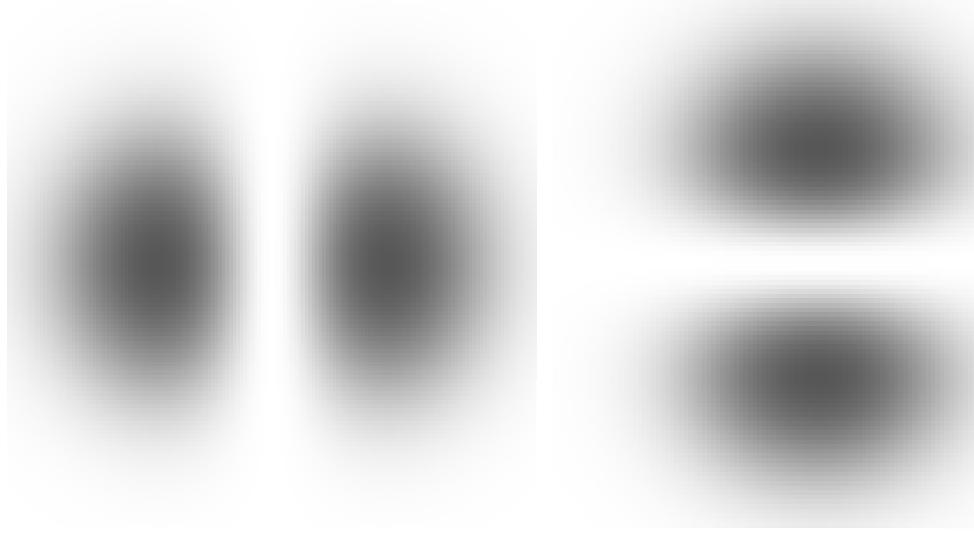


Figure 2.17: Wave functions ψ_{100} and ψ_{010} .

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 6.1.4.) Also, while in each case there are two blobs occupied by a single particle, the particle will never 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 2.18 shows eigenfunction 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 6.6.4.

The basic reason that quantum mechanics is so slow is simple. To analyze,

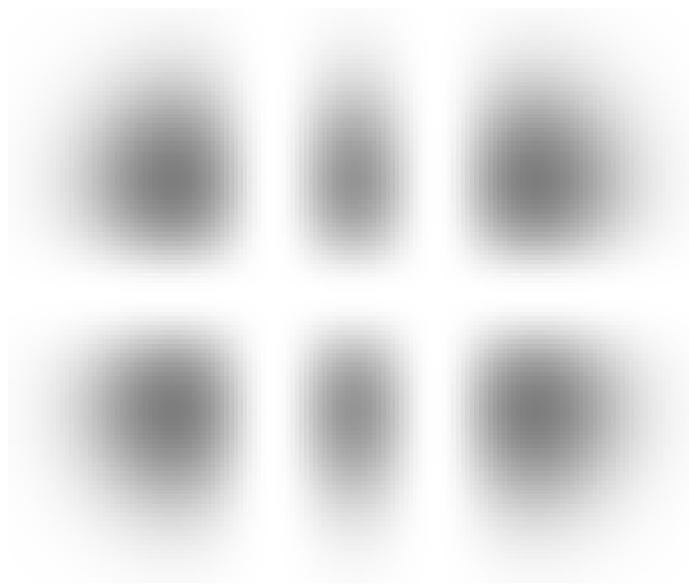


Figure 2.18: Energy eigenfunction ψ_{213} .

say the x -motion, classical physics says: “the *value* of the total energy E_x is

$$E_x = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}cx^2,$$

now go analyze the motion!”. Quantum mechanics says: “the total energy *operator* H_x is

$$H_x = \frac{1}{2}m \left(\frac{\hbar}{im} \frac{\partial}{\partial x} \right)^2 + \frac{1}{2}c\hat{x}^2,$$

now first figure out the possible energy *values* E_{x0}, E_{x1}, \dots before you can even start thinking about analyzing the motion.”

Key Points

- o-□ The ground state wave function is spherically symmetric: it looks the same seen from any angle.
 - o-□ In energy eigenstates the particle position is uncertain.
-

2.6.4 Review Questions

- 1 Write out the ground state wave function and show that it is indeed spherically symmetric.

- 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.
 - 3** Write down the explicit expression for the eigenstate ψ_{213} using table 2.1, then verify that it looks like figure 2.18 when looking along the z -axis, with the x -axis horizontal and the y -axis vertical.
-

2.6.5 Degeneracy

As the energy spectrum figure 2.15 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 specific 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 2.17, 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 2.6.2. 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.

2.6.5 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 2.1, then $(\psi_{100} + \psi_{010})/\sqrt{2}$ 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.

2.6.6 Non-eigenstates

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:

$$\Psi(x, y, z, t) = \sum_{n_x=0}^{\infty} \sum_{n_y=0}^{\infty} \sum_{n_z=0}^{\infty} c_{n_x n_y n_z} \psi_{n_x n_y n_z} \quad (2.44)$$

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 2.19.

Figure 2.19: Arbitrary wave function (not an energy eigenfunction).

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.

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:

$$c_{n_x n_y n_z} = \langle \psi_{n_x n_y n_z} | \Psi \rangle \quad (2.45)$$

The reason this works is orthonormality of the eigenfunctions. As an example, consider the case of coefficient c_{100} :

$$c_{100} = \langle \psi_{100} | \Psi \rangle = \langle \psi_{100} | c_{000} \psi_{000} + c_{100} \psi_{100} + c_{010} \psi_{010} + c_{001} \psi_{001} + c_{200} \psi_{200} + \dots \rangle$$

Now proper eigenfunctions of Hermitian operators are orthonormal; the inner product between different eigenfunctions is zero, and between identical eigenfunctions is one:

$$\langle \psi_{100} | \psi_{000} \rangle = 0 \quad \langle \psi_{100} | \psi_{100} \rangle = 1 \quad \langle \psi_{100} | \psi_{010} \rangle = 0 \quad \langle \psi_{100} | \psi_{001} \rangle = 0 \quad \dots$$

So, the inner product above must indeed produce c_{100} .

Chapter 6.1 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 2.19, 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:

$$E_{x2} = \frac{5}{2}\hbar\omega \quad E_{y1} = \frac{3}{2}\hbar\omega \quad E_{z3} = \frac{7}{2}\hbar\omega$$

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 2.18.

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

- o□ The basic ideas of quantum mechanics were illustrated using an example.
 - o□ The energy eigenfunctions are not the only game in town. Their seemingly lowly coefficients are important too.
 - o□ 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.
-

Chapter 3

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.

Before the hydrogen atom can be discussed however, 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 no definite 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.

3.1 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.

3.1.1 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 $\hbar\nabla/\text{i}$ for the linear momentum, so the angular momentum operator becomes:

$$\hat{\vec{L}} = \frac{\hbar}{\text{i}} \hat{\vec{r}} \times \nabla \quad \hat{\vec{r}} \equiv (\hat{x}, \hat{y}, \hat{z}) \quad \nabla \equiv \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \quad (3.1)$$

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 (3.1) has been identified.
-

3.1.2 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 [17] or [34]), and that is what will be done here.

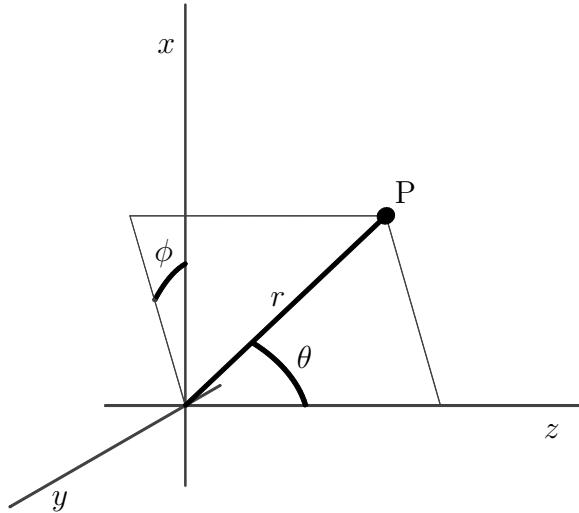


Figure 3.1: Spherical coordinates of an arbitrary point P.

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 3.1. 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:

$$\boxed{\hat{L}_z \equiv \frac{\hbar}{i} \frac{\partial}{\partial \phi}} \quad (3.2)$$

This can be verified by looking up the gradient operator ∇ in spherical coordinates in [28, pp. 124-126] and then taking the component of $\vec{r} \times \nabla$ in the z -direction.

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/i times the derivative with respect to the coordinate z along the z -axis, then surely the logical equivalent for z -angular momentum is \hbar/i times the derivative with respect to the angle ϕ around the z -axis?

Anyway, the eigenfunctions of the operator \hat{L}_z above turn out to be exponentials in ϕ . More precisely, the eigenfunctions are of the form

$$C(r, \theta) e^{im\phi} \quad (3.3)$$

where m is a constant and $C(r, \theta)$ can be any arbitrary function of r and θ . The number m is called the “magnetic quantum number”. It must be an integer, one of $\dots, -2, -1, 0, 1, 2, 3, \dots$. 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 (3.3) must again be the same, but that is only the case if m is an integer, as can be verified from the Euler formula (1.5).

The above solution is easily verified directly, and the eigenvalue L_z identified, by substitution into the eigenvalue problem $\hat{L}_z C e^{im\phi} = L_z C e^{im\phi}$ using the expression for \hat{L}_z above:

$$\frac{\hbar}{i} \frac{\partial C e^{im\phi}}{\partial \phi} = L_z C e^{im\phi} \implies \frac{\hbar}{i} i m C e^{im\phi} = L_z C e^{im\phi}$$

It follows that every eigenvalue is of the form:

$$L_z = m\hbar \text{ for } m \text{ an integer} \quad (3.4)$$

So the angular momentum in a given direction cannot just take on any value: it must be a whole multiple m , (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, chapter 12.1.6)

Key Points

- o 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.
- o Spherical coordinates based on the chosen z -axis are needed in this and subsequent analysis. They are defined in figure 3.1.
- o The operator for the z -component of angular momentum is (3.2), where ϕ is the angle around the z -axis.
- o The eigenvalues, or measurable values, of angular momentum in any arbitrary direction are whole multiples m , possibly negative, of \hbar .
- o The whole multiple m is called the magnetic quantum number.

3.1.2 Review Questions

- 1 If the angular momentum in a given direction is a multiple of $\hbar = 1.05457 \cdot 10^{-34}$ J s, then \hbar should have units of angular momentum. Verify that.
- 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.
- 3 Actually, based on the derived eigenfunction, $C(r, \theta) e^{im\phi}$, 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?

3.1.3 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 \hat{L}^2 operator and its eigenvalues.

Like the \hat{L}_z operator of the previous subsection, \hat{L}^2 can be written in terms of spherical coordinates. To do so, note first that, {A.14},

$$\hat{\vec{L}} \cdot \hat{\vec{L}} = \frac{\hbar}{i}(\vec{r} \times \nabla) \cdot \frac{\hbar}{i}(\vec{r} \times \nabla) = -\hbar^2 \vec{r} \cdot (\nabla \times (\vec{r} \times \nabla))$$

and then look up the gradient and the curl in [28, pp. 124-126]. The result is:

$$\hat{L}^2 \equiv -\frac{\hbar^2}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) - \frac{\hbar^2}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \quad (3.5)$$

Obviously, this result is not as intuitive as the \hat{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.

$Y_0^0 = \sqrt{\frac{1}{4\pi}}$	$Y_1^0 = \sqrt{\frac{3}{4\pi}} \cos(\theta)$	$Y_2^0 = \sqrt{\frac{5}{16\pi}} (3 \cos^2 \theta - 1)$
	$Y_1^1 = -\sqrt{\frac{3}{8\pi}} \sin \theta e^{i\phi}$	$Y_2^1 = -\sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{i\phi}$
	$Y_1^{-1} = \sqrt{\frac{3}{8\pi}} \sin \theta e^{-i\phi}$	$Y_2^{-1} = \sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{-i\phi}$
		$Y_2^2 = \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{2i\phi}$
		$Y_2^{-2} = \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{-2i\phi}$

Table 3.1: The first few spherical harmonics.

The solution to the problem may be summarized as follows. First, the non uniqueness is removed by demanding that the eigenfunctions are *also* eigenfunctions of \hat{L}_z , the operator of angular momentum in the z -direction. This makes the problem solvable, {A.15}, and the resulting eigenfunctions are called the “spherical harmonics” $Y_l^m(\theta, \phi)$. The first few are given explicitly in table 3.1. In case you need more of them for some reason, there is a generic expression (A.28) in note {A.15}.

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:

$$\int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} Y_l^m(\theta, \phi)^* Y_l^m(\theta, \phi) \sin \theta d\theta d\phi = \begin{cases} 1 & \text{if } l = \underline{l} \text{ and } m = \underline{m} \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

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 $e^{im\phi}$, as an eigenfunction of \hat{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 3.2. 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.

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

$$L_z = m\hbar \quad (3.7)$$

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

$$L^2 = l(l+1)\hbar^2 \quad (3.8)$$

where l is also an integer, and is called the “azimuthal quantum number”. It is maybe a weird result, (why not simply $l^2\hbar^2$?) but that is what square angular momentum turns out to be.

$Y_0^0 = \sqrt{\frac{1}{4\pi}}$	$rY_1^0 = \sqrt{\frac{3}{4\pi}}z$	$r^2Y_2^0 = \sqrt{\frac{5}{16\pi}}(2z^2 - x^2 - y^2)$
	$rY_1^1 = -\sqrt{\frac{3}{8\pi}}(x + iy)$	$r^2Y_2^1 = -\sqrt{\frac{15}{8\pi}}z(x + iy)$
	$rY_1^{-1} = \sqrt{\frac{3}{8\pi}}(x - iy)$	$r^2Y_2^{-1} = \sqrt{\frac{15}{8\pi}}z(x - iy)$
		$r^2Y_2^2 = \sqrt{\frac{15}{32\pi}}(x + iy)^2$
		$r^2Y_2^{-2} = \sqrt{\frac{15}{32\pi}}(x - iy)^2$

Table 3.2: The first few spherical harmonics rewritten.

The azimuthal quantum number is at least as large as the magnitude of the magnetic quantum number m :

$$\boxed{l \geq |m|} \quad (3.9)$$

The reason is that $\hat{L}^2 = \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2$ must be at least as large as \hat{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 $l \geq |m|$, 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 (3.5).
- 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 $0, 1, 2, 3, \dots$, 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 .

3.1.3 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

$$\int_{r=0}^{\infty} R(r)^* R(r) r^2 dr = 1.$$

- 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?
- 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 ?

3.1.4 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 10.1.1). The wave function will be in a state where L_x and L_y have a range of possible values $m_1\hbar, m_2\hbar, \dots$, 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

- o According to quantum mechanics, an exact nonzero angular momentum vector will never exist. If one component of angular momentum has a value, then the other two components will be uncertain.

3.2 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.

3.2.1 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

$$V = -\frac{e^2}{4\pi\epsilon_0 r} \quad (3.10)$$

with r the distance from the nucleus. The constant

$$e = 1.6 \cdot 10^{-19} \text{ C} \quad (3.11)$$

is the magnitude of the electric charges of the electron and proton, and the constant

$$\epsilon_0 = 8.85 \cdot 10^{-12} \text{ C}^2/\text{J m} \quad (3.12)$$

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 3.1. 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 \hat{T} of chapter 2.3, which involves the Laplacian. The Laplacian in spherical coordinates is readily available in table books, [28, p. 126], and the Hamiltonian is thus found to be:

$$H = -\frac{\hbar^2}{2m_e r^2} \left\{ \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right\} - \frac{e^2}{4\pi\epsilon_0 r} \quad (3.13)$$

where

$$m_e = 9.109 \cdot 10^{-31} \text{ kg} \quad (3.14)$$

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 $9.1044 \cdot 10^{-31}$ kg, {A.16}. This makes the solution exact, except for extremely small errors due to relativistic effects. (These are discussed in chapter 12.1.6.)

Key Points

- To analyze the hydrogen atom, you must use spherical coordinates.
- The Hamiltonian in spherical coordinates has been written down. It is (3.13).

3.2.2 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:

$$\begin{aligned} \left[-\frac{\hbar^2}{2m_e r^2} \left\{ \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right\} - \frac{e^2}{4\pi\epsilon_0} \frac{1}{r} \right] R\Theta\Phi \\ = ER\Theta\Phi \end{aligned}$$

To reduce this problem, premultiply by $2m_e r^2 / R\Theta\Phi$ and then separate the various terms:

$$\begin{aligned} -\frac{\hbar^2}{R} \frac{\partial}{\partial r} \left(r^2 \frac{\partial R}{\partial r} \right) + \frac{1}{\Theta\Phi} \left\{ -\frac{\hbar^2}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) - \frac{\hbar^2}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right\} \Theta\Phi \\ - \frac{2m_e r^2 e^2}{4\pi\epsilon_0} \frac{1}{r} = 2m_e r^2 E \quad (3.15) \end{aligned}$$

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 (3.15) 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 $E_{\theta\phi}$:

$$\left[-\frac{\hbar^2}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) - \frac{\hbar^2}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right] \Theta\Phi = E_{\theta\phi} \Theta\Phi \quad (3.16)$$

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 \hat{L}_z of section 3.1.2:

$$-\hbar^2 \left(\frac{\partial^2}{\partial\phi^2} \right) \Phi = \left(\frac{\hbar}{i} \frac{\partial}{\partial\phi} \right)^2 \Phi = \hat{L}_z^2 \Phi$$

So this equation must have the same eigenfunctions as the operator \hat{L}_z ,

$$\Phi_m = e^{im\phi}$$

and must have the square eigenvalues

$$E_\phi = (m\hbar)^2$$

(each application of \hat{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 (3.16) for $\Theta\Phi$ is even easier; it is exactly the one for the square angular momentum L^2 of section 3.1.3. (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 (3.15), 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, {A.17}.

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:

$n > l \geq |m|$

(3.17)

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:

$\psi_{nlm} = R_{nl}(r)Y_l^m(\theta, \phi)$

(3.18)

where the spherical harmonics Y_l^m were described in section 3.1.3. The brand new radial wave functions R_{nl} can be found written out in table 3.3 for small values of n and l , or in note {A.17}, (A.30), 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

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} \approx 0.529\,177\,10^{-10} \text{ m} \quad (3.19)$$

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.

$R_{10} = \frac{2}{\sqrt{a_0^3}} e^{-\rho}$	$R_{20} = \frac{2-\rho}{2\sqrt{2a_0^3}} e^{-\rho/2}$	$R_{30} = \frac{54 - 36\rho + 4\rho^2}{81\sqrt{3a_0^3}} e^{-\rho/3}$
	$R_{21} = \frac{\rho}{2\sqrt{6a_0^3}} e^{-\rho/2}$	$R_{31} = \frac{24\rho - 4\rho^2}{81\sqrt{6a_0^3}} e^{-\rho/3}$
		$R_{32} = \frac{4\rho^2}{81\sqrt{30a_0^3}} e^{-\rho/3}$
$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2}$		$\rho = \frac{r}{a_0}$

Table 3.3: The first few radial wave functions for hydrogen.

The energy eigenvalues are much simpler and more interesting than the eigenfunctions; they are

$$E_n = -\frac{\hbar^2}{2m_e a_0^2} \frac{1}{n^2} = \frac{E_1}{n^2} \quad n = 1, 2, 3, \dots \quad E_1 = -\frac{\hbar^2}{2m_e a_0^2} = -13.605\,7 \text{ eV} \quad (3.20)$$

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.56}, 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

- Skip 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 $n > l \geq |m|$. 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} .

3.2.2 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: $\int_0^\infty e^{-2u} u^2 du = \frac{1}{4}$.

- 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.

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 $1.602 \cdot 10^{-19}$ J. Values for the physical constants can be found at the start of this section and in the notations section.

3.2.3 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:

$$E_n = -\frac{\hbar^2}{2m_e a_0^2} \frac{1}{n^2} \quad n = 1, 2, 3, \dots$$

This subsection discusses the physical consequences of this result.

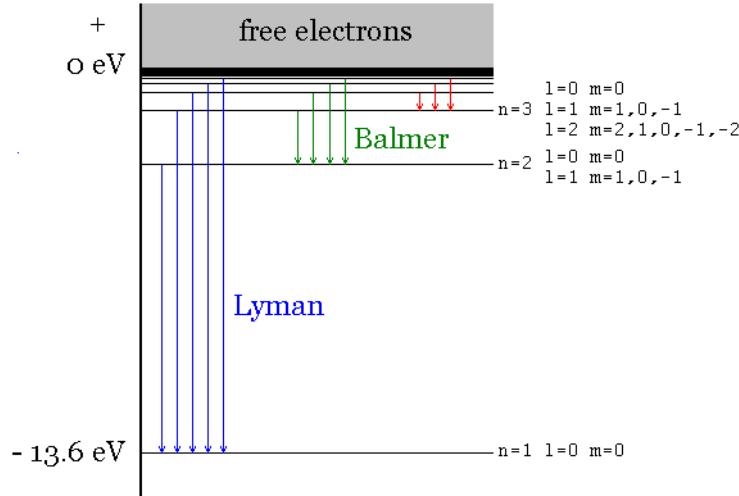


Figure 3.2: Spectrum of the hydrogen atom.

To aid the discussion, the allowed energies are plotted in the form of an energy spectrum in figure 3.2. 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 6.1.7, 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 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

$$E_1 = -13.6 \text{ eV} \quad (3.21)$$

(an eV or “electron volt” is $1.6 \cdot 10^{-19} \text{ J}$). 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 6.3. For now, it can be pointed out that different transitions are possible, as indicated by the arrows in figure 3.2. 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.

Atoms and molecules may also absorb electromagnetic energy of the same frequencies to enter an excited state and eventually emit it again in a different direction, chapter 6.3. In this way, they can remove these frequencies from light that passes them on its way to earth, resulting in an absorption spectrum. Since hydrogen is so prevalent in the universe, its energy levels as derived here are particularly important in astronomy.

Key Points

- o 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.
- o The ground state has a energy 13.6 eV below this zero level.
- o 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.
- o 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

$$\hbar\omega = E_{n_1} - E_{n_2}$$

- o Similarly, atoms with energy E_{n_2} may absorb electromagnetic energy of such a frequency.

3.2.3 Review Questions

- 1 If there are infinitely many energy levels $E_1, E_2, E_3, E_4, E_5, E_6, \dots$, where did they all go in the energy spectrum?
- 2 What is the value of energy level E_2 ? And E_3 ?
- 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.
- 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 ?

3.2.4 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 3.2.2 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

$$n > l \geq |m|$$

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 3.2.2):

$$\psi_{100}(r) = \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0} \quad (3.22)$$

where a_0 is called the “Bohr radius”,

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} = 0.53 \times 10^{-10} \text{ m} \quad (3.23)$$

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 3.3; the electron may be found within a blob size that is about thrice the Bohr radius, or roughly an Ångstrom, (10^{-10} m), in diameter.

Figure 3.3: 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, \dots are all degenerate; as the spectrum figure 3.2 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 3.4 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.



Figure 3.4: Eigenfunction ψ_{200} .

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 “spherically symmetric.” Similarly, the ground state ψ_{100} is commonly indicated as “1s”, having the lowest energy E_1 .

States which have azimuthal quantum number $l = 1$ are called “p” states, for some historical reason. In particular, the ψ_{21m} states are called “2p” states. As first example of such a state, figure 3.5 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 function squeezes close to the z axis, this state is often more specifically referred to as the “ $2p_z$ ” state. Think “points along the z -axis.”

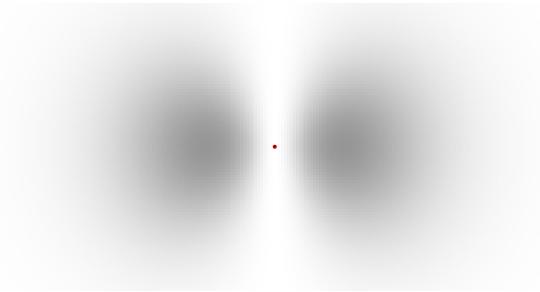
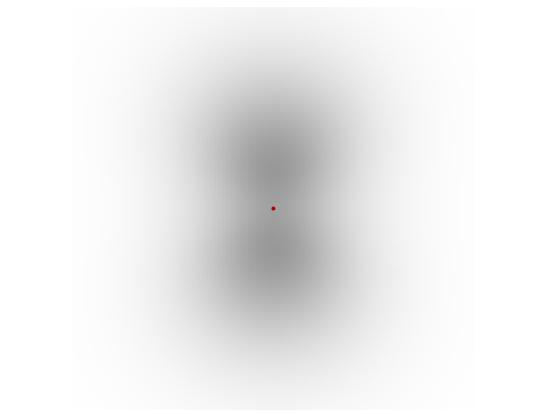
Figure 3.5: Eigenfunction ψ_{210} , or $2p_z$.

Figure 3.6 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.

Figure 3.6: Eigenfunction ψ_{211} (and ψ_{21-1}).

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.

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 $2p_x$ and $2p_y$; they are the combination states:

$$2p_x: \frac{1}{\sqrt{2}} (-\psi_{211} + \psi_{21-1}) \quad 2p_y: \frac{i}{\sqrt{2}} (\psi_{211} + \psi_{21-1}) \quad (3.24)$$

These two states are shown in figure 3.7; they look exactly like the “pointer” state $2p_z$ of figure 3.5, 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, $2p_y$ looks rotationally symmetric. Seen from the side, it would look like p_z in figure 3.5.)

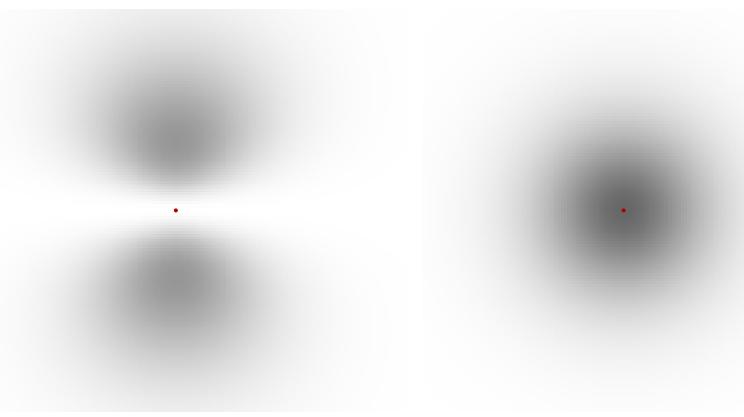


Figure 3.7: Eigenfunctions $2p_x$, left, and $2p_y$, right.

Note that unlike the two original states ψ_{211} and ψ_{21-1} , the states $2p_x$ and $2p_y$ 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 4.11.4, 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 .

- o- All higher energy levels are degenerate, there is more than one eigenstate producing that energy.
 - o- 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.
 - o- States of the form ψ_{n1m} are called p states. The basic 2p states are ψ_{21-1} , ψ_{210} , and ψ_{211} .
 - o- The state ψ_{210} is also called the $2p_z$ state, since it squeezes itself around the z -axis.
 - o- 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} .
 - o- 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.
 - o- However, since the E_2 energy level is degenerate, eigenstates of still different shapes are likely to show up in applications.
-

3.2.4 Review Questions

- 1 At what distance from the nucleus r , expressed as a multiple of the Bohr radius a_0 , becomes the square of the ground state wave function less than one percent of its value at the nucleus? What is that expressed in Å?
- 2 Check from the conditions

$$n > l \geq |m|$$

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 $2p_x$ and $2p_y$, 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} .)

- 3 Check that the states

$$2p_x = \frac{1}{\sqrt{2}} (-\psi_{211} + \psi_{21-1}) \quad 2p_y = \frac{i}{\sqrt{2}} (\psi_{211} + \psi_{21-1})$$

are properly normalized.

3.3 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 \hat{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.

- 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.
-

3.3.1 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:

$$\sigma = [\frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2 +$$

$$\begin{aligned} & \frac{1}{6}(4 - 3.5)^2 + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2]^{1/2} \\ & = 1.71 \end{aligned}$$

On average then, the throws are 1.71 points off from 3.5.

Key Points

- o The expectation value is obtained by summing the possible values times their probabilities.
- o To get the standard deviation, first find the average square deviation from the expectation value, then take a square root of that.

3.3.1 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?
- 2 Continuing this example, what will be the maximum deviation?
- 3 Continuing this example, what will be the standard deviation?
- 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?

3.3.2 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

$$a_1, a_2, a_3, \dots$$

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 $\alpha_1, \alpha_2, \alpha_3, \dots$ are the eigenfunctions of operator A , and the wave function is

$$\Psi = c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \dots$$

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:

$$\langle a \rangle = |c_1|^2 a_1 + |c_2|^2 a_2 + |c_3|^2 a_3 + \dots$$

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

$$\Psi = c_{100}\psi_{100} + c_{200}\psi_{200} + c_{210}\psi_{210} + c_{211}\psi_{211} + c_{21-1}\psi_{21-1} + c_{300}\psi_{300} + \dots$$

then the expectation value for energy will be, noting that $E_1 = -13.6$ eV, $E_2 = -3.4$ eV, ...:

$$\langle E \rangle = -|c_{100}|^2 13.6 \text{ eV} - |c_{200}|^2 3.4 \text{ eV} - |c_{210}|^2 3.4 \text{ eV} - |c_{211}|^2 3.4 \text{ eV} - \dots$$

Also, the expectation value of the square angular momentum will be, recalling that its eigenvalues are $l(l+1)\hbar^2$,

$$\langle L^2 \rangle = |c_{100}|^2 0 + |c_{200}|^2 0 + |c_{210}|^2 2\hbar^2 + |c_{211}|^2 2\hbar^2 + |c_{21-1}|^2 2\hbar^2 + |c_{300}|^2 0 + \dots$$

Also, the expectation value of the z -component of angular momentum will be, recalling that its eigenvalues are $m\hbar$,

$$\langle L_z \rangle = |c_{100}|^2 0 + |c_{200}|^2 0 + |c_{210}|^2 0 + |c_{211}|^2 \hbar - |c_{21-1}|^2 \hbar + |c_{300}|^2 0 + \dots$$

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.

3.3.2 Review Questions

- 1** The $2p_x$ pointer state of the hydrogen atom was defined as

$$\frac{1}{\sqrt{2}} (-\psi_{211} + \psi_{21-1}).$$

What are the expectation values of energy, square angular momentum, and z -angular momentum for this state?

- 2** Continuing the previous question, what are the standard deviations in energy, square angular momentum, and z -angular momentum?

3.3.3 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:

$$\langle A \rangle = \langle \Psi | A | \Psi \rangle. \quad (3.25)$$

(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 note {A.18}.

The simplified expression for the expectation value can also be used to find the standard deviation, σ_A or σ_a :

$$\sigma_A = \sqrt{\langle (A - \langle A \rangle)^2 \rangle} \quad (3.26)$$

where $\langle (A - \langle A \rangle)^2 \rangle$ is the inner product $\langle \Psi | (A - \langle A \rangle)^2 | \Psi \rangle$.

Key Points

- o The expectation value of a quantity a with operator A can be found as $\langle A \rangle = \langle \Psi | A | \Psi \rangle$.
- o Similarly, the standard deviation can be found using the expression $\sigma_A = \sqrt{\langle (A - \langle A \rangle)^2 \rangle}$.

3.3.3 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 \hat{L}^2 with eigenvalue $2\hbar^2$; however, ψ_{211} is an eigenfunction of z -angular momentum \hat{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.)

- 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.
-

3.3.4 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:

$$\langle x \rangle = \langle \psi_{100} | \hat{x} \psi_{100} \rangle = \iiint x |\psi_{100}|^2 dx dy dz.$$

This integral is zero. The reason is that $|\psi_{100}|^2$, shown as grey scale in figure 3.3, 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 3.3, 3.4, 3.5, 3.6, as well as the combination states $2p_x$ and $2p_y$ of figure 3.7, 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

$$\sigma_x = \sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\langle x^2 \rangle} = \sqrt{\iiint x^2 |\psi_{100}(x, y, z)|^2 dx dy dz}$$

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 $\hat{p}_x = \hbar\partial/\text{i}\partial x$:

$$\langle p_x \rangle = \langle \psi_{100} | \hat{p}_x \psi_{100} \rangle = \iiint \psi_{100} \frac{\hbar}{\text{i}} \frac{\partial \psi_{100}}{\partial x} dx dy dz = \frac{\hbar}{\text{i}} \iiint \frac{\partial^{\frac{1}{2}} \psi_{100}^2}{\partial x} dx dy dz$$

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 6.1. 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 6.

Key Points

- Examples of certain and uncertain quantities were given for example wave functions.
- A quantity is certain when the wave function is an eigenfunction of that quantity.

3.4 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.

3.4.1 Commuting operators

First, note that in many cases there is no fundamental prohibition against more than one quantity having 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 precise values at the same time.

More generally, two different quantities with operators A and B have precise values if the wave function is an eigenfunction of both A and B . So, the question whether two quantities can be certain 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, {A.19}:

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 2.6.2 commute:

$$\begin{aligned} H_x H_y \Psi &= \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2} cx^2 \right] \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial y^2} + \frac{1}{2} cy^2 \right] \Psi \\ &= \left(\frac{\hbar^2}{2m} \right)^2 \frac{\partial^4 \Psi}{\partial x^2 \partial y^2} - \frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \frac{1}{2} cy^2 \Psi - \frac{1}{2} cx^2 \frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial y^2} + \frac{1}{2} cx^2 \frac{1}{2} cy^2 \Psi \\ &= H_y H_x \Psi \end{aligned}$$

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 2.6.2.

Similarly, for the hydrogen atom, the total energy Hamiltonian H , the square angular momentum operator \hat{L}^2 and the z -component of angular momentum \hat{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 , \hat{L}^2 , and the x -component of angular momentum \hat{L}_x also all commute, and they too have a common set of

eigenfunctions. But that will *not* be the ψ_{nlm} , since \hat{L}_x and \hat{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

- o Operators commute if you can change their order, as in $AB = BA$.
- o For commuting operators, a common set of eigenfunctions exists.
- o For those eigenfunctions, the physical quantities corresponding to the commuting operators all have precise values at the same time.

3.4.1 Review Questions

- 1 The pointer state

$$2p_x = \frac{1}{\sqrt{2}} (-\psi_{211} + \psi_{21-1}).$$

is one of the eigenstates that H , \hat{L}^2 , and \hat{L}_x have in common. Check that it is not an eigenstate that H , \hat{L}^2 , and \hat{L}_z have in common.

3.4.2 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 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” $[A, B]$:

$$[A, B] \equiv AB - BA \tag{3.27}$$

A nonzero commutator $[A, B]$ demands a minimum amount of uncertainty in the corresponding quantities a and b . It can be shown, {A.20}, that the uncertainties, or standard deviations, σ_a in a and σ_b in b are at least so large that:

$$\sigma_a \sigma_b \geq \frac{1}{2} | \langle [A, B] \rangle | \tag{3.28}$$

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.

3.4.3 The Heisenberg uncertainty relationship

This section will work out the uncertainty relationship 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 \hat{x} , and the linear momentum operator $\hat{p}_x = \hbar\partial/\mathrm{i}\partial x$. These two operators do not commute, $\hat{p}_x\hat{x}\Psi$ is simply not the same as $\hat{x}\hat{p}_x\Psi$: $\hat{p}_x\hat{x}\Psi$ means multiply function Ψ by x to get the product function $x\Psi$ and then apply \hat{p}_x on that product, while $\hat{x}\hat{p}_x\Psi$ means apply \hat{p}_x on Ψ and then multiply the resulting function by x . The difference is found from writing it out:

$$\hat{p}_x\hat{x}\Psi = \frac{\hbar}{\mathrm{i}} \frac{\partial x\Psi}{\partial x} = \frac{\hbar}{\mathrm{i}}\Psi + \frac{\hbar}{\mathrm{i}}x\frac{\partial\Psi}{\partial x} = -\mathrm{i}\hbar\Psi + \hat{x}\hat{p}_x\Psi$$

the second equality resulting from differentiating out the product.

Comparing start and end shows that the difference between $\hat{x}\hat{p}_x$ and $\hat{p}_x\hat{x}$ is not zero, but $\mathrm{i}\hbar$. By definition, this difference is their commutator:

$$[\hat{x}, \hat{p}_x] = \mathrm{i}\hbar \tag{3.29}$$

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:

$$\sigma_x\sigma_{p_x} \geq \frac{1}{2}\hbar \tag{3.30}$$

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 $\Delta p_x \Delta x \geq \frac{1}{2}\hbar$ 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, $\Delta E \Delta t \geq \frac{1}{2}\hbar$, 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. One common definition will be given later in chapter 6.1.7.

Key Points

- The canonical commutator $[\hat{x}, \hat{p}_x]$ equals $i\hbar$.
- 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 $\frac{1}{2}\hbar$.

3.4.3 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.

3.4.4 Commutator reference [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 :

$$[A, B] = AB - BA \quad (3.31)$$

By this very definition, the commutator is zero for any two operators A_1 and A_2 that commute, (whose order can be interchanged):

$$[A_1, A_2] = 0 \quad \text{if } A_1 \text{ and } A_2 \text{ commute; } A_1 A_2 = A_2 A_1. \quad (3.32)$$

If operators all commute, all their products commute too:

$$[A_1 A_2 \dots A_k, A_{k+1} \dots A_n] = 0 \quad \text{if } A_1, A_2, \dots, A_k, A_{k+1}, \dots, A_n \text{ all commute.} \quad (3.33)$$

Everything commutes with itself, of course:

$$[A, A] = 0, \quad (3.34)$$

and everything commutes with a numerical constant; if A is an operator and a is some number, then:

$$[A, a] = [a, A] = 0. \quad (3.35)$$

The commutator is “antisymmetric”; or in simpler words, if you interchange the sides; it will change the sign, {A.21}:

$$[B, A] = -[A, B]. \quad (3.36)$$

For the rest however, linear combinations multiply out just like you would expect:

$$[aA + bB, cC + dD] = ac[A, C] + ad[A, D] + bc[B, C] + bd[B, D], \quad (3.37)$$

(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:

$$\underbrace{[AB, \dots]}_{\leftarrow \rightarrow} = A[\dots, B] + [\dots, A]B, \quad \underbrace{[\dots, AB]}_{\leftarrow \rightarrow} = A[\dots, B] + [\dots, A]B. \quad (3.38)$$

So, if A or B commutes with the other side of the operator, it can simply be taken out at its side; (the second commutator will be zero.) For example,

$$[A_1 B, A_2] = A_1 [B, A_2], \quad [B A_1, A_2] = [B, A_2] A_1$$

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:

$$[\hat{x}, \hat{y}] = [\hat{y}, \hat{z}] = [\hat{z}, \hat{x}] = 0 \quad (3.39)$$

as do position-dependent operators such as a potential energy $V(x, y, z)$:

$$[\hat{x}, V(x, y, z)] = [\hat{y}, V(x, y, z)] = [\hat{z}, V(x, y, z)] = 0 \quad (3.40)$$

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:

$$[\hat{p}_x, \hat{p}_y] = [\hat{p}_y, \hat{p}_z] = [\hat{p}_z, \hat{p}_x] = 0 \quad (3.41)$$

However, position operators and linear momentum operators in the same direction do *not* commute; instead:

$$[\hat{x}, \hat{p}_x] = [\hat{y}, \hat{p}_y] = [\hat{z}, \hat{p}_z] = i\hbar \quad (3.42)$$

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:

$$[\hat{x}, \hat{p}_y] = [\hat{x}, \hat{p}_z] = [\hat{y}, \hat{p}_z] = [\hat{y}, \hat{p}_x] = [\hat{z}, \hat{p}_x] = [\hat{z}, \hat{p}_y] = 0 \quad (3.43)$$

A generalization that is frequently very helpful is:

$$[f, \hat{p}_x] = i\hbar \frac{\partial f}{\partial x} \quad [f, \hat{p}_y] = i\hbar \frac{\partial f}{\partial y} \quad [f, \hat{p}_z] = i\hbar \frac{\partial f}{\partial z} \quad (3.44)$$

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:”

$$[\hat{L}_x, \hat{L}_y] = i\hbar \hat{L}_z \quad [\hat{L}_y, \hat{L}_z] = i\hbar \hat{L}_x \quad [\hat{L}_z, \hat{L}_x] = i\hbar \hat{L}_y \quad (3.45)$$

Note the $\dots xyzxyz\dots$ order of the indices that produces positive signs; a reversed $\dots zyxzy\dots$ order adds a minus sign. For example $[\hat{L}_z, \hat{L}_y] = -i\hbar \hat{L}_x$ because y following z is in reversed order.

The angular momentum components do all commute with the square angular momentum operator:

$$[\hat{L}_x, \hat{L}^2] = [\hat{L}_y, \hat{L}^2] = [\hat{L}_z, \hat{L}^2] = 0 \quad \text{where } \hat{L}^2 = \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2 \quad (3.46)$$

Just the opposite of the situation for linear momentum, position and angular momentum operators in the same direction commute,

$$[\hat{x}, \hat{L}_x] = [\hat{y}, \hat{L}_y] = [\hat{z}, \hat{L}_z] = 0 \quad (3.47)$$

but those in different directions do not:

$$[\hat{x}, \hat{L}_y] = [\hat{L}_x, \hat{y}] = i\hbar \hat{z} \quad [\hat{y}, \hat{L}_z] = [\hat{L}_y, \hat{z}] = i\hbar \hat{x} \quad [\hat{z}, \hat{L}_x] = [\hat{L}_z, \hat{x}] = i\hbar \hat{y} \quad (3.48)$$

Square position commutes with all components of angular momentum,

$$[\hat{r}^2, \hat{L}_x] = [\hat{r}^2, \hat{L}_y] = [\hat{r}^2, \hat{L}_z] = [\hat{r}^2, \hat{L}^2] = 0 \quad (3.49)$$

The commutator between position and square angular momentum is, using vector notation for conciseness,

$$[\hat{r}, \hat{L}^2] = -2\hbar^2\hat{r} - 2i\hbar\hat{r} \times \hat{L} = -2\hbar^2\hat{r} + 2i\hbar(\hat{r} \cdot \hat{r})\hat{p} - 2i\hbar\hat{r}(\hat{r} \cdot \hat{p}) \quad (3.50)$$

The commutators between linear and angular momentum are very similar to the ones between position and angular momentum:

$$[\hat{p}_x, \hat{L}_x] = [\hat{p}_y, \hat{L}_y] = [\hat{p}_z, \hat{L}_z] = 0 \quad (3.51)$$

$$[\hat{p}_x, \hat{L}_y] = [\hat{L}_x, \hat{p}_y] = i\hbar\hat{p}_z \quad [\hat{p}_y, \hat{L}_z] = [\hat{L}_y, \hat{p}_z] = i\hbar\hat{p}_x \quad [\hat{p}_z, \hat{L}_x] = [\hat{L}_z, \hat{p}_x] = i\hbar\hat{p}_y \quad (3.52)$$

$$[\hat{p}^2, \hat{L}_x] = [\hat{p}^2, \hat{L}_y] = [\hat{p}^2, \hat{L}_z] = [\hat{p}^2, \hat{L}^2] = 0 \quad (3.53)$$

$$[\hat{p}, \hat{L}^2] = -2\hbar^2\hat{p} - 2i\hbar\hat{p} \times \hat{L} = 2\hbar^2\hat{p} + 2i\hbar(\hat{r} \cdot \hat{p})\hat{p} - 2i\hbar\hat{r}(\hat{p} \cdot \hat{p}) \quad (3.54)$$

The following commutators are also useful:

$$[\vec{r} \times \hat{\vec{L}}, \hat{L}^2] = 2i\hbar\vec{r}\hat{L}^2 \quad [[\vec{r}, \hat{L}^2], \hat{L}^2] = 2\hbar^2(\vec{r}\hat{L}^2 + \hat{L}^2\vec{r}) \quad (3.55)$$

Commutators involving spin are discussed in a later chapter, 4.5.3.

Key Points

- □ Rules for evaluating commutators were given.
- □ Return to this subsection if you need to figure out some commutator or the other.

3.5 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 H₂⁺-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.”

3.5.1 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 3.2.1, there are now two terms to the potential energy, the electron experiencing attraction to both protons:

$$H = -\frac{\hbar^2}{2m_e} \nabla^2 - \frac{e^2}{4\pi\epsilon_0 r_l} - \frac{e^2}{4\pi\epsilon_0 r_r} \quad (3.56)$$

where r_l and r_r are the distances from the electron to the left and right protons,

$$r_l \equiv |\vec{r} - \vec{r}_{lp}| \quad r_r \equiv |\vec{r} - \vec{r}_{rp}| \quad (3.57)$$

with \vec{r}_{lp} the position of the left proton and \vec{r}_{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.

3.5.2 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 3.8. 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 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 (3.22), but the distance should be measured



Figure 3.8: Hydrogen atom plus free proton far apart.

from the position \vec{r}_{lp} of the left proton instead of from the origin:

$$\psi = \psi_{100}(|\vec{r} - \vec{r}_{lp}|)$$

To shorten the notations, this wave function will be denoted by ψ_l :

$$\psi_l(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{lp}|) \quad (3.58)$$

Similarly the wave function that would describe the electron as being in the ground state around the right proton will be denoted as ψ_r , with

$$\psi_r(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{rp}|) \quad (3.59)$$

Key Points

- When the protons are far apart, there are two lowest energy states, ψ_l and ψ_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.

3.5.3 Energy when closer together



Figure 3.9: Hydrogen atom plus free proton 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 3.9. 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.

3.5.4 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,

$$\psi = a\psi_l + b\psi_r \quad (3.60)$$

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

$$\langle a\psi_l + b\psi_r | a\psi_l + b\psi_r \rangle = a^2 + b^2 + 2ab\langle \psi_l | \psi_r \rangle = 1 \quad (3.61)$$

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 3.10 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

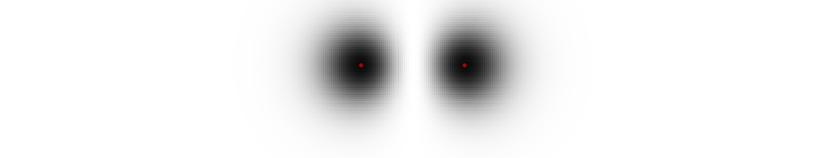


Figure 3.10: The electron being anti-symmetrically shared.

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 $|\psi_l|^2$. Similarly, the probability of finding the electron in the second state is given by $|\psi_r|^2$. But for the combined state nature does not do the logical thing of adding the two probabilities together to come up with $\frac{1}{2}|\psi_l|^2 + \frac{1}{2}|\psi_r|^2$.

Instead of adding physically *observable* probabilities, nature squares the *unobservable* wave function $a\psi_l - a\psi_r$ to find the new probability distribution. The squaring adds a cross term, $-2a^2\psi_l\psi_r$, 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:

$$\Psi \approx a [\psi_{100}(|\vec{r} - \vec{r}_{lp}|) + \psi_{100}(|\vec{r} - \vec{r}_{rp}|)] \quad \psi_{100}(r) \equiv \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0} \quad (3.62)$$

where $a_0 = 0.53 \text{ \AA}$ 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 3.11.



Figure 3.11: 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.

3.5.5 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, [15], 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.

3.5.6 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 $a\psi_l + b\psi_r$,

where ψ_l described the state where the electron was in the ground state around the left proton, and ψ_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, {A.22}.

Note that all that can be changed in the approximation $a\psi_l + b\psi_r$ 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 (3.61), 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 note {A.23}.

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.
-
-

3.5.6 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 2.5. 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 2.5.3, 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: $\int_0^\ell x(\ell - x) dx = \ell^3/6$ and $\int_0^\ell x^2(\ell - x)^2 dx = \ell^5/30$)

3.5.7 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.24}.

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 7.2.3 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.25}; 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.26}. 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 3.5.5, 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 4

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 functions 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.

4.1 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

$$\Psi(\vec{r}_1, \vec{r}_2; t) \tag{4.1}$$

depending on both particle positions. The value of $|\Psi(\vec{r}_1, \vec{r}_2; t)|^2 d^3\vec{r}_1 d^3\vec{r}_2$ gives the probability of simultaneously finding particle 1 within a vicinity $d^3\vec{r}_1$ of \vec{r}_1 and particle 2 within a vicinity $d^3\vec{r}_2$ of \vec{r}_2 .

The wave function must be normalized to express that the electrons must be somewhere:

$$\langle \Psi | \Psi \rangle_6 = \iint |\Psi(\vec{r}_1, \vec{r}_2; t)|^2 d^3\vec{r}_1 d^3\vec{r}_2 = 1 \tag{4.2}$$

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 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 note {A.16}.

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.

4.1 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 .
- 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.

4.2 The Hydrogen Molecule

This section uses similar approximations as for the hydrogen molecular ion of chapter 3.5 to examine the neutral H₂ 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.

4.2.1 The Hamiltonian

Just like for the hydrogen molecular ion of chapter 3.5, 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:

$$H = -\frac{\hbar^2}{2m_e} (\nabla_1^2 + \nabla_2^2) - \frac{e^2}{4\pi\epsilon_0} \left(\frac{1}{r_{1l}} + \frac{1}{r_{1r}} + \frac{1}{r_{2l}} + \frac{1}{r_{2r}} - \frac{1}{|\vec{r}_1 - \vec{r}_2|} \right) \quad (4.3)$$

In this expression, the Laplacians of the first two, kinetic energy, terms are with respect to the position coordinates of the two electrons:

$$\nabla_1^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2} \quad \nabla_2^2 = \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y_2^2} + \frac{\partial^2}{\partial z_2^2}$$

The next four terms in the Hamiltonian (4.3) are the attractive potentials between the electrons and the protons, with r_{1l} , r_{2l} , r_{1r} , and r_{2r} 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.

4.2.1 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.

- 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?

4.2.2 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 3.5, 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 ψ_l will again be used for the wave function of a *single* electron that in the ground state around the left proton and ψ_r for one that is in the ground state around the right hand one:

$$\psi_l(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{lp}|) \quad \psi_r(\vec{r}) \equiv \psi_{100}(|\vec{r} - \vec{r}_{rp}|)$$

where ψ_{100} is the hydrogen atom ground state (3.22), and \vec{r}_{lp} and \vec{r}_{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:

$$\psi(\vec{r}_1, \vec{r}_2) = \psi_l(\vec{r}_1)\psi_r(\vec{r}_2)$$

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:

$$|\psi_l(\vec{r}_1)|^2 d^3\vec{r}_1 \times |\psi_r(\vec{r}_2)| d^3\vec{r}_2$$

or in words:

$$\begin{aligned} & [\text{probability of finding 1 at } \vec{r}_1 \text{ unaffected by where 2 is}] \\ & \times [\text{probability of finding 2 at } \vec{r}_2 \text{ unaffected by where 1 is}] \end{aligned}$$

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 ψ_l and ψ_r .
 - The complete wave function for that case is $\psi_l(\vec{r}_1)\psi_r(\vec{r}_2)$, assuming that electron 1 is around the left proton and electron 2 around the right one.
-

4.2.2 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?
 - 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?
-

4.2.3 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 $d^3\vec{r}$ of the point. This probability is found as

$$n(\vec{r}) = \int |\Psi(\vec{r}, \vec{r}_2)|^2 d^3\vec{r}_2 + \int |\Psi(\vec{r}_1, \vec{r})|^2 d^3\vec{r}_1 \quad (4.4)$$

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 $d^3\vec{r}$ 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 $\psi_l(\vec{r}_1)\psi_r(\vec{r}_2)$ with electron 1 around the left proton and electron 2 around the right one is shown in figure 4.1. Of course the probability density for the state $\psi_r(\vec{r}_1)\psi_l(\vec{r}_2)$ with the electrons exchanged would look exactly the same.

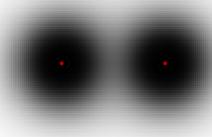


Figure 4.1: State with two neutral atoms.

Key Points

- The probability density is the probability per unit volume of finding an electron, whichever one, near a given point.

4.2.3 Review Questions

- 1 Suppose, given the wave function $\psi_l(\vec{r}_1)\psi_r(\vec{r}_2)$, 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?

4.2.4 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 $\psi_l(\vec{r}_1)\psi_r(\vec{r}_2)$ which describes that electron 1 is around the left proton and electron 2 around the right one will be indicated by $\psi_l\psi_r$, 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 $\psi_r\psi_l$:

$$\psi = a\psi_l\psi_r + b\psi_r\psi_l \quad (4.5)$$

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

$$\langle\psi|\psi\rangle_6 = \langle a\psi_l\psi_r + b\psi_r\psi_l | a\psi_l\psi_r + b\psi_r\psi_l \rangle = a^2 + b^2 + 2ab\langle\psi_l|\psi_r\rangle^2 = 1 \quad (4.6)$$

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:

$$n = \psi_l^2 + \psi_r^2 + 2ab\langle\psi_l|\psi_r\rangle \left\{ 2\psi_l\psi_r - \langle\psi_l|\psi_r\rangle(\psi_l^2 + \psi_r^2) \right\} \quad (4.7)$$

The most important combination state is the one with $b = a$:

$$\psi(\vec{r}_1, \vec{r}_2) = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] \quad (4.8)$$

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

$$\Psi \approx a [\psi_{100}(|\vec{r}_1 - \vec{r}_{lp}|)\psi_{100}(|\vec{r}_2 - \vec{r}_{rp}|) + \psi_{100}(|\vec{r}_1 - \vec{r}_{rp}|)\psi_{100}(|\vec{r}_2 - \vec{r}_{lp}|)] \quad (4.9)$$

with $\psi_{100}(r) \equiv e^{-r/a_0}/\sqrt{\pi a_0^3}$, where $a_0 = 0.53 \text{ \AA}$ is the Bohr radius, and \vec{r}_1 , \vec{r}_2 , \vec{r}_{1p} , and \vec{r}_{2p} are again the position vectors of the electrons and protons.

The probability density of this wave function looks like figure 4.2. It has increased likelihood for electrons to be found in between the protons, compared to figure 4.1 in which each proton had its own electron.

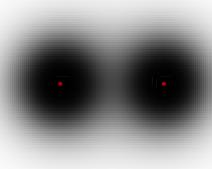


Figure 4.2: Symmetric sharing of the electrons.

The state with $b = -a$,

$$\psi(\vec{r}_1, \vec{r}_2) = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) - \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] \quad (4.10)$$

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 4.3, the antisymmetric state has decreased likelihood for electrons to be found in between the protons.

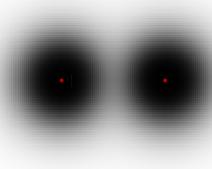


Figure 4.3: Antisymmetric sharing of the electrons.

Key Points

- In state $\psi_l\psi_r$, the electron numbered 1 is around the left proton and 2 around the right one.
- In state $\psi_r\psi_l$, the electron numbered 1 is around the right proton and 2 around the left one.

- In the symmetric state $a(\psi_l\psi_r + \psi_r\psi_l)$ 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 $a(\psi_l\psi_r - \psi_r\psi_l)$ 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.
-
-

4.2.4 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 6-dimensional wave function, but visualizing 6-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.

- 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?
-

4.2.5 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 3.5.6, 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 $a\psi_1\psi_r + b\psi_r\psi_1$; 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, {A.27}.

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.

4.2.6 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.25,A.26}.

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 10.13.) 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; {A.27}.

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 8.1, 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.

4.3 Two-State Systems

The protons in the H_2^+ hydrogen molecular ion of chapter 3.5 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. The main purpose of this section is to shed some light on the question why chemical bonds involving a single electron are 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, $\psi_1 = \psi_l$, described that the electron was around the left proton, the other, $\psi_2 = \psi_r$, that it was around the right one. For the hydrogen

molecule, $\psi_1 = \psi_l\psi_r$ had electron 1 around the left proton and electron 2 around the right one; $\psi_2 = \psi_r\psi_l$ was the same, except with 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.

The C₆H₆ “benzene molecular ring” consists of a hexagon of 6 carbon atoms that are held together by 9 covalent bonds. The way that the 9 bonds between the 6 atoms can be arranged is to make every second bond a double one. However, that still leaves two possibilities, by swapping the locations of the single and double bonds, hence two different states ψ_1 and ψ_2 .

The NH₃ “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 4.11.3, 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 4.11.3.) 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 states ψ_1 and ψ_2 .

In each case described above, there are two logical physical states ψ_1 and ψ_2 . The peculiarities of two state systems arise from states that are combinations of these two states, as in

$$\Psi = a\psi_1 + b\psi_2$$

Note that according to the ideas of quantum mechanics, the square magnitude of the first coefficient of the combined state, $|a|^2$, represents the probability of being in state ψ_1 and $|b|^2$ the probability of being in state ψ_2 . Of course, the total probability of being in one of the states should be one:

$$|a|^2 + |b|^2 = 1$$

(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 what combination of states has the lowest energy. The expectation value of energy is

$$\langle E \rangle = \langle a\psi_1 + b\psi_2 | H | a\psi_1 + b\psi_2 \rangle$$

This can be multiplied out as, (remember that numerical factors come out of the left of an inner product as complex conjugates,)

$$\langle E \rangle = a^*aH_{11} + a^*bH_{12} + b^*aH_{21} + b^*bH_{22}$$

where the shorthand notation

$$H_{11} = \langle \psi_1 | H \psi_1 \rangle, \quad H_{12} = \langle \psi_1 | H \psi_2 \rangle, \quad H_{21} = \langle \psi_2 | H \psi_1 \rangle, \quad H_{22} = \langle \psi_2 | H \psi_2 \rangle$$

was used. Note that H_{11} and H_{22} are real, (1.16), and the states will be ordered so that H_{11} is less or equal to H_{22} . Normally, H_{12} and H_{21} are not real but complex conjugates, (1.16), but you can always change the definition of, say, ψ_1 by a 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. Also note that $a^*a = |a|^2$ and $b^*b = |b|^2$.

The above expression for the expectation energy consists of two kinds of terms, which will be called:

$$\text{the averaged energy: } |a|^2 H_{11} + |b|^2 H_{22} \quad (4.11)$$

$$\text{the twilight terms: } (a^*b + b^*a) H_{12} \quad (4.12)$$

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 average of the 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 lower than the energy H_{11} of the lowest component state. (To see that, just take a and b positive real numbers and b small enough that b^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 and have the same energy. This is the case for the hydrogen examples and most of the others mentioned. For such states, the ground state will occur for an *equal* mixture of states, $a = b = \sqrt{\frac{1}{2}}$, because then the twilight terms are most negative. In that case, the lowest energy, call it E_L , is an amount H_{12} below the energy $H_{11} = H_{22}$ of the component states.

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 for $|a| \approx 1$ and $|b| \approx 0$. (This assumes that the twilight terms are not big enough to dominate the energy.) In that case $ab \approx 0$, which pretty much takes the twilight terms (4.12) 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 be 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 forge 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 polarization of the hydrogen atom by the lithium ion, instead of as 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 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 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 ring that the nitrogen atom is at. Since these two states are physically equivalent, there is again a significant lowering of the energy E_L for the symmetric combination $a = b$. Similarly, there is a significant raising of the energy E_H for the antisymmetric combination $a = -b$. 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, and 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 6.3.2 for the operating principle of lasers.

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. The equivalent states here merely have the identical electrons interchanged.

4.3 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

$$\bar{\psi}_1 = \alpha (\psi_1 - \varepsilon \psi_2) \quad \bar{\psi}_2 = \alpha (\psi_2 - \varepsilon \psi_1).$$

For normalized ψ_1 and ψ_2 the Cauchy-Schwartz inequality says 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^{ic} and be^{-ic} , with a , b , and c real, to get states that do.)

- 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.

4.4 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 be referred to as “orbital” angular momentum. As was discussed in chapter 3.1, 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:

$$S^2 = s(s+1)\hbar^2 \quad (4.13)$$

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

$$S_z = m\hbar \quad (4.14)$$

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 chapter 12.2.3.)

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 $\frac{1}{2}$, which allows baryons to have spin $\frac{3}{2}$ or $\frac{1}{2}$. (It is not self-evident, but spin values can be additive or subtractive within the confines of their discrete allowable values; see chapter 10.1.) 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 $\frac{1}{2}$. 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 $|1/2\ 1/2\rangle$ or \uparrow . Similarly, the spin-down state is indicated by $|1/2\ -1/2\rangle$ or \downarrow .

4.4 Review Questions

- 1 Delta particles have spin $s = \frac{3}{2}$. What values can their spin angular momentum in a given direction have?

- 2** Delta particles have spin $\frac{3}{2}$. What value does their total square angular momentum have?
-

4.5 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.

4.5.1 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:

$$\boxed{\Psi \equiv \Psi(\vec{r}, S_z; t)} \quad (4.15)$$

The value of $|\Psi(\vec{r}, S_z; t)|^2 d^3\vec{r}$ gives the probability of finding the particle within a vicinity $d^3\vec{r}$ 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:

$$\boxed{\Psi_+(\vec{r}; t) \equiv \Psi(\vec{r}, \frac{1}{2}\hbar; t) \quad \text{and} \quad \Psi_-(\vec{r}; t) \equiv \Psi(\vec{r}, -\frac{1}{2}\hbar; t)} \quad (4.16)$$

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 10.2 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 note {A.96}.) 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:

$$\Psi(\vec{r}, S_z; t) = \Psi_+(\vec{r}; t)\uparrow(S_z) + \Psi_-(\vec{r}; t)\downarrow(S_z). \quad (4.17)$$

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{i} and \hat{j} in normal vector analysis; they have by definition the following values:

$$\uparrow(\frac{1}{2}\hbar) = 1 \quad \uparrow(-\frac{1}{2}\hbar) = 0 \quad \downarrow(\frac{1}{2}\hbar) = 0 \quad \downarrow(-\frac{1}{2}\hbar) = 1.$$

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 $s = \frac{1}{2}$ 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

$$\uparrow(\frac{1}{2}\hbar) = 1 \quad \uparrow(-\frac{1}{2}\hbar) = 0 \quad \downarrow(\frac{1}{2}\hbar) = 0 \quad \downarrow(-\frac{1}{2}\hbar) = 1$$

and represent the pure spin-up, respectively spin-down states.

4.5.1 Review Questions

- 1 What is the normalization requirement of the wave function of a spin $\frac{1}{2}$ particle in terms of Ψ_+ and Ψ_- ?
-

4.5.2 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

$$\langle \Psi_1 | \Psi_2 \rangle = \sum_{S_z=\pm\frac{1}{2}\hbar} \int_{\text{all } \vec{r}} \Psi_1^*(\vec{r}, S_z; t) \Psi_2(\vec{r}, S_z; t) d^3\vec{r}$$

or writing out the two-term sum,

$$\langle \Psi_1 | \Psi_2 \rangle = \int_{\text{all } \vec{r}} \Psi_1^*(\vec{r}, \frac{1}{2}\hbar; t) \Psi_2(\vec{r}, \frac{1}{2}\hbar; t) d^3\vec{r} + \int_{\text{all } \vec{r}} \Psi_1^*(\vec{r}, -\frac{1}{2}\hbar; t) \Psi_2(\vec{r}, -\frac{1}{2}\hbar; t) d^3\vec{r}$$

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:

$$\langle \Psi_1 | \Psi_2 \rangle = \int_{\text{all } \vec{r}} \Psi_{1+}^*(\vec{r}; t) \Psi_{2+}(\vec{r}; t) d^3\vec{r} + \int_{\text{all } \vec{r}} \Psi_{1-}^*(\vec{r}; t) \Psi_{2-}(\vec{r}; t) d^3\vec{r}$$

In other words, the inner product with spin evaluates as

$$\boxed{\langle \Psi_{1+} \uparrow + \Psi_{1-} \downarrow | \Psi_{2+} \uparrow + \Psi_{2-} \downarrow \rangle = \langle \Psi_{1+} | \Psi_{2+} \rangle + \langle \Psi_{1-} | \Psi_{2-} \rangle} \quad (4.18)$$

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,

$$\boxed{\langle \uparrow | \uparrow \rangle = 1 \quad \langle \uparrow | \downarrow \rangle = \langle \downarrow | \uparrow \rangle = 0 \quad \langle \downarrow | \downarrow \rangle = 1} \quad (4.19)$$

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:

$$\begin{aligned} & \langle \Psi_{1+}\uparrow + \Psi_{1-}\downarrow | \Psi_{2+}\uparrow + \Psi_{2-}\downarrow \rangle \\ &= \langle \Psi_{1+} | \Psi_{2+} \rangle \langle \uparrow | \uparrow \rangle + \langle \Psi_{1+} | \Psi_{2-} \rangle \langle \uparrow | \downarrow \rangle + \langle \Psi_{1-} | \Psi_{2+} \rangle \langle \downarrow | \uparrow \rangle + \langle \Psi_{1-} | \Psi_{2-} \rangle \langle \downarrow | \downarrow \rangle \\ &= \langle \Psi_{1+} | \Psi_{2+} \rangle + \langle \Psi_{1-} | \Psi_{2-} \rangle \end{aligned}$$

Key Points

o In inner products, you must sum over the spin states.

o For spin $\frac{1}{2}$ 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.

o The spin-up and spin-down states \uparrow and \downarrow are an orthonormal pair.

4.5.2 Review Questions

- 1 Show that the normalization requirement for the wave function of a spin $\frac{1}{2}$ particle in terms of Ψ_+ and Ψ_- requires its norm $\sqrt{\langle \Psi | \Psi \rangle}$ to be one.
 - 2 Show that if ψ_l and ψ_r are normalized spatial wave functions, then a combination like $(\psi_l\uparrow + \psi_r\downarrow) / \sqrt{2}$ is a normalized wave function with spin.
-

4.5.3 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:

$$[\hat{S}_x, \hat{S}_y] = i\hbar \hat{S}_z \quad [\hat{S}_y, \hat{S}_z] = i\hbar \hat{S}_x \quad [\hat{S}_z, \hat{S}_x] = i\hbar \hat{S}_y \quad (4.20)$$

These equations are called the “fundamental commutation relations.” As will be shown in chapter 10.1, 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 \hat{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 \hat{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 \hat{S}_z only affects the variation in this additional direction. For example, for a particle with spin one half, \hat{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 \hat{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.

4.5.3 Review Questions

- 1 Are not some commutators missing from the fundamental commutation relationship? For example, what is the commutator $[\hat{S}_y, \hat{S}_x]$?

4.5.4 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

$$\boxed{\Psi = \Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t)} \quad (4.21)$$

The value of $|\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t)|^2 d^3\vec{r}_1 d^3\vec{r}_2$ gives the probability of simultaneously finding particle 1 within a vicinity $d^3\vec{r}_1$ of \vec{r}_1 with spin angular momentum in the z -direction S_{z1} , and particle 2 within a vicinity $d^3\vec{r}_2$ 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

$$\boxed{\begin{aligned}\Psi_{++}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, +\frac{1}{2}\hbar, \vec{r}_2, +\frac{1}{2}\hbar; t) \\ \Psi_{+-}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, +\frac{1}{2}\hbar, \vec{r}_2, -\frac{1}{2}\hbar; t) \\ \Psi_{-+}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, -\frac{1}{2}\hbar, \vec{r}_2, +\frac{1}{2}\hbar; t) \\ \Psi_{--}(\vec{r}_1, \vec{r}_2; t) &\equiv \Psi(\vec{r}_1, -\frac{1}{2}\hbar, \vec{r}_2, -\frac{1}{2}\hbar; t)\end{aligned}} \quad (4.22)$$

For example, $|\Psi_{+-}(\vec{r}_1, \vec{r}_2; t)|^2 d^3\vec{r}_1 d^3\vec{r}_2$ gives the probability of finding particle 1 within a vicinity $d^3\vec{r}_1$ of \vec{r}_1 with spin up, and particle 2 within a vicinity $d^3\vec{r}_2$ of \vec{r}_2 with spin down.

The wave function can be written using purely spatial functions and purely spin functions as

$$\begin{aligned}\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) &= \Psi_{++}(\vec{r}_1, \vec{r}_2; t) \uparrow(S_{z1}) \uparrow(S_{z2}) + \Psi_{+-}(\vec{r}_1, \vec{r}_2; t) \uparrow(S_{z1}) \downarrow(S_{z2}) \\ &\quad + \Psi_{-+}(\vec{r}_1, \vec{r}_2; t) \downarrow(S_{z1}) \uparrow(S_{z2}) + \Psi_{--}(\vec{r}_1, \vec{r}_2; t) \downarrow(S_{z1}) \downarrow(S_{z2})\end{aligned}$$

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

$$\langle \Psi_1 | \Psi_2 \rangle =$$

$$\sum_{S_{z1}=\pm\frac{1}{2}\hbar} \sum_{S_{z2}=\pm\frac{1}{2}\hbar} \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} \Psi_1^*(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) \Psi_2(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) d^3\vec{r}_1 d^3\vec{r}_2$$

This can be written in terms of the purely spatial components as

$$\boxed{\langle \Psi_1 | \Psi_2 \rangle = \langle \Psi_{1++} | \Psi_{2++} \rangle + \langle \Psi_{1+-} | \Psi_{2+-} \rangle + \langle \Psi_{1-+} | \Psi_{2-+} \rangle + \langle \Psi_{1--} | \Psi_{2--} \rangle} \quad (4.23)$$

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.
-
-

4.5.4 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} .
- 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:

$$\int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} f(\vec{r}_1)g(\vec{r}_2) d^3\vec{r}_1 d^3\vec{r}_2 =$$

$$\left[\int_{\text{all } \vec{r}_1} f(\vec{r}_1) d^3\vec{r}_1 \right] \left[\int_{\text{all } \vec{r}_2} g(\vec{r}_2) d^3\vec{r}_2 \right]$$

4.5.5 Example: the hydrogen molecule

As an example, this section considers the ground state of the hydrogen molecule. It was found in section 4.2 that the ground state electron wave function must be of the approximate form

$$\psi_{\text{gs},0} = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)]$$

where ψ_l was the electron ground state of the left hydrogen atom, and ψ_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

$$\psi_{\text{gs}} = \psi_{++}\uparrow\uparrow + \psi_{+-}\uparrow\downarrow + \psi_{-+}\downarrow\uparrow + \psi_{--}\downarrow\downarrow.$$

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 $\psi_{\text{gs},0}$ above. Anything else would have more than the lowest possible energy, {A.28}.

So the approximate ground state including spin must take the form

$$\psi_{\text{gs}} = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] [a_{++}\uparrow\uparrow + a_{+-}\uparrow\downarrow + a_{-+}\downarrow\uparrow + a_{--}\downarrow\downarrow] \quad (4.24)$$

where a_{++} , a_{+-} , a_{-+} , and a_{--} are constants.

Key Points

- The electron wave function $\psi_{\text{gs},0}$ for the hydrogen molecule derived previously ignored spin.
- In the full electron wave function, each spatial component must separately be proportional to $a(\psi_l\psi_r + \psi_r\psi_l)$.

4.5.5 Review Questions

- 1 Show that the normalization requirement for ψ_{gs} means that

$$|a_{++}|^2 + |a_{+-}|^2 + |a_{-+}|^2 + |a_{--}|^2 = 1$$

4.5.6 Triplet and singlet states

In the case of two particles with spin $1/2$, it is often more convenient to use slightly different basis states to describe the spin states than the four arrow

combinations $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, and $\downarrow\downarrow$. The more convenient basis states can be written in $|s m\rangle$ ket notation, and they are:

$$\boxed{\begin{array}{lll} |1 1\rangle = \uparrow\uparrow & |1 0\rangle = \frac{1}{\sqrt{2}}(\uparrow\downarrow + \downarrow\uparrow) & |1 -1\rangle = \downarrow\downarrow \\ \underbrace{\hspace{10em}}_{\text{the triplet states}} & & \underbrace{\hspace{10em}}_{\text{the singlet state}} \\ & & |0 0\rangle = \frac{1}{\sqrt{2}}(\uparrow\downarrow - \downarrow\uparrow) \end{array}} \quad (4.25)$$

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 $\uparrow\downarrow$ and $\downarrow\uparrow$ states can be written as

$$\uparrow\downarrow = \frac{1}{\sqrt{2}}(|1 0\rangle + |0 0\rangle) \quad \downarrow\uparrow = \frac{1}{\sqrt{2}}(|1 0\rangle - |0 0\rangle)$$

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 $\uparrow\downarrow$ and $\downarrow\uparrow$ cannot be written in $|s m\rangle$ ket notation; there is no value for s .

Incidentally, note that z -components of angular momentum simply add up, as the Newtonian analogy suggests. For example, for $\uparrow\downarrow$, 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 10.1 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 $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, and $\downarrow\downarrow$ 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.

4.5.6 Review Questions

- 1 Like the states $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, and $\downarrow\downarrow$; 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.

4.6 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:

$$\Psi(\vec{r}_1, S_{z1}, \dots, \vec{r}_i, S_{zi}, \dots, \vec{r}_j, S_{zj}, \dots) = \Psi(\vec{r}_1, S_{z1}, \dots, \vec{r}_j, S_{zj}, \dots, \vec{r}_i, S_{zi}, \dots) \quad (4.26)$$

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:

$$\Psi(\vec{r}_1, S_{z1}, \dots, \vec{r}_i, S_{zi}, \dots, \vec{r}_j, S_{zj}, \dots) = -\Psi(\vec{r}_1, S_{z1}, \dots, \vec{r}_j, S_{zj}, \dots, \vec{r}_i, S_{zi}, \dots) \quad (4.27)$$

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

$$\psi_{\text{gs}} = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] [a_{++}\uparrow\uparrow + a_{+-}\uparrow\downarrow + a_{-+}\downarrow\uparrow + a_{--}\downarrow\downarrow] \quad (4.28)$$

were ψ_l was the ground state of the left hydrogen atom, ψ_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

$$-\psi_{\text{gs}} = a [\psi_l(\vec{r}_2)\psi_r(\vec{r}_1) + \psi_r(\vec{r}_2)\psi_l(\vec{r}_1)] [a_{++}\uparrow\uparrow + a_{+-}\downarrow\uparrow + a_{-+}\uparrow\downarrow + a_{--}\downarrow\downarrow];$$

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 (4.28), reorder the terms back to the same order:

$$-\psi_{\text{gs}} = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] [a_{++}\uparrow\uparrow + a_{-+}\uparrow\downarrow + a_{+-}\downarrow\uparrow + a_{--}\downarrow\downarrow]$$

The spatial factor is seen to be the same as the nonexchanged version in (4.28); 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,

$$\psi_{\text{gs}} = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] a_{+-} [\uparrow\downarrow - \downarrow\uparrow]$$

or after normalization, noting that a factor of magnitude one is always arbitrary,

$$\psi_{\text{gs}} = a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] \frac{\uparrow\downarrow - \downarrow\uparrow}{\sqrt{2}}$$

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 13.3 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.
-

4.6 Review Questions

- 1 Check that indeed any linear combination of the triplet states is unchanged under particle exchange.
- 2 Suppose the electrons of the hydrogen molecule are in the excited antisymmetric spatial state

$$a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) - \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)].$$

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.

4.7 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”,

$$\psi_1^P(\vec{r}, S_z), \psi_2^P(\vec{r}, S_z), \dots$$

An example is provided by the approximate ground state of the hydrogen molecule from the previous section,

$$a [\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)] \frac{\uparrow\downarrow - \downarrow\uparrow}{\sqrt{2}}.$$

This can be multiplied out to be

$$\frac{a}{\sqrt{2}} \left[\psi_l(\vec{r}_1) \uparrow(S_{z1}) \psi_r(\vec{r}_2) \downarrow(S_{z2}) + \psi_r(\vec{r}_1) \uparrow(S_{z1}) \psi_l(\vec{r}_2) \downarrow(S_{z2}) \right. \\ \left. - \psi_l(\vec{r}_1) \downarrow(S_{z1}) \psi_r(\vec{r}_2) \uparrow(S_{z2}) - \psi_r(\vec{r}_1) \downarrow(S_{z1}) \psi_l(\vec{r}_2) \uparrow(S_{z2}) \right]$$

and consists of four single-particle functions:

$$\begin{aligned} \psi_1^P(\vec{r}, S_z) &= \psi_l(\vec{r}) \uparrow(S_z) & \psi_2^P(\vec{r}, S_z) &= \psi_l(\vec{r}) \downarrow(S_z) \\ \psi_3^P(\vec{r}, S_z) &= \psi_r(\vec{r}) \uparrow(S_z) & \psi_4^P(\vec{r}, S_z) &= \psi_r(\vec{r}) \downarrow(S_z). \end{aligned}$$

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

$$\frac{a}{\sqrt{2}} \psi_1^P(\vec{r}_1, S_{z1}) \psi_4^P(\vec{r}_2, S_{z2}) + \frac{a}{\sqrt{2}} \psi_3^P(\vec{r}_1, S_{z1}) \psi_2^P(\vec{r}_2, S_{z2}) \\ - \frac{a}{\sqrt{2}} \psi_2^P(\vec{r}_1, S_{z1}) \psi_3^P(\vec{r}_2, S_{z2}) - \frac{a}{\sqrt{2}} \psi_4^P(\vec{r}_1, S_{z1}) \psi_1^P(\vec{r}_2, S_{z2})$$

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:

$$\begin{aligned} \Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) = & a_{11} \psi_1^P(\vec{r}_1, S_{z1}) \psi_1^P(\vec{r}_2, S_{z2}) + a_{12} \psi_1^P(\vec{r}_1, S_{z1}) \psi_2^P(\vec{r}_2, S_{z2}) + \\ & a_{13} \psi_1^P(\vec{r}_1, S_{z1}) \psi_3^P(\vec{r}_2, S_{z2}) + a_{14} \psi_1^P(\vec{r}_1, S_{z1}) \psi_4^P(\vec{r}_2, S_{z2}) + \\ & a_{21} \psi_2^P(\vec{r}_1, S_{z1}) \psi_1^P(\vec{r}_2, S_{z2}) + a_{22} \psi_2^P(\vec{r}_1, S_{z1}) \psi_2^P(\vec{r}_2, S_{z2}) + \\ & a_{23} \psi_2^P(\vec{r}_1, S_{z1}) \psi_3^P(\vec{r}_2, S_{z2}) + a_{24} \psi_2^P(\vec{r}_1, S_{z1}) \psi_4^P(\vec{r}_2, S_{z2}) + \\ & a_{31} \psi_3^P(\vec{r}_1, S_{z1}) \psi_1^P(\vec{r}_2, S_{z2}) + a_{32} \psi_3^P(\vec{r}_1, S_{z1}) \psi_2^P(\vec{r}_2, S_{z2}) + \\ & a_{33} \psi_3^P(\vec{r}_1, S_{z1}) \psi_3^P(\vec{r}_2, S_{z2}) + a_{34} \psi_3^P(\vec{r}_1, S_{z1}) \psi_4^P(\vec{r}_2, S_{z2}) + \\ & a_{41} \psi_4^P(\vec{r}_1, S_{z1}) \psi_1^P(\vec{r}_2, S_{z2}) + a_{42} \psi_4^P(\vec{r}_1, S_{z1}) \psi_2^P(\vec{r}_2, S_{z2}) + \\ & a_{43} \psi_4^P(\vec{r}_1, S_{z1}) \psi_3^P(\vec{r}_2, S_{z2}) + a_{44} \psi_4^P(\vec{r}_1, S_{z1}) \psi_4^P(\vec{r}_2, S_{z2}) \end{aligned}$$

This can be written much more concisely using summation indices as

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) = \sum_{n_1=1}^4 \sum_{n_2=1}^4 a_{n_1 n_2} \psi_{n_1}^P(\vec{r}_1, S_{z1}) \psi_{n_2}^P(\vec{r}_2, S_{z2})$$

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:

$$\begin{aligned}
 \text{I : } & a_{11}\psi_1^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2}) \\
 \text{II : } & a_{22}\psi_2^P(\vec{r}_1, S_{z1})\psi_2^P(\vec{r}_2, S_{z2}) \\
 \text{III : } & a_{33}\psi_3^P(\vec{r}_1, S_{z1})\psi_3^P(\vec{r}_2, S_{z2}) \\
 \text{IV : } & a_{44}\psi_4^P(\vec{r}_1, S_{z1})\psi_4^P(\vec{r}_2, S_{z2}) \\
 \text{V : } & a_{12}\psi_1^P(\vec{r}_1, S_{z1})\psi_2^P(\vec{r}_2, S_{z2}) + a_{21}\psi_2^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2}) \\
 \text{VI : } & a_{13}\psi_1^P(\vec{r}_1, S_{z1})\psi_3^P(\vec{r}_2, S_{z2}) + a_{31}\psi_3^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2}) \\
 \text{VII : } & a_{14}\psi_1^P(\vec{r}_1, S_{z1})\psi_4^P(\vec{r}_2, S_{z2}) + a_{41}\psi_4^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2}) \\
 \text{VIII : } & a_{23}\psi_2^P(\vec{r}_1, S_{z1})\psi_3^P(\vec{r}_2, S_{z2}) + a_{32}\psi_3^P(\vec{r}_1, S_{z1})\psi_2^P(\vec{r}_2, S_{z2}) \\
 \text{IX : } & a_{24}\psi_2^P(\vec{r}_1, S_{z1})\psi_4^P(\vec{r}_2, S_{z2}) + a_{42}\psi_4^P(\vec{r}_1, S_{z1})\psi_2^P(\vec{r}_2, S_{z2}) \\
 \text{X : } & a_{34}\psi_3^P(\vec{r}_1, S_{z1})\psi_4^P(\vec{r}_2, S_{z2}) + a_{43}\psi_4^P(\vec{r}_1, S_{z1})\psi_3^P(\vec{r}_2, S_{z2})
 \end{aligned}$$

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:

$$\begin{aligned}
 \Psi = & a_{12} \left| \begin{array}{cc} \psi_1^P(\vec{r}_1, S_{z1}) & \psi_2^P(\vec{r}_1, S_{z1}) \\ \psi_1^P(\vec{r}_2, S_{z2}) & \psi_2^P(\vec{r}_2, S_{z2}) \end{array} \right| + a_{13} \left| \begin{array}{cc} \psi_1^P(\vec{r}_1, S_{z1}) & \psi_3^P(\vec{r}_1, S_{z1}) \\ \psi_1^P(\vec{r}_2, S_{z2}) & \psi_3^P(\vec{r}_2, S_{z2}) \end{array} \right| + \\
 & a_{14} \left| \begin{array}{cc} \psi_1^P(\vec{r}_1, S_{z1}) & \psi_4^P(\vec{r}_1, S_{z1}) \\ \psi_1^P(\vec{r}_2, S_{z2}) & \psi_4^P(\vec{r}_2, S_{z2}) \end{array} \right| + a_{23} \left| \begin{array}{cc} \psi_2^P(\vec{r}_1, S_{z1}) & \psi_3^P(\vec{r}_1, S_{z1}) \\ \psi_2^P(\vec{r}_2, S_{z2}) & \psi_3^P(\vec{r}_2, S_{z2}) \end{array} \right| + \\
 & a_{24} \left| \begin{array}{cc} \psi_2^P(\vec{r}_1, S_{z1}) & \psi_4^P(\vec{r}_1, S_{z1}) \\ \psi_2^P(\vec{r}_2, S_{z2}) & \psi_4^P(\vec{r}_2, S_{z2}) \end{array} \right| + a_{34} \left| \begin{array}{cc} \psi_3^P(\vec{r}_1, S_{z1}) & \psi_4^P(\vec{r}_1, S_{z1}) \\ \psi_3^P(\vec{r}_2, S_{z2}) & \psi_4^P(\vec{r}_2, S_{z2}) \end{array} \right|
 \end{aligned}$$

$$a_{24} \begin{vmatrix} \psi_2^P(\vec{r}_1, S_{z1}) & \psi_4^P(\vec{r}_1, S_{z1}) \\ \psi_2^P(\vec{r}_2, S_{z2}) & \psi_4^P(\vec{r}_2, S_{z2}) \end{vmatrix} + a_{34} \begin{vmatrix} \psi_3^P(\vec{r}_1, S_{z1}) & \psi_4^P(\vec{r}_1, S_{z1}) \\ \psi_3^P(\vec{r}_2, S_{z2}) & \psi_4^P(\vec{r}_2, S_{z2}) \end{vmatrix}$$

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 4.2, the coefficients a_{13} and a_{24} must be zero since their Slater determinants produce the excited antisymmetric spatial state $\psi_1\psi_r - \psi_r\psi_1$ 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 $\psi_1\psi_r + \psi_r\psi_1$ 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 $\psi_1^P(\vec{r}, S_z), \psi_2^P(\vec{r}, S_z), \dots, \psi_N^P(\vec{r}, S_z)$ are used to describe I , instead of 2, particles. Then the most general possible wave

function assumes the form:

$$\Psi = \sum_{n_1=1}^N \sum_{n_2=1}^N \dots \sum_{n_I=1}^N a_{n_1 n_2 \dots n_I} \psi_{n_1}^P(\vec{r}_1, S_{z1}) \psi_{n_2}^P(\vec{r}_2, S_{z2}) \dots \psi_{n_I}^P(\vec{r}_I, S_{zI}) \quad (4.29)$$

where the $a_{n_1 n_2 \dots n_I}$ 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 $a_{n_1 n_2 \dots n_I}$ 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 ψ_1^P consists of the single term

$$a_{11\dots 1} \psi_1^P(\vec{r}_1, S_{z1}) \psi_1^P(\vec{r}_2, S_{z2}) \dots \psi_1^P(\vec{r}_I, S_{zI}).$$

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 ψ_n^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 $\psi_{n_1}^p, \psi_{n_2}^p, \dots, \psi_{n_I}^p$ reduces under the antisymmetrization requirement to a single Slater determinant of the form

$$\frac{1}{\sqrt{I!}} \begin{vmatrix} \psi_{n_1}^p(\vec{r}_1, S_{z1}) & \psi_{n_2}^p(\vec{r}_1, S_{z1}) & \psi_{n_3}^p(\vec{r}_1, S_{z1}) & \cdots & \psi_{n_I}^p(\vec{r}_1, S_{z1}) \\ \psi_{n_1}^p(\vec{r}_2, S_{z2}) & \psi_{n_2}^p(\vec{r}_2, S_{z2}) & \psi_{n_3}^p(\vec{r}_2, S_{z2}) & \cdots & \psi_{n_I}^p(\vec{r}_2, S_{z2}) \\ \psi_{n_1}^p(\vec{r}_3, S_{z3}) & \psi_{n_2}^p(\vec{r}_3, S_{z3}) & \psi_{n_3}^p(\vec{r}_3, S_{z3}) & \cdots & \psi_{n_I}^p(\vec{r}_3, S_{z3}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_{n_1}^p(\vec{r}_I, S_{zI}) & \psi_{n_2}^p(\vec{r}_I, S_{zI}) & \psi_{n_3}^p(\vec{r}_I, S_{zI}) & \cdots & \psi_{n_I}^p(\vec{r}_I, S_{zI}) \end{vmatrix} \quad (4.30)$$

multiplied by a single unknown coefficient. The normalization factor $1/\sqrt{I!}$ has been thrown in merely to ensure that if the functions ψ_n^p 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 $a_{n_1 n_2 \dots n_I}$ to just one, $a_{12\dots I}$; 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 (4.29). Even after reduction to Slater determinants, there would still be a prohibitive $3 \cdot 10^{26}$ or so unknown coefficients left. The precise expression for the number of Slater determinants is called “ N choose I ,” it is given by

$$\binom{N}{I} = \frac{N!}{(N-I)!I!} = \frac{N(N-1)(N-2)\dots(N-I+1)}{I!},$$

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. {A.29}.

The basic “Hartree-Fock” approach, discussed in chapter 7.3, 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.

4.7 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?
- 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?

4.8 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 (4.29) 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

$$\Psi = a_1 \psi_1^S(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_I, S_{zI}) + a_2 \psi_2^S(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_I, S_{zI}) + \dots$$

or in index notation:

$$\Psi = \sum_{q=1}^Q a_q \psi_q^S(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_I, S_{zI}). \quad (4.31)$$

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 ψ_q^S are allowed to be anything; individual (Hartree) products of single-particle wave functions for distinguishable particles as in (4.29), 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 $\psi_n^p(\vec{r})$ 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:

$$\sum_{q=1}^Q H a_q \psi_q^S = \sum_{q=1}^Q E a_q \psi_q^S$$

The trick is now to take the inner product of both sides of this equation with each function ψ_q^S in the set of wave functions in turn. In other words, take an inner product with $\langle \psi_1^S |$ to get one equation, then take an inner product with $\langle \psi_2^S |$ 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,

$$\begin{aligned} H_{11}a_1 + H_{12}a_2 + \dots + H_{1Q}a_Q &= Ea_1 \\ H_{21}a_1 + H_{22}a_2 + \dots + H_{2Q}a_Q &= Ea_2 \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ H_{\underline{q}1}a_1 + H_{\underline{q}2}a_2 + \dots + H_{\underline{q}Q}a_Q &= Ea_{\underline{q}} \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ H_{Q1}a_1 + H_{Q2}a_2 + \dots + H_{QQ}a_Q &= Ea_Q \end{aligned}$$

where

$$H_{11} = \langle \psi_1^S | H \psi_1^S \rangle, \quad H_{12} = \langle \psi_1^S | H \psi_2^S \rangle, \quad \dots, \quad H_{QQ} = \langle \psi_Q^S | H \psi_Q^S \rangle.$$

are the matrix coefficients, or Hamiltonian coefficients.

This can again be written more compactly in index notation:

$$\sum_{q=1}^Q H_{\underline{q}\underline{q}} a_{\underline{q}} = Ea_{\underline{q}} \quad \text{for } \underline{q} = 1, 2, \dots, Q \quad \text{with } H_{\underline{q}\underline{q}} = \langle \psi_{\underline{q}}^S | H \psi_{\underline{q}}^S \rangle \quad (4.32)$$

which is just a finite-size matrix eigenvalue problem.

Since the functions ψ_q^S are known, chosen, functions, and the Hamiltonian H is also known, the matrix coefficients $H_{\underline{q}\underline{q}}$ can be determined. The eigenvalues E and corresponding eigenvectors (a_1, a_2, \dots) can then be found using linear algebra procedures. Each eigenvector produces a corresponding approximate eigenfunction $a_1\psi_1^S + a_2\psi_2^S + \dots$ 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.

4.8 Review Questions

- 1 As a relatively simple example, work out the above ideas for the $Q = 2$ hydrogen molecule spatial states $\psi_1^S = \psi_l \psi_r$ and $\psi_2^S = \psi_l \psi_r$. Write the matrix eigenvalue problem and identify the two eigenvalues and eigenvectors. Compare with the results of section 4.3.

Assume that ψ_l and ψ_r have been slightly adjusted to be orthonormal. Then so are ψ_1^S and ψ_2^S orthonormal, since the various six-dimensional

inner product integrals, like

$$\langle \psi_1^S | \psi_2^S \rangle \equiv \langle \psi_l \psi_r | \psi_r \psi_l \rangle \equiv \\ \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} \psi_l(\vec{r}_1) \psi_r(\vec{r}_2) \psi_r(\vec{r}_1) \psi_l(\vec{r}_2) d^3 \vec{r}_1 d^3 \vec{r}_2$$

can according to the rules of calculus be factored into three-dimensional integrals as

$$\begin{aligned} & \langle \psi_1^S | \psi_2^S \rangle \\ &= \left[\int_{\text{all } \vec{r}_1} \psi_l(\vec{r}_1) \psi_r(\vec{r}_1) d^3 \vec{r}_1 \right] \left[\int_{\text{all } \vec{r}_2} \psi_r(\vec{r}_2) \psi_l(\vec{r}_2) d^3 \vec{r}_2 \right] \\ &= \langle \psi_l | \psi_r \rangle \langle \psi_r | \psi_l \rangle \end{aligned}$$

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 4.2 noted, that can only be done numerically. Instead just refer to H_{11} as J and to H_{12} as $-L$:

$$H_{11} \equiv \langle \psi_1^S | H \psi_1^S \rangle \equiv \langle \psi_l \psi_r | H \psi_l \psi_r \rangle \equiv J$$

$$H_{12} \equiv \langle \psi_1^S | H \psi_2^S \rangle \equiv \langle \psi_l \psi_r | H \psi_r \psi_l \rangle \equiv -L.$$

Next note that you also have

$$H_{22} \equiv \langle \psi_2^S | H \psi_2^S \rangle \equiv \langle \psi_r \psi_l | H \psi_r \psi_l \rangle = J$$

$$H_{21} \equiv \langle \psi_2^S | H \psi_1^S \rangle \equiv \langle \psi_r \psi_l | H \psi_l \psi_r \rangle = -L$$

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 $\psi_l \psi_r$ or $\psi_r \psi_l$.

- 2** Find the eigenstates for the same problem, but now including spin. As section 4.7 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

$$\begin{aligned} \psi_1^S &= \frac{\psi_l \psi_r \uparrow\uparrow - \psi_r \psi_l \uparrow\uparrow}{\sqrt{2}} & \psi_2^S &= \frac{\psi_l \psi_r \uparrow\downarrow - \psi_r \psi_l \uparrow\downarrow}{\sqrt{2}} \\ \psi_3^S &= \frac{\psi_l \psi_r \downarrow\uparrow - \psi_r \psi_l \downarrow\uparrow}{\sqrt{2}} & \psi_4^S &= \frac{\psi_l \psi_r \downarrow\downarrow - \psi_r \psi_l \downarrow\downarrow}{\sqrt{2}} \end{aligned}$$

Note that the Hamiltonian does not involve spin, to the approximation used in most of this book, so that, following the techniques of

section 4.5, an inner product like $H_{23} = \langle \psi_2^S | H \psi_3^S \rangle$ can be written out like

$$\begin{aligned} H_{23} &= \frac{1}{2} \langle \psi_1 \psi_r \uparrow \downarrow - \psi_r \psi_1 \downarrow \uparrow | H (\psi_1 \psi_r \downarrow \uparrow - \psi_r \psi_1 \uparrow \downarrow) \rangle \\ &= \frac{1}{2} \langle \psi_1 \psi_r \uparrow \downarrow - \psi_r \psi_1 \downarrow \uparrow | (H \psi_1 \psi_r) \downarrow \uparrow - (H \psi_r \psi_1) \uparrow \downarrow \rangle \end{aligned}$$

and then multiplied out into inner products of matching spin components to give

$$H_{23} = -\frac{1}{2} \langle \psi_1 \psi_r | H \psi_r \psi_1 \rangle - \frac{1}{2} \langle \psi_r \psi_1 | H \psi_1 \psi_r \rangle = L.$$

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.

4.9 Heavier Atoms [Descriptive]

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.

4.9.1 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 3.2. The total energy Hamiltonian for the

electrons of an element with atomic number Z with is:

$$H = \sum_{i=1}^Z \left[-\frac{\hbar^2}{2m_e} \nabla_i^2 - \frac{e^2}{4\pi\epsilon_0} \frac{Z}{r_i} + \frac{1}{2} \sum_{\substack{i=1 \\ i \neq i}}^Z \frac{e^2}{4\pi\epsilon_0} \frac{1}{|\vec{r}_i - \vec{r}_{\underline{i}}|} \right] \quad (4.33)$$

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:

$$H\psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_Z, S_{zZ}) = E\psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_Z, S_{zZ})$$

Key Points

- The Hamiltonian for the electron structure has been written down.

4.9.2 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 3.2.

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 ,

$$h_i = -\frac{\hbar^2}{2m} \nabla_i^2 - \frac{e^2}{4\pi\epsilon_0 r_i}.$$

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:

$$\psi^Z = \psi_1^P(\vec{r}_1, S_{z1}) \psi_2^P(\vec{r}_2, S_{z2}) \dots \psi_Z^P(\vec{r}_Z, S_{zZ}).$$

Substitution of this assumption into the eigenvalue problem $\sum_i h_i \psi^Z = E \psi^Z$ and dividing by ψ^Z produces

$$\frac{1}{\psi_1^P(\vec{r}_1, S_{z1})} h_1 \psi_1^P(\vec{r}_1, S_{z1}) + \frac{1}{\psi_2^P(\vec{r}_2, S_{z2})} h_2 \psi_2^P(\vec{r}_2, S_{z2}) + \dots = E$$

since h_1 only does anything to the factor $\psi_1^P(\vec{r}_1, S_{z1})$, h_2 only does anything to the factor $\psi_2^P(\vec{r}_2, S_{z2})$, 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 $\psi_1^P(\vec{r}_1, S_{z1})$ itself does, since none of the other terms in the equation varies with those variables. That means that

$$h_1 \psi_1^P(\vec{r}_1, S_{z1}) = \epsilon_1 \psi_1^P(\vec{r}_1, S_{z1}),$$

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 3.2; allowing for spin, the possible solutions are,

$$\psi_{100}(\vec{r}_1)\uparrow(S_{z1}), \psi_{100}(\vec{r}_1)\downarrow(S_{z1}), \psi_{200}(\vec{r}_1)\uparrow(S_{z1}), \psi_{200}(\vec{r}_1)\downarrow(S_{z1}), \dots$$

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 $\psi_{nlm}\uparrow$ hydrogen atom ones, (where \uparrow 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,

$$\psi_{n_1 l_1 m_1}(\vec{r}_1)\uparrow(S_{z1}) \psi_{n_2 l_2 m_2}(\vec{r}_2)\uparrow(S_{z2}) \dots \psi_{n_Z l_Z m_Z}(\vec{r}_Z)\uparrow(S_{zZ})$$

and the total energy is the sum of all the corresponding hydrogen atom energy levels,

$$E_{n_1} + E_{n_2} + \dots + E_{n_Z}.$$

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 4.7. 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).

4.9.3 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 3.2. The lowest energy state, or ground state, is ψ_{100} , (3.22), 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:

$$\Psi(\vec{r}_1, S_{z1}) = a_1 \psi_{100}(\vec{r}_1) \uparrow(S_{z1}) + a_2 \psi_{100}(\vec{r}_1) \downarrow(S_{z1}) = \psi_{100}(\vec{r}_1) (a_1 \uparrow(S_{z1}) + a_2 \downarrow(S_{z1})) \quad (4.34)$$

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 3.2.

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 4.7):

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}; t) = \frac{a}{\sqrt{2}} \begin{vmatrix} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) \end{vmatrix} \quad (4.35)$$

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 4.5.6.

Figure 4.4 shows the approximate probability density for the first two elements, indicating where electrons are most likely to be found. In actuality, the shielding approximation underestimates the nuclear attraction and the shown helium atom is much too big.



Figure 4.4: Approximate solutions for hydrogen (left) and helium (right) atoms.

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 significantly as they do here, their mutual repulsion is a lot less than you would naively expect, (compare figure 10.13). As a result, the second proton is only partly shielded, and the electron is held much more tightly than the analysis predicts. See chapter 12.1.2 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 {A.30}.

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.”

4.9.4 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 $\psi_{2lm\uparrow}$ 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 4.5, lithium starts a new row.

For the third electron of the lithium atom, the available states with theoretical energy E_2 are the $\psi_{200\uparrow}$ “2s” states and the $\psi_{211\uparrow}$, $\psi_{210\uparrow}$, and $\psi_{21-1\uparrow}$ “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

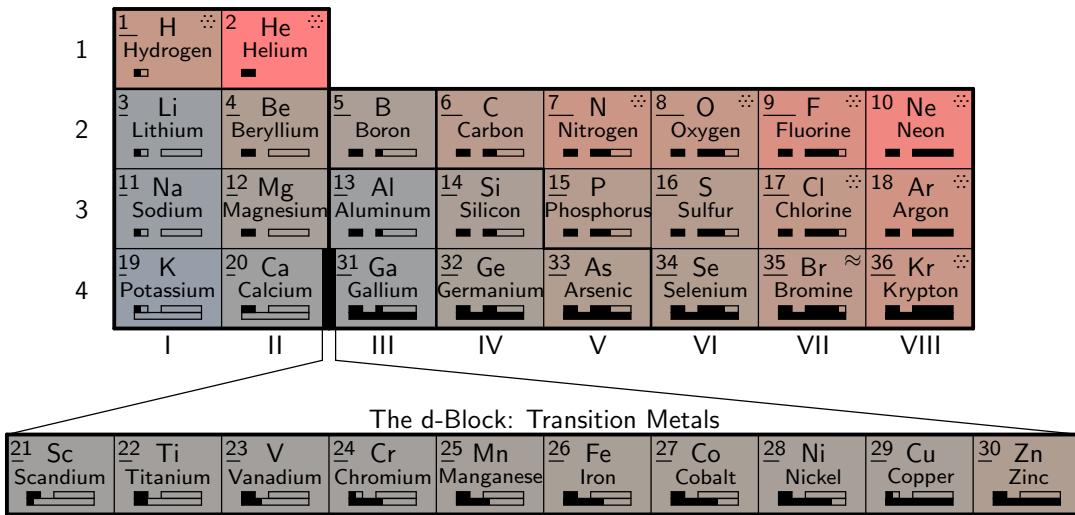


Figure 4.5: Abbreviated periodic table of the elements. Boxes below the element names indicate the quantum states being filled with electrons in that row. Cell color indicates ionization energy. The length of a bar below an atomic number indicates electronegativity. A dot pattern indicates that the element is a gas under normal conditions and wavy lines a liquid.

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 {A.31}.

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.

$$\begin{aligned}
 & \frac{a_1}{\sqrt{6}} \left| \begin{array}{ccc} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) & \psi_{200}(\vec{r}_1)\uparrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) & \psi_{200}(\vec{r}_2)\uparrow(S_{z2}) \\ \psi_{100}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{100}(\vec{r}_3)\downarrow(S_{z3}) & \psi_{200}(\vec{r}_3)\uparrow(S_{z3}) \end{array} \right| \\
 & + \frac{a_2}{\sqrt{6}} \left| \begin{array}{ccc} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) & \psi_{200}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) & \psi_{200}(\vec{r}_2)\downarrow(S_{z2}) \\ \psi_{100}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{100}(\vec{r}_3)\downarrow(S_{z3}) & \psi_{200}(\vec{r}_3)\downarrow(S_{z3}) \end{array} \right|
 \end{aligned} \quad (4.36)$$

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

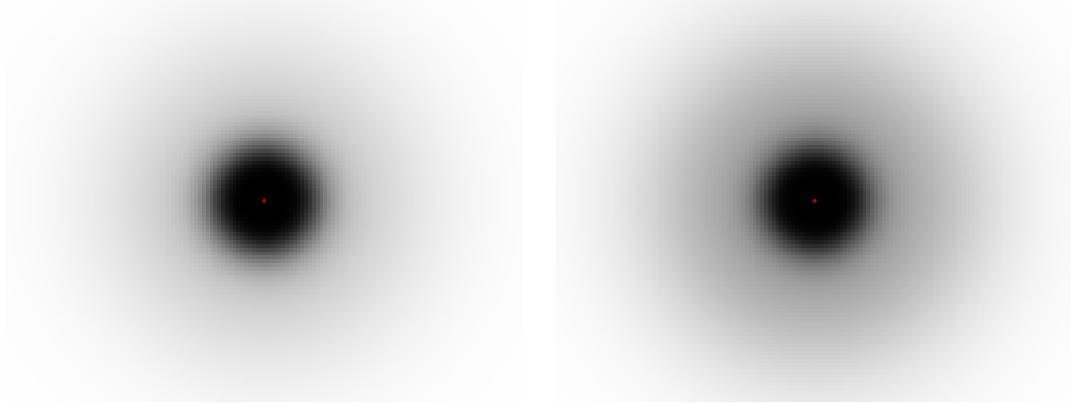


Figure 4.6: Approximate solutions for lithium (left) and beryllium (right).

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;

$$\frac{a}{\sqrt{24}} \begin{vmatrix} \psi_{100}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{100}(\vec{r}_1)\downarrow(S_{z1}) & \psi_{200}(\vec{r}_1)\uparrow(S_{z1}) & \psi_{200}(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_{100}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{100}(\vec{r}_2)\downarrow(S_{z2}) & \psi_{200}(\vec{r}_2)\uparrow(S_{z2}) & \psi_{200}(\vec{r}_2)\downarrow(S_{z2}) \\ \psi_{100}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{100}(\vec{r}_3)\downarrow(S_{z3}) & \psi_{200}(\vec{r}_3)\uparrow(S_{z3}) & \psi_{200}(\vec{r}_3)\downarrow(S_{z3}) \\ \psi_{100}(\vec{r}_4)\uparrow(S_{z4}) & \psi_{100}(\vec{r}_4)\downarrow(S_{z4}) & \psi_{200}(\vec{r}_4)\uparrow(S_{z4}) & \psi_{200}(\vec{r}_4)\downarrow(S_{z4}) \end{vmatrix} \quad (4.37)$$

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 4.7 shows the approximate solution in which the ψ_{210} , or “ $2p_z$ ” state is occupied. It may be recalled from figure 3.5 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, indicating

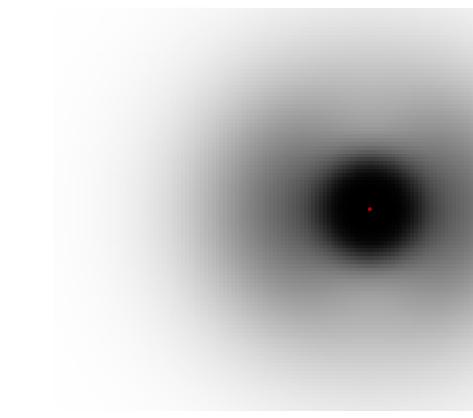


Figure 4.7: Example approximate solution for boron.

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 actuality, 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 3.2, $2p_x$ 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 $\psi_{200}\downarrow$ “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 $\psi_{2lm}\downarrow$ are called the “L shell.”
-

4.9.5 Sodium to argon

Starting with sodium (sodium), 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 4.5 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 $\psi_{32m}\downarrow$ 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 $\psi_{3lm}\downarrow$ are called the “M shell.”

4.9.6 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.

Key Points

- Unlike what the approximate theory says, in real life the 4s states $\psi_{400\downarrow}$ have less energy than the $\psi_{32m\downarrow}$ 3d states, and are filled first.
 - After that, the transition metals fill the skipped 3d states before the old logic resumes.
 - The states $\psi_{4lm\downarrow}$ are called the “N shell.” It all spells KLM Netherlands.
 - The substates are of course called “s,” “p,” “d,” “f,” ...
-

4.9.7 Full periodic table

Continuing to still heavier elements, the energy levels get even more confused. This discussion will stop while it is still ahead.

However, a complete periodic table is shown in figure 4.8. Note how the rows expand through an additional block of lanthanides and actinides when $l = 3, f$, states have to be filled.

The color of each cell indicates the ionization energy, increasing from bluish to reddish. The number in the top left corner is the atomic number Z . The length of the bar below the 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.

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 $n + 2$.

Periodic table figure 4.8 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 11 on nuclei, like figures 11.1 and 11.2, are much more useful.

It should be noted that periodic table figure 4.8 deviates in a number of aspects from the normal conventions. Figure 4.8 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 put helium in group VIII instead of group II. The intention is to try to get elements with similar properties together in the same group. But from a quantum mechanics point of view, it does not make any sense. The physical explanation why helium is a noble gas is not because it is in column VIII of a period, but because it is at the *end* of a period.

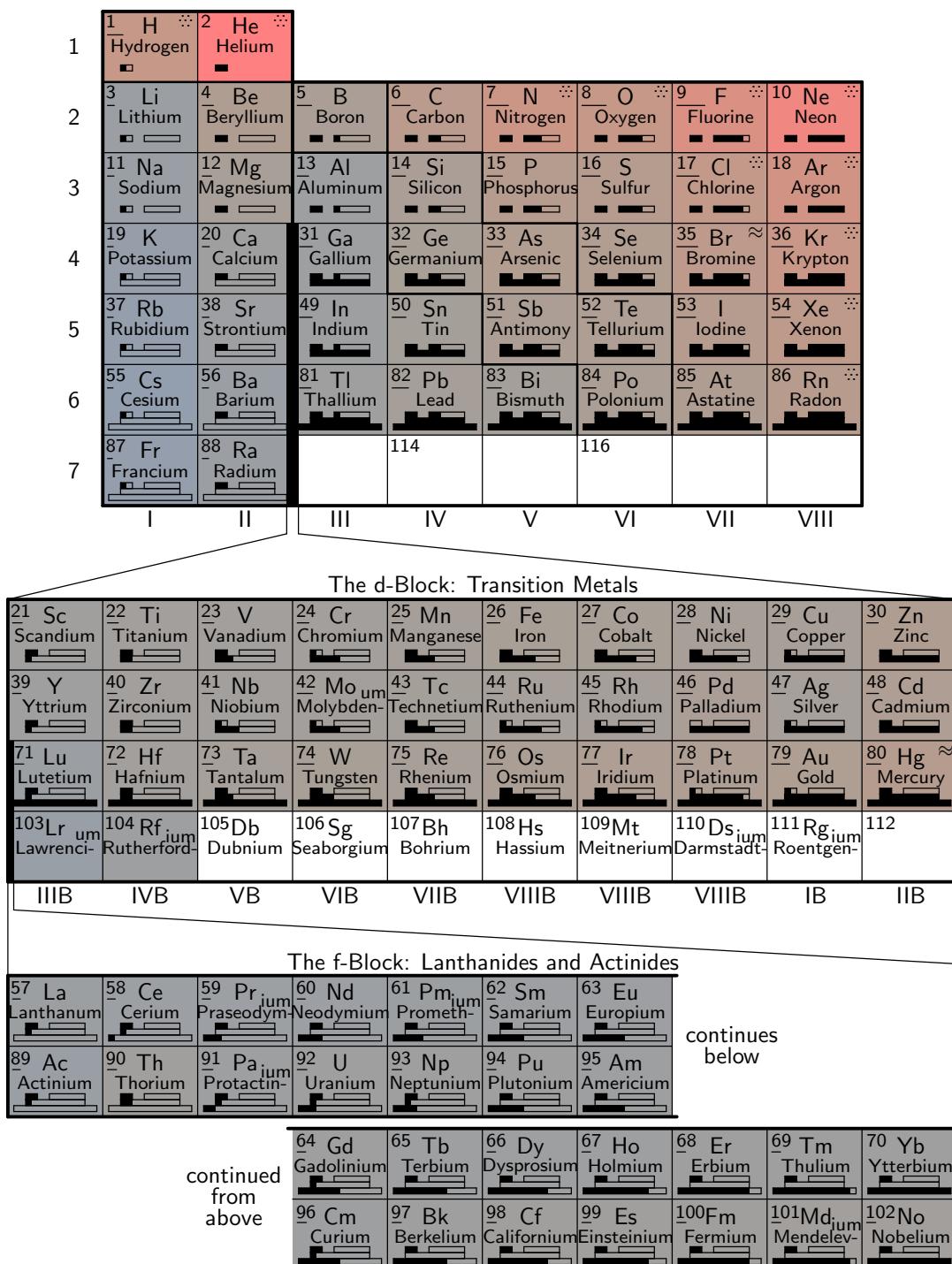


Figure 4.8: Periodic table of the elements.

Helium does not physically have 8 electrons or any other feature that can justify a number 8.

And surely you cannot reasonably simply ignore that and put helium in column VIII anyway and then *not* put hydrogen in column VII. Now, if this was a book on chemistry instead of quantum mechanics, you would probably see hydrogen and helium in groups VII and VIII.

The top of the shown table is called the main block. For some reason however, hydrogen is not included in the term. The group I elements except hydrogen are called the alkali metals. The group II elements except helium are called the alkaline metals. The group VII elements are the halogens. The group VIII elements plus helium are the noble gases.

The term transition elements 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 elements.

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 4.8 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 in the lutetium-lawrencium column.

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 4.8. 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.)

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 fattened cell boundaries in figure 4.8. It extends from boron to polonium. The metals are found to the left of this band and the nonmetals to the right.

A nice recent example of a more conventional periodic table by an authoritative source is from NIST¹. An alternate link can be found in the web version of this document². The hieroglyphs found in the NIST table are explained in chapter 8.7.1.

Periodic table figure 4.8 was based on data from various sources. Shell fillings and ionization energies agree with the NIST listing and table. The uncertain shell fillings at atomic numbers 103 and 104 were left out. 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.

4.10 Pauli Repulsion [Descriptive]

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.

¹<http://physics.nist.gov/PhysRefData/PerTable/periodic-table.pdf>

²<http://www.eng.fsu.edu/~dommelen/quansup/periodic-table.pdf>

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.
-

4.11 Chemical Bonds [Descriptive]

The electron states, or “atomic orbitals”, of the elements discussed in section 4.9 form the basis for the “valence bond” description of chemical bonds. This section summarizes some of the basic ideas involved.

4.11.1 Covalent sigma bonds

As pointed out in section 4.9, 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 4.2, 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 H₂ molecule stays together. A sketch of the shared 1s electrons was given in figure 4.2.

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 4.9 (all other electrons have been omitted.) This bond between p electrons looks quite

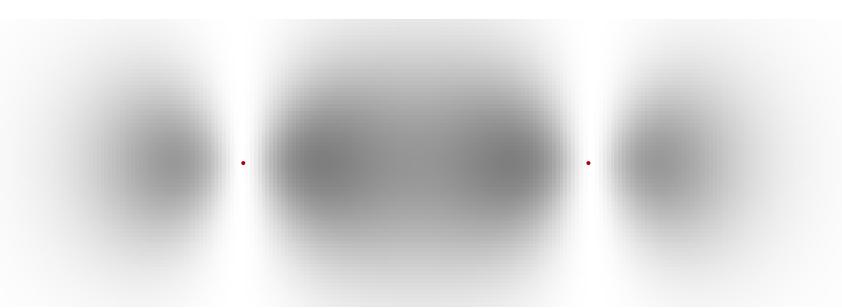


Figure 4.9: Covalent sigma bond consisting of two $2p_z$ states.

different from the H_2 bond between s electrons in figure 4.2, 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 .)

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.

4.11.2 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 4.10 illustrates this for the

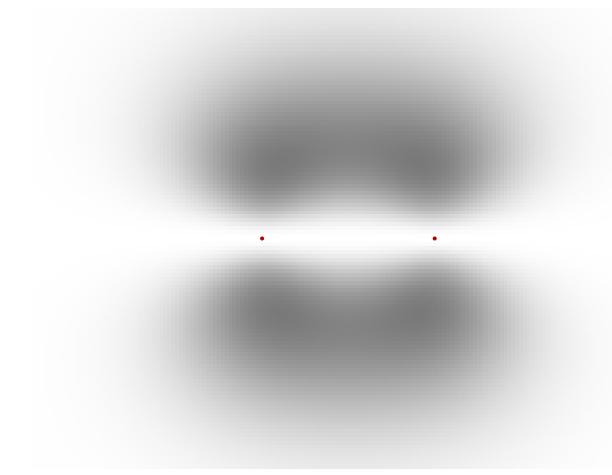


Figure 4.10: Covalent pi bond consisting of two $2p_x$ states.

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.
-

4.11.3 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 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 4.11. Similarly, the lone $2p_y$ 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 axis are orthogonal, the two hydrogen atoms in water should be at a 90 degree angle from each other, relative to the oxygen nucleus.

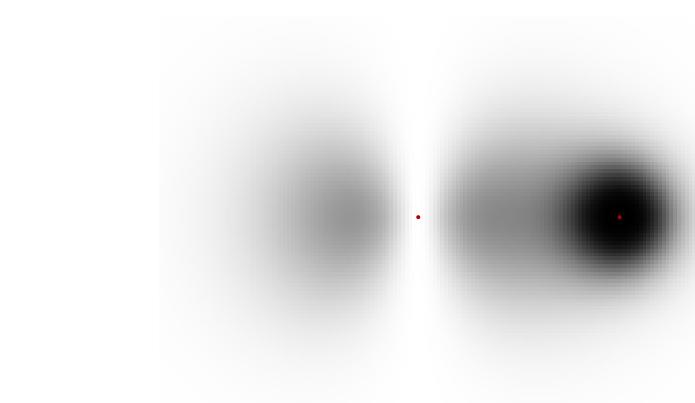


Figure 4.11: Covalent sigma bond consisting of a $2p_z$ and a $1s$ state.

(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.
-

4.11.4 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 CH₄ 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 CH₄ 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 CH₄ 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 3.2 to produce the equivalent 2p_x and 2p_y pointer states.)

In case of CH₄, the carbon converts the 2s, 2p_x, 2p_y, and 2p_z states into four new states. These are called sp³ states, since they are formed from one s and

three p states. They are given by:

$$|sp_a^3\rangle = \frac{1}{2}(|2s\rangle + |2p_x\rangle + |2p_y\rangle + |2p_z\rangle)$$

$$|sp_b^3\rangle = \frac{1}{2}(|2s\rangle + |2p_x\rangle - |2p_y\rangle - |2p_z\rangle)$$

$$|sp_c^3\rangle = \frac{1}{2}(|2s\rangle - |2p_x\rangle + |2p_y\rangle - |2p_z\rangle)$$

$$|sp_d^3\rangle = \frac{1}{2}(|2s\rangle - |2p_x\rangle - |2p_y\rangle + |2p_z\rangle)$$

where the kets denote the wave functions of the indicated states.

All four sp^3 hybrids have the same shape, shown in figure 4.12. The asym-

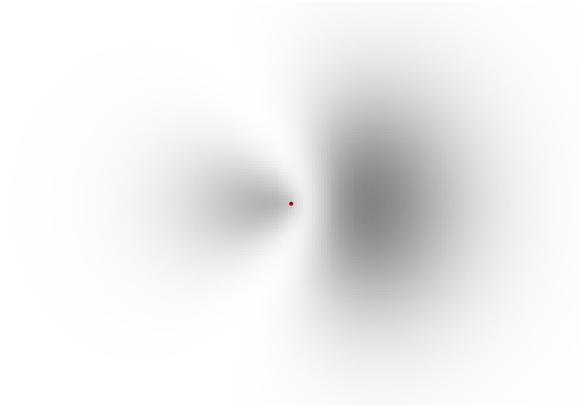


Figure 4.12: Shape of an sp^3 hybrid state.

metrical 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

plus the 2s state can then be combined into three sp^2 states:

$$\begin{aligned} |sp_a^2\rangle &= \frac{1}{\sqrt{3}}|2s\rangle + \frac{2}{\sqrt{6}}|2p_z\rangle \\ |sp_b^2\rangle &= \frac{1}{\sqrt{3}}|2s\rangle - \frac{1}{\sqrt{6}}|2p_z\rangle + \frac{1}{\sqrt{2}}|2p_y\rangle \\ |sp_c^2\rangle &= \frac{1}{\sqrt{3}}|2s\rangle - \frac{1}{\sqrt{6}}|2p_z\rangle - \frac{1}{\sqrt{2}}|2p_y\rangle \end{aligned}$$

Each is shaped as shown in figure 4.13. These planar hybrids are under 120

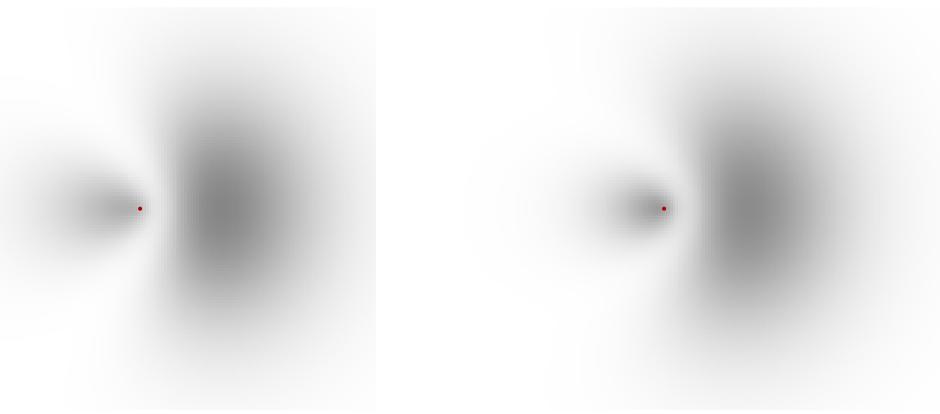


Figure 4.13: Shapes of the sp^2 (left) and sp (right) hybrids.

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:

$$\begin{aligned} |sp_a\rangle &= \frac{1}{\sqrt{2}}(|2s\rangle + |2p_z\rangle) \\ |sp_b\rangle &= \frac{1}{\sqrt{2}}(|2s\rangle - |2p_z\rangle) \end{aligned}$$

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.
-

4.11.5 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 8.2 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.
-

4.11.6 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 O₂ 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 5

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.

5.1 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 crystal 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:

$$\begin{aligned}\psi_{\vec{n}_1, \vec{n}_2, \dots, \vec{n}_I}^S(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_I, S_{zi}, \dots, \vec{r}_I, S_{zI}) = \\ \psi_{\vec{n}_1}^P(\vec{r}_1, S_{z1}) \times \psi_{\vec{n}_2}^P(\vec{r}_2, S_{z2}) \times \dots \times \psi_{\vec{n}_i}^P(\vec{r}_i, S_{zi}) \times \dots \times \psi_{\vec{n}_I}^P(\vec{r}_I, S_{zI})\end{aligned}\quad (5.1)$$

In such a system eigenfunction, particle number i out of I is in a single-particle energy eigenfunction $\psi_{\vec{n}_i}^P(\vec{r}_i, S_{zi})$. 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 2.5 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. More specifically, 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 4.7. Simple Hartree product energy eigenfunctions of the form (5.1) 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.
 - □ (Anti)symmetrization requirements further restrict the system energy eigenfunctions.
-

5.2 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 2.5 under the guise of a pipe with a rectangular cross section. The single-particle energy eigenfunctions are:

$$\psi_{n_x n_y n_z}^p(\vec{r}) = \sqrt{\frac{8}{\mathcal{V}}} \sin(k_x x) \sin(k_y y) \sin(k_z z) \quad (5.2)$$

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:

$$k_x = n_x \frac{\pi}{\ell_x} \quad k_y = n_y \frac{\pi}{\ell_y} \quad k_z = n_z \frac{\pi}{\ell_z} \quad (5.3)$$

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 ψ_{111}^p , 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 5.1.

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

$$E^p = \frac{\hbar^2}{2m} k^2 \quad k \equiv \sqrt{k_x^2 + k_y^2 + k_z^2} \quad (5.4)$$

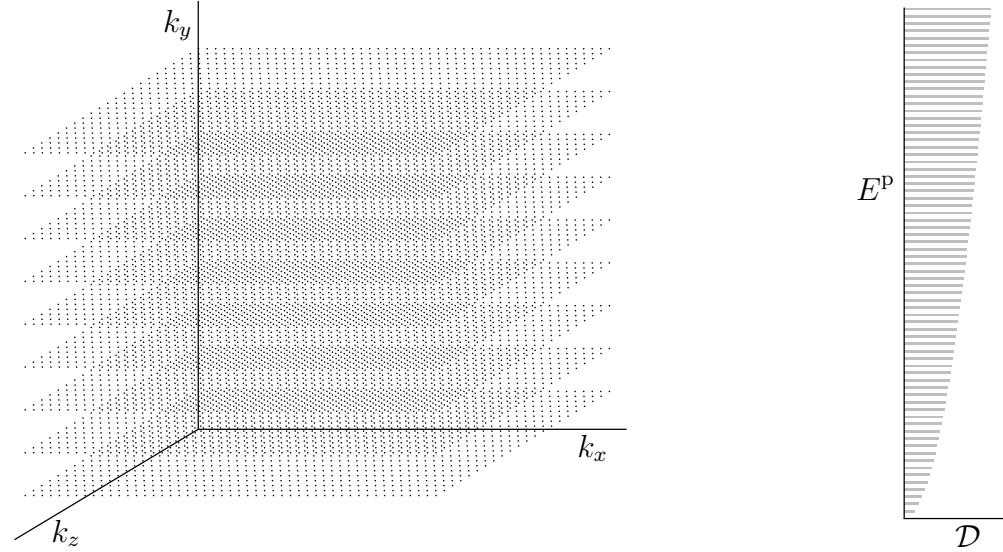


Figure 5.1: Allowed wave number vectors, left, and energy spectrum, right.

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.)

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:

$$\begin{aligned}\psi_{n_x n_y n_z, \frac{1}{2}}^p(\vec{r}, S_z) &= \sqrt{\frac{8}{\mathcal{V}}} \sin(k_x x) \sin(k_y y) \sin(k_z z) \uparrow(S_z) \\ \psi_{n_x n_y n_z, -\frac{1}{2}}^p(\vec{r}, S_z) &= \sqrt{\frac{8}{\mathcal{V}}} \sin(k_x x) \sin(k_y y) \sin(k_z z) \downarrow(S_z)\end{aligned}$$

That means that each point in the wave number space figure 5.1 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 chapter 12.2.3 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 5.1 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 $2s + 1$.
-

5.3 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 3.2.

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 $7 \cdot 10^6$ cones and $1.3 \cdot 10^8$ 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 5.1. The spacing between energy levels is however very many orders of magnitude tighter than the hatching can indicate.

	helium atom	electron	photon
E_{111}^p , eV:	$1.5 \cdot 10^{-18}$	$1.1 \cdot 10^{-14}$	$1.1 \cdot 10^{-4}$
T_{equiv} , K:	$1.2 \cdot 10^{-14}$	$8.7 \cdot 10^{-11}$	0.83

Table 5.1: 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 cm³ cubic box in table 5.1. The table gives the lowest energy as computed using the formulae given in the previous section. The lowest energy occurs for the state ψ_{111}^P 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 $1.6 \cdot 10^{-19}$ 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, minuscule as it might be.

The next question is why quantum mechanics is needed here at all. Classical non-quantum 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 dN of single-particle states in a small wave number range from k to $k + dk$ is given by, {A.32},

$$dN = \mathcal{V} \mathcal{D}_k dk \quad \mathcal{D}_k = \frac{2s+1}{2\pi^2} k^2 \quad (5.5)$$

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 dk , 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 dk should be small. However, dk should still be large enough that there are a lot of states in the range dk ; otherwise \mathcal{D}_k cannot be approximated by a simple continuous function. If the spacing dk 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 $E^p = \hbar^2 k^2 / 2m$, where m is the particle mass. That gives:

$$dN = \mathcal{V} \mathcal{D} dE^p \quad \mathcal{D} = \frac{2s+1}{4\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \sqrt{E^p} \quad (5.6)$$

The requirements on the energy range dE^p are like those on dk .

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 5.1, the density of states is indicated by means of the width of the spectrum.

Note that the density of states grows like $\sqrt{E^p}$: 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 5.1 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, $E^p = \hbar\omega$, where the electromagnetic frequency is $\omega = ck$ with c the speed of light. In addition, as mentioned in section 5.2, 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 (5.5) and $2s+1$ by 2 gives

$$dN = \mathcal{V} \mathcal{D}_\omega d\omega \quad \mathcal{D}_\omega = \frac{1}{\pi^2 c^3} \omega^2 \quad (5.7)$$

The factor \mathcal{D}_ω 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 dE^p is $\mathcal{V} \mathcal{D} dE^p$.

- To use this expression, the energy range dE^P should be small. However, dE^P should still be large enough that there are a lot of states in the range.
 - For photons, use the density of modes.
-

5.4 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 $\psi_{111}^P(\vec{r})$ of lowest energy. That makes the system energy eigenfunction for spinless bosons equal to:

$$\psi_{\text{gs, bosons}} = \psi_{111}^P(\vec{r}_1) \times \psi_{111}^P(\vec{r}_2) \times \dots \times \psi_{111}^P(\vec{r}_I) \quad (5.8)$$

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_{111}^P , the number of bosons times the single-particle ground state energy.

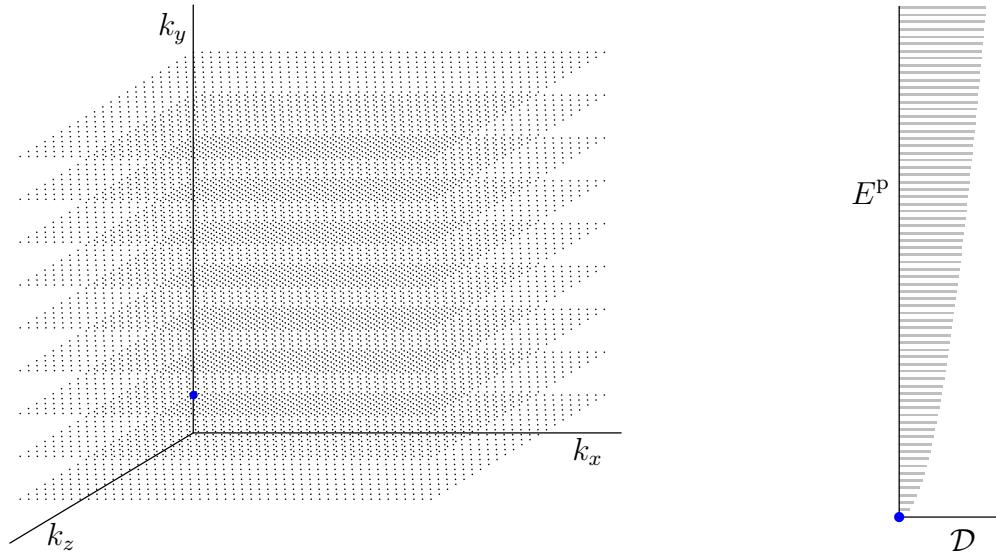


Figure 5.2: Ground state of a system of noninteracting bosons in a box.

Graphically, the single-particle ground state ψ_{111}^P is the point closest to the origin in wave number space. It is shown as a fat blue dot in figure 5.2 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_{111}^p 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: ψ_{111}^p 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.

5.5 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 $-273.15\text{ }^\circ\text{C}$ in degrees Celsius (Centigrade) or $-459.67\text{ }^\circ\text{F}$ 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 $-273.15\text{ }^\circ\text{C}$ or $-459.67\text{ }^\circ\text{F}$. 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 $-272.15\text{ }^\circ\text{C}$; both are one degree above absolute zero. Normal ambient temperatures are near 300 K . More precisely, 300 K is equal to $27.15\text{ }^\circ\text{C}$ or $80.6\text{ }^\circ\text{F}$.

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 $\frac{3}{2}k_B T$, with $k_B = 1.38 \cdot 10^{-23}$ 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.42}, 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 $\frac{3}{2}k_B T$, where k_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.

5.6 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 5.2. 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 5.3 and 5.4. So far, that all seems plausible enough.

But something weird occurs for identical bosons:

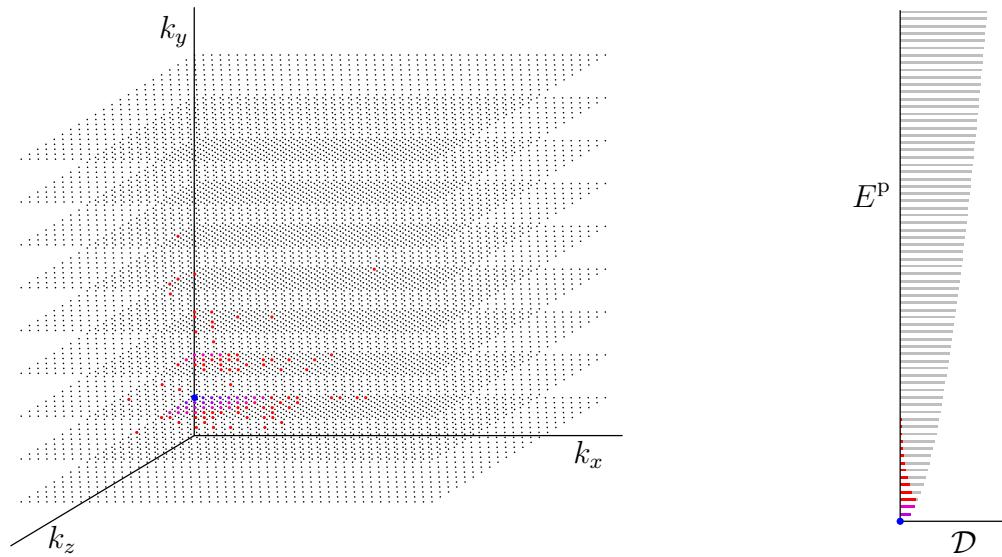


Figure 5.3: The system of bosons at a very low temperature.

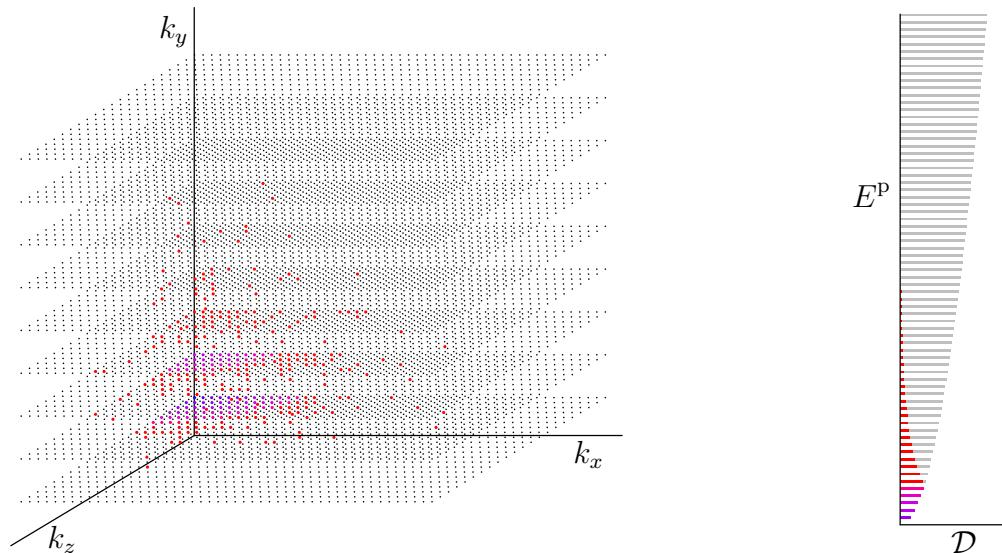


Figure 5.4: The system of bosons at a relatively low temperature.

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 5.3. 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 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 5.4. 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 hard.

A process similar to Bose-Einstein condensation might also occur in liquid helium when it turns into a superfluid below 2.17 K. However, the evidence is ambiguous, {A.72}. 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 the lowest-energy single-particle state. Unlike for Bose-Einstein condensation, for helium it is believed that the number of atoms in the single-particle ground state remains limited. At absolute zero only about 8% of the atoms end up in that state.

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.

5.6.1 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 9. However, the idea can be explained qualitatively by examining a very simple system

Assume that there are just three different single-particle energy levels, with values E_1^P , $2E_1^P$, and $3E_1^P$. Also assume that there is just one single-particle state with energy E_1^P , but two with energy $2E_1^P$ and 3 with energy $3E_1^P$. 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 5.5. Assume also that there are just three particles and for now take them to be distinguishable. Figure 5.5 then shows the system ground state in which every particle is in the single-particle ground state with energy E_1^P . That makes the total system energy $3E_1^P$.

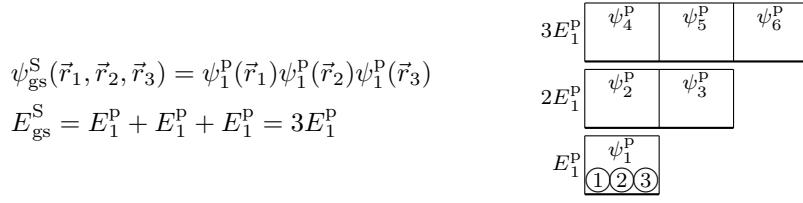


Figure 5.5: Ground state system energy eigenfunction for a simple model system with only 3 single-particle energy levels, 6 single-particle states, and 3 distinguishable spinless particles. Left: mathematical form. Right: graphical representation. All three particles are in the single-particle ground state.

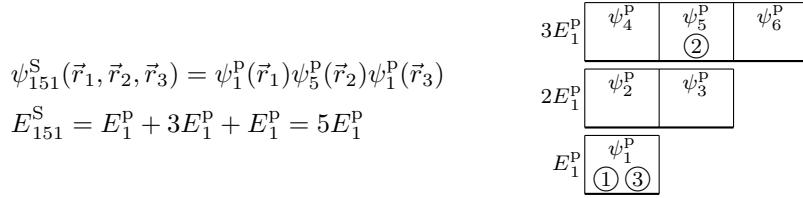


Figure 5.6: Example system energy eigenfunction with five times the single-particle ground state energy.

However, now assume that the system is at a nonzero temperature. In particular, assume that the total system energy is $5E_1^{\text{P}}$. An example system energy eigenfunction with that energy is illustrated in figure 5.6.

But there are a lot more system eigenfunctions with energy $5E_1^{\text{P}}$. There are two general ways to achieve that energy:

Energy distribution A: Two particles in the ground state with energy E_1^{P} and one in a state with energy $3E_1^{\text{P}}$.

Energy distribution B: One particle in the ground state with energy E_1^{P} and two in states with energy $2E_1^{\text{P}}$.

As figures 5.7 and 5.8 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

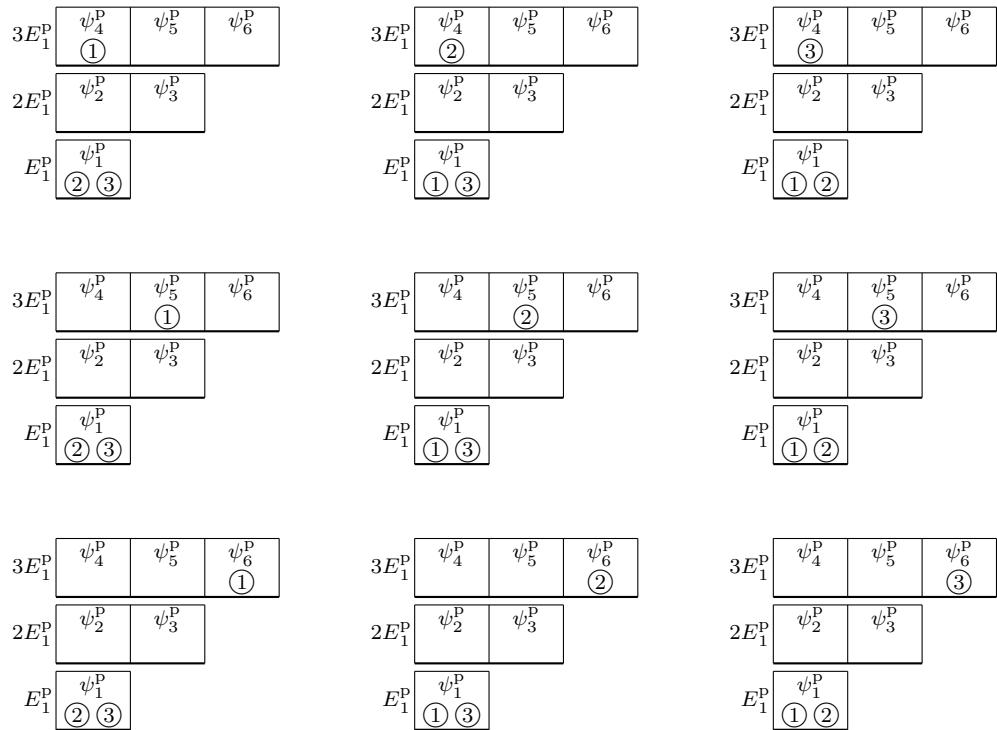


Figure 5.7: For distinguishable particles, there are 9 system energy eigenfunctions that have energy distribution A.

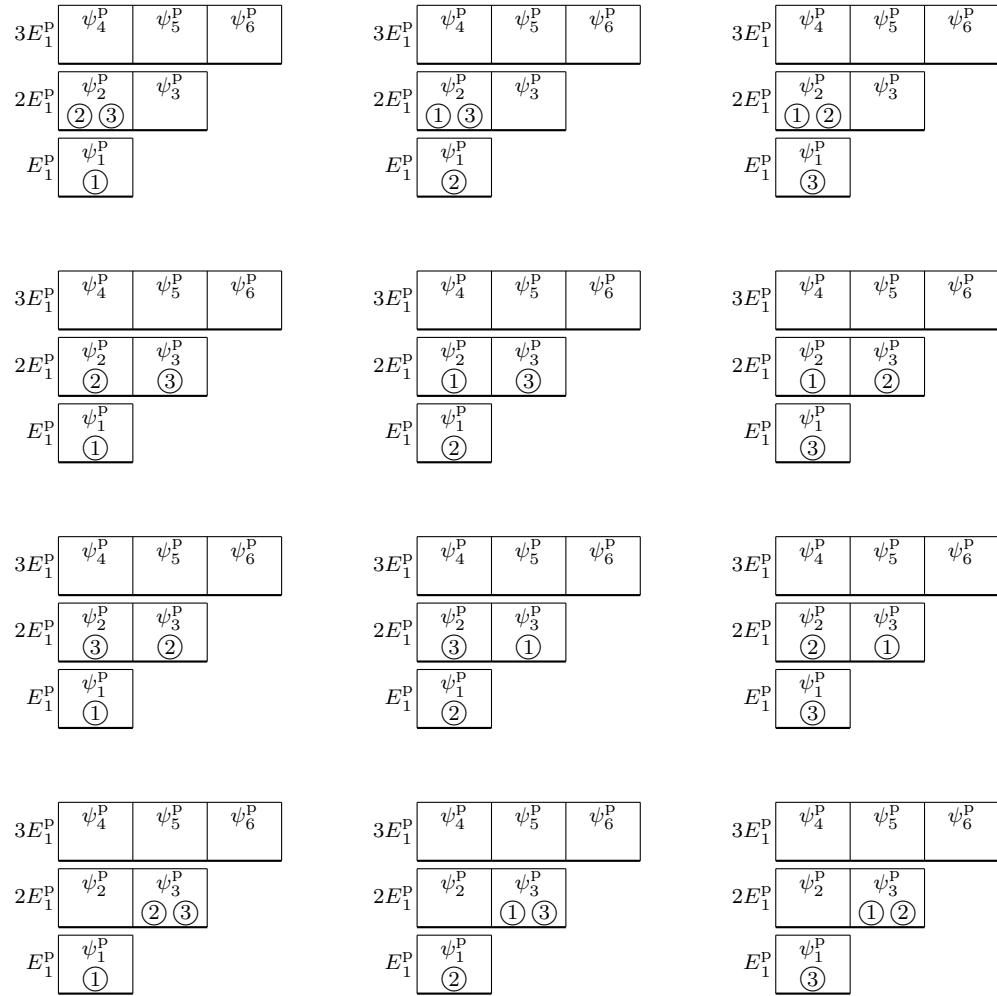


Figure 5.8: For distinguishable particles, there are 12 system energy eigenfunctions that have energy distribution B.

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.

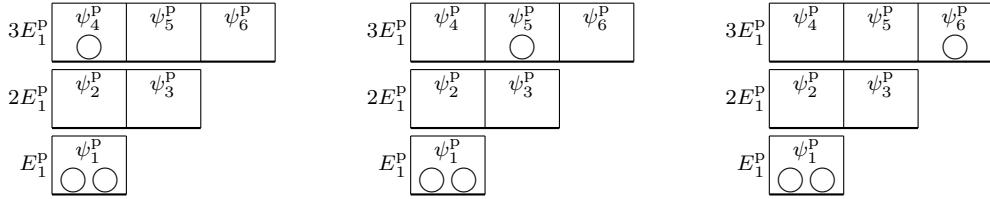


Figure 5.9: For identical bosons, there are only 3 system energy eigenfunctions that have energy distribution A.

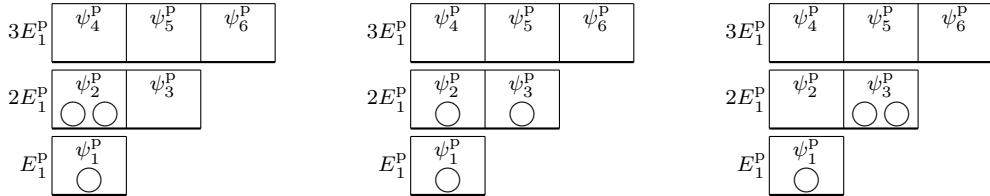


Figure 5.10: For identical bosons, there are also only 3 system energy eigenfunctions that have energy distribution B.

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 4.7. 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 5.7 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 be combined together into a single eigenfunction. That leaves only the three system eigenfunctions shown in figure 5.9.

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 5.9. But what good would that do?

Comparing figures 5.7 and 5.9, 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 5.8, 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 5.10.

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 5.7 and 5.8 with 5.9 and 5.10 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, [13, 26]. 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.
-

5.7 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:

$$\text{Bose-Einstein distribution: } \nu^b = \frac{1}{e^{(E^p - \mu)/k_B T} - 1} \quad (5.9)$$

Here ν^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 9.12.

The Bose-Einstein distribution is derived in chapter 9. 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_{gs}^p . Just check it out using the formula above: if μ would be greater than E_{gs}^p , then the number of particles ν^b in the lowest single-particle state would be negative. Negative numbers of particles do not exist. Similarly, if μ would equal E_{gs}^p then the number of particles in the lowest single-particle state would be infinite.

The fact that μ must stay less than E_{gs}^p means that the number of particles in anything but the lowest single-particle state has a limit. It cannot become greater than

$$\nu_{\max}^b = \frac{1}{e^{(E^p - E_{gs}^p)/k_B T} - 1}$$

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 τ^b in all states. However, that would lower the total number of particles, which is kept constant. To compensate, μ inches closer to E_{gs}^p . 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 9.14.1.

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.

5.8 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

$$\rho(\omega) = \frac{\omega^2}{\pi^2 c^3} k_B T$$

where $k_B = 1.38 \cdot 10^{-23}$ 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

$$\nu_\gamma^b = \frac{1}{e^{E^p/k_B T} - 1} \quad (5.10)$$

where γ is the standard symbol for a photon. Note the missing chemical potential. As discussed in chapter 9, 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.)

To get the energy of the photons in a small frequency range $d\omega$, simply multiply the number of single particle states in that range, (5.7), by the number of photons per state ν_γ^b above, and that by the single-photon energy $E^p = \hbar\omega$.

That gives the radiation energy per unit volume of the box and per unit energy range as

$$\rho(\omega) = \frac{\omega^2}{\pi^2 c^3} \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1}$$

(5.11)

This expression is known as “Planck’s blackbody spectrum.”

For low frequencies, the final ratio is about $k_B T$, 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, $dE = \rho(\omega) d\omega$, to $dE = \bar{\rho}(f) df$. 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 , {A.33}. That gives for the radiation energy emitted per unit area, per unit frequency range, and per unit time:

$$\boxed{\mathcal{I}(\omega) = \frac{\omega^2}{4\pi^2 c^2} \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1}} \quad (5.12)$$

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

$$dE = \frac{\omega^2}{4\pi^3 c^2} \frac{A_e \cos \theta_e A_r \cos \theta_r}{r^2} \frac{\hbar\omega}{e^{\hbar\omega/k_B T} - 1} d\omega dt \quad (5.13)$$

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 (5.12) over all values of the frequency. You will want to make a change of integration variable to $\hbar\omega/k_B T$ while doing this and then use a table book like [28, 18.80, p. 132]. The result is called the “Stefan-Boltzmann law:

$$dE_{\text{total emitted}} = A\sigma_B T^4 dt \quad \sigma_B = \frac{\pi^2 k_B^4}{60\hbar^3 c^2} \approx 5.67 \cdot 10^{-8} \text{ W/m}^2 \text{ K}^4 \quad (5.14)$$

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, {A.34}. 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-versa. 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 (5.13) 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.

5.9 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 5.2. 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.

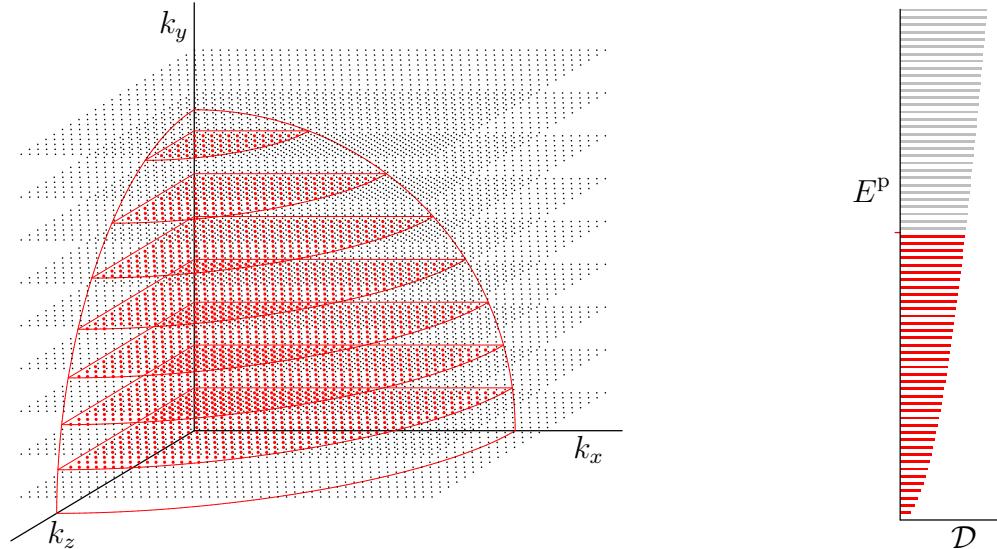


Figure 5.11: Ground state of a system of noninteracting electrons, or other fermions, in a box.

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 5.11. 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 5.11 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.

5.10 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. Therefor, 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. Therefor 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. Therefor, 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 5.11, 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 5.11. 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 5.1. The lowest single-particle energy is, assuming that the box is cubic

$$E_{111}^p = 3\pi^2 \frac{\hbar^2}{2m_e} \frac{1}{V^{2/3}} \quad (5.15)$$

where m_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 5.11 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_{111}^p .

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:

$$E_F^p = \left(3\pi^2\right)^{2/3} \frac{\hbar^2}{2m_e} \left(\frac{I}{\mathcal{V}}\right)^{2/3} \quad (5.16)$$

To check this relationship, integrate the density of states (5.6) given in section 5.3 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

$$\frac{I^{2/3}}{(3\pi^2)^{1/3}}$$

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 $8.5 \cdot 10^{28}$ atoms/m³ by dividing the mass density, 9 000 kg/m³, by the molar mass, 63.5 kg/kmol, and then multiplying that by Avogadro's number, $6.02 \cdot 10^{26}$ particles/kmol. Plugging it in (5.16) 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:

$$E_{\text{average}}^p = \frac{3}{5} E_F^p \quad (5.17)$$

To verify this expression, find the total energy $E = \int E^p \mathcal{V} \mathcal{D} dE^p$ of the electrons using (5.6) and divide by the number of electrons $I = \int \mathcal{V} \mathcal{D} dE^p$. 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 $e/\frac{3}{2}k_B$ 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.

5.11 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

$$P_d = \frac{2}{5} (3\pi^2)^{2/3} \frac{\hbar^2}{2m_e} \left(\frac{I}{V}\right)^{5/3} \quad (5.18)$$

This may be verified by equating the work $-P_d dV$ done when compressing the volume a bit to the increase in the total kinetic energy E^S of the electrons:

$$-P_d dV = dE^S$$

The energy E^S is I times the average energy per electron. According to section 5.10, that is $\frac{3}{5}I$ times the Fermi energy (5.16).

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 $40 \cdot 10^9$ 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 $E^p = p^2/2m$, but by the Planck-Einstein relationship $E^p = pc$ 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 $\frac{1}{2}$ fermions but 2000 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.

5.12 Confinement and the DOS

The motion of a single particle in a confining box was described in chapter 2.5.9. 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 dN in an energy range from E^p to $E^p + dE^p$ as

$$dN = \mathcal{V}\mathcal{D} dE^p$$

where \mathcal{V} is the volume of the box containing the particles. To use this expression, the size of the energy range dE^p should be small, but still big enough that the number of states dN in it remains large.

For a box that is not confining, the density of states is proportional to $\sqrt{E^p}$. 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 5.11 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 $\sqrt{E^p}$, (5.4), so

$$N = (\text{some constant}) \left(E^p \right)^{3/2}$$

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 $E^p + dE^p$, take a differential:

$$dN = (\text{some other constant}) \sqrt{E^p} dE^p$$

So the density of states is proportional to $\sqrt{E^P}$. (The constant of proportionality is worked out in note {A.32}.) This density of states is shown as the width of the energy spectrum to the right in figure 5.11.

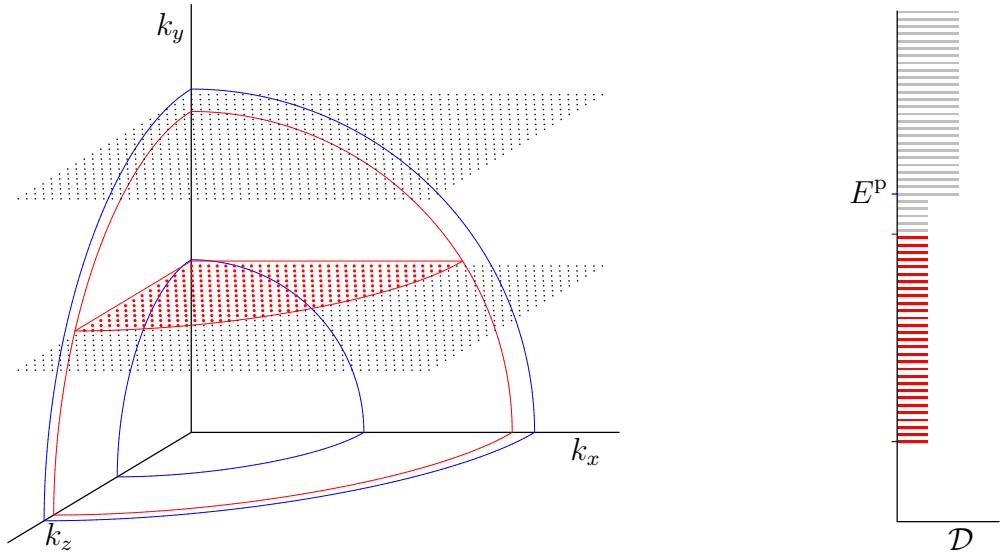


Figure 5.12: Severe confinement in the y -direction, as in a quantum well.

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 5.12. The red states are again the ones with an energy below some given example value E^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 5.12.

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 5.12, 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 note {A.32}.)

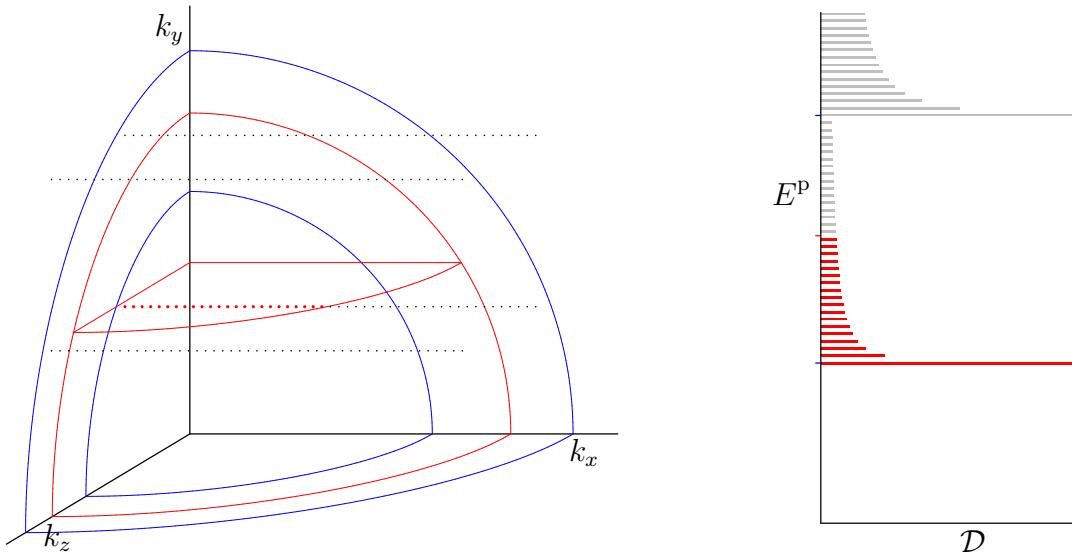


Figure 5.13: Severe confinement in both the y - and z -directions, as in a quantum wire.

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 5.13. 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.

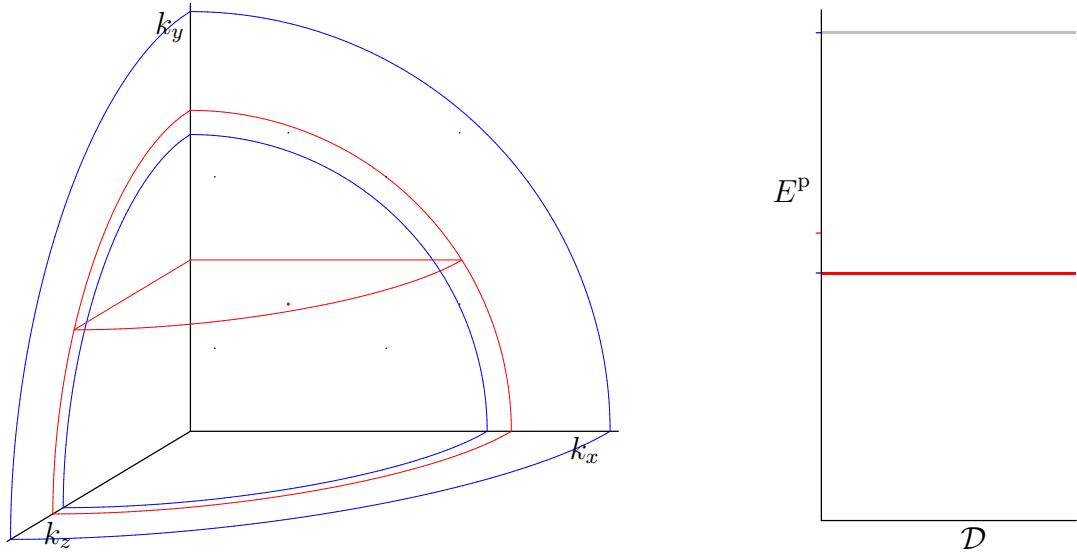


Figure 5.14: Severe confinement in all three directions, as in a quantum dot or artificial atom.

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 5.14. The density of states is now zero unless the energy sphere exactly hits one of the individual points, in which 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 dE^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 a 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.

5.13 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 ι^f per single-particle state can be found from the so-called

$$\boxed{\text{Fermi-Dirac distribution: } \iota^f = \frac{1}{e^{(E^p - \mu)/k_B T} + 1}} \quad (5.19)$$

This distribution is derived in chapter 9. Like the Bose-Einstein distribution for bosons, it depends on the energy E^p of the single-particle state, the absolute temperature T , the Boltzmann constant $k_B = 1.38 \cdot 10^{-23}$ 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 ι^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^p below the chemical potential μ . Similarly the states that end up empty are the ones with E^p above μ .

To see why, note that for $E^p - \mu < 0$, in the limit $T \rightarrow 0$ the argument of the exponential in the Fermi-Dirac distribution becomes minus infinity. That makes the exponential zero, and ι^f is then equal to one. Conversely, for $E^p - \mu > 0$, in the limit $T \rightarrow 0$ the argument of the exponential in the Fermi-Dirac distribution

becomes positive infinity. That makes the exponential infinite, and ν^f is then zero.

The correct ground state, as pictured earlier in figure 5.11, has one electron per state below the Fermi energy E_F^p 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:

$$\mu = E_F^p \quad \text{at} \quad T = 0 \quad (5.20)$$

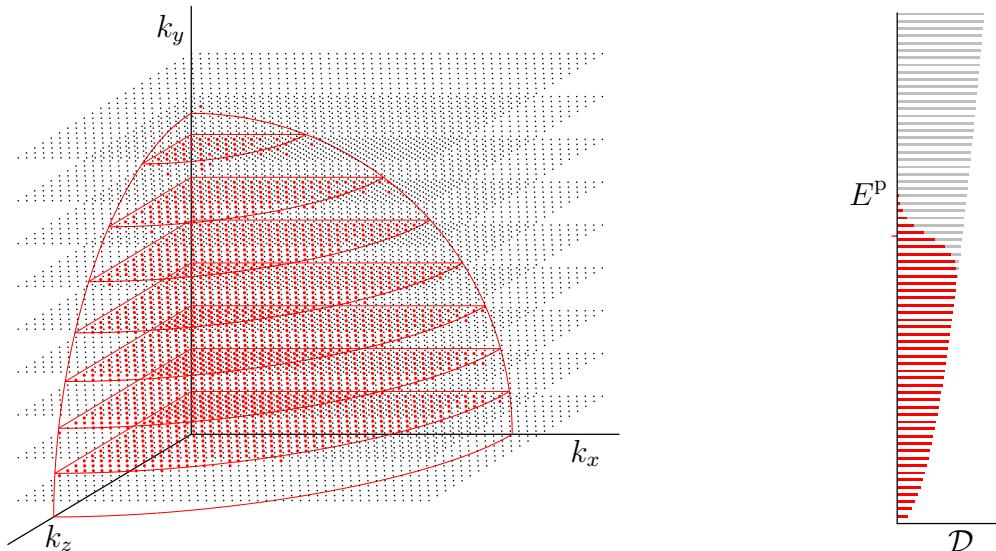


Figure 5.15: A system of fermions at a nonzero temperature.

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 $E^p - \mu$ is comparable to $k_B T$. Only in a energy range of order $k_B T$ 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 5.11 to the one at a nonzero temperature shown in figure 5.15. 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 5.15 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 $k_B T$.

You may at first hardly notice the effect in the wave number space shown in figure 5.15. 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 $k_B T$ 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 temper-

ature 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 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 non-zero 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.

5.14 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

$$\text{Maxwell-Boltzmann distribution: } \nu^d = \frac{1}{e^{(E^p - \mu)/k_B T}} \quad \nu^d \ll 1 \quad (5.21)$$

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 (5.9) and (5.19) can be ignored.

Figure 5.16 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.

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 ν^d refers to distinguishable.

The Maxwell-Boltzmann distribution was already known before quantum mechanics. The factor $e^{-E^p/k_B T}$ 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 $e^{-mgh/k_B T}$ 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

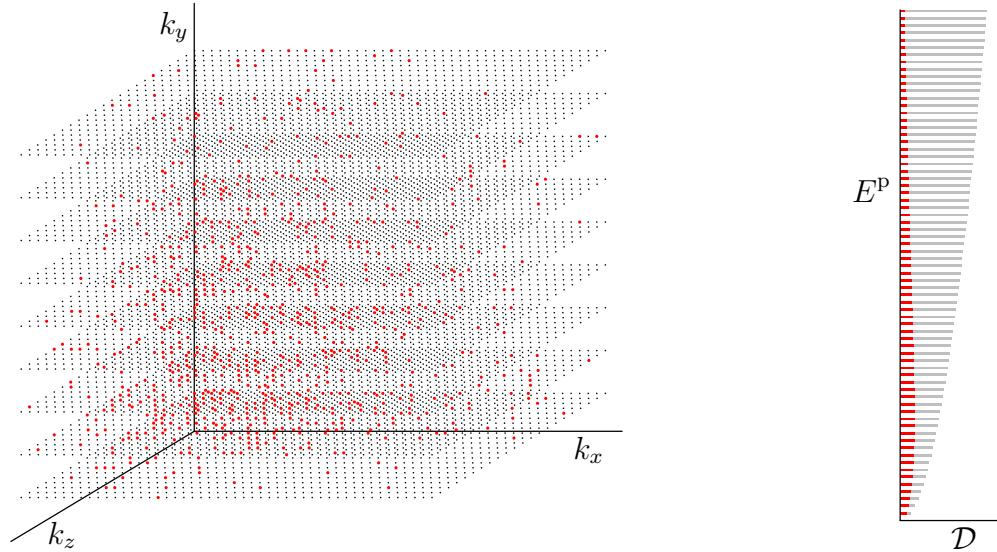


Figure 5.16: Particles at high-enough temperature and low-enough particle density.

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:

$$E^p = mgh + E_i^p \quad \mu = mgh + \mu_i$$

Since $E^p - \mu = E_i^p - \mu_i$, 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_i^p is $1/e^{(E_i^p+mgh-\mu)/k_B T}$. That decreases with height proportional to $e^{-mgh/k_B T}$, 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 $e^{\mu/k_B T}$. 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 (5.21) 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 :

$$\frac{I_2}{I_1} = \frac{8}{2} e^{-(E_2 - E_1)/k_B T}$$

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 $k_B T$ 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.

5.15 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, {A.35},

$$j = AT^2 e^{-e\varphi_w/k_B T} \quad (5.22)$$

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

$$A_{\text{theory}} = \frac{m_e e k_B}{2\pi^2 \hbar^3} \approx 1.2 \cdot 10^6 \text{ amp/m}^2 \cdot \text{K}^2 \quad (5.23)$$

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, {A.35}.

Thermionic emission can be helped along by applying an additional electric field E_{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, {A.35},

$$\sqrt{\frac{eE_{\text{ext}}}{4\pi\epsilon_0}} \quad (5.24)$$

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.

5.16 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 5.28.2 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.

5.17 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 6.4 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 $\hat{\vec{p}}$. 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.

5.18 Periodic Single-Particle States

The single-particle quantum states, or energy eigenfunctions, for noninteracting particles in a closed box were given in section 5.2, (5.2). 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:

$$\psi_{n_x n_y n_z}^P(\vec{r}) = \mathcal{V}^{-\frac{1}{2}} e^{i(k_x x + k_y y + k_z z)} = \mathcal{V}^{-\frac{1}{2}} e^{i\vec{k} \cdot \vec{r}} \quad (5.25)$$

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 eigenfunctions 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 $\hbar \partial / i\partial x$ on the eigenfunction above. The same for the other two components of linear momentum, so:

$$p_x = \hbar k_x \quad p_y = \hbar k_y \quad p_z = \hbar k_z \quad \vec{p} = \hbar \vec{k} \quad (5.26)$$

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

$$E^P = \frac{\hbar^2}{2m} k^2 \quad k \equiv \sqrt{k_x^2 + k_y^2 + k_z^2} \quad (5.27)$$

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 momemtum 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, $e^{ik_x \ell_x}$ must be the same as $e^{ik_x 0}$. That restricts $k_x \ell_x$ to be an integer multiple of 2π , (1.5). The same for the other two components of the wave number vector, so:

$$k_x = n_x \frac{2\pi}{\ell_x} \quad k_y = n_y \frac{2\pi}{\ell_y} \quad k_z = n_z \frac{2\pi}{\ell_z} \quad (5.28)$$

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, $e^{-ik_x x}$ is not just a multiple of $e^{ik_x x}$ but a fundamentally different function.

Figure 5.17 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 5.11; 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.

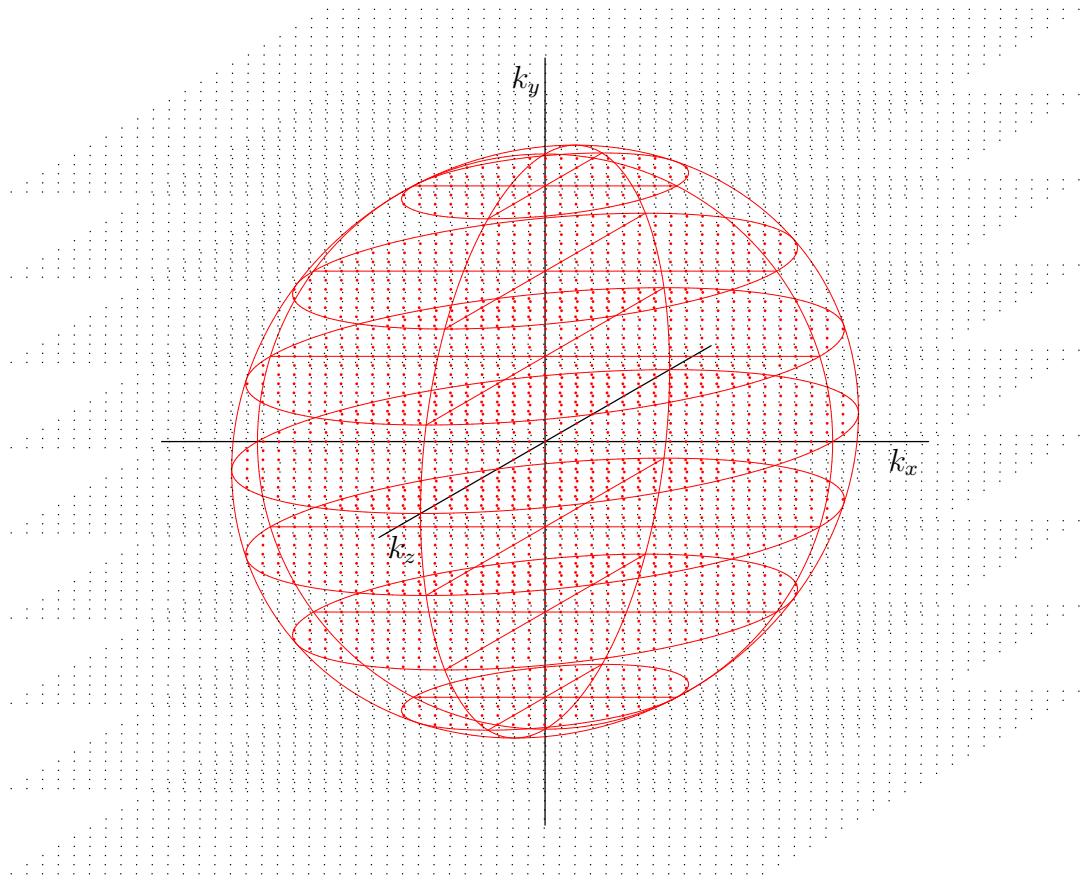


Figure 5.17: Ground state of a system of noninteracting electrons, or other fermions, in a periodic box.

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:

$$\psi_{n_x n_y n_z}^P(\vec{r}) = \psi_{k_x k_y k_z}^P(\vec{r}) = \psi_{p_x p_y p_z}^P(\vec{r})$$

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 $\psi_{000}^P(\vec{r})$; 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.
-

5.19 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 5.3 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 (5.3) and (5.28) 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 (5.5). That gives the number of single-particle states dN in a momentum range of size dp as:

$$dN = \mathcal{V}\mathcal{D}_p dp \quad \mathcal{D}_p = \frac{2s+1}{2\pi^2\hbar^3} p^2 \quad (5.29)$$

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 5.12, 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.
-

5.20 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 5.18 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 $E^P = \frac{1}{2}m_e v^P x^2$ for the kinetic energy: for the single-particle states on the outer boundary, the velocity is purely in the x -direction.

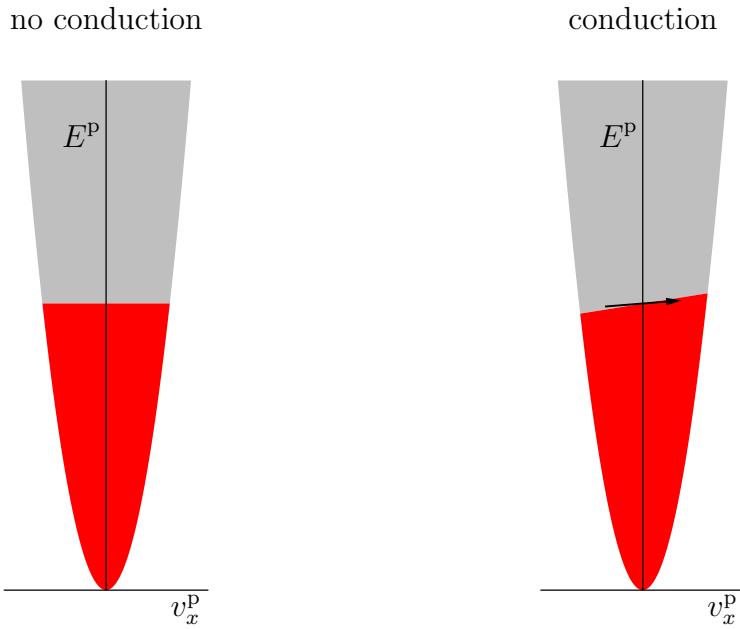


Figure 5.18: Conduction in the free-electron gas model.

In the system ground state, shown to the left in figure 5.18, 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 5.18. 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.

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 5.17, 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 5.18. 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 note {A.52} 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 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 Δv_x^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, non-quantum, 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 $\sqrt{k_B T/m_e}$, with k_B the Boltzmann constant, T the absolute temperature, and m_e the electron mass. In addition to this random thermal motion in all directions, the electrons would also have a small organized drift velocity Δv_x^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_e , times the electric charge $-e$ that each carries, times the small organized drift velocity Δv_x^P in the x -direction that each has:

$$j_x = -i_e e \Delta v_x^P \quad (5.30)$$

The drift velocity Δv_x^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 $-eE_x$ where 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,

$$\sigma \sim \frac{i_e e^2 \ell}{m_e v} \quad (5.31)$$

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 5.18. 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 5.10. 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 5.22.3.

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.

5.21 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.
-

5.21.1 Metals and insulators

To understand electrical conduction in solids requires consideration their electron energy levels.

Typical energy spectra are sketched in figure 5.19. The spectrum of a free-electron gas, noninteracting electrons in a box, is shown to the left. The energy E^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.

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

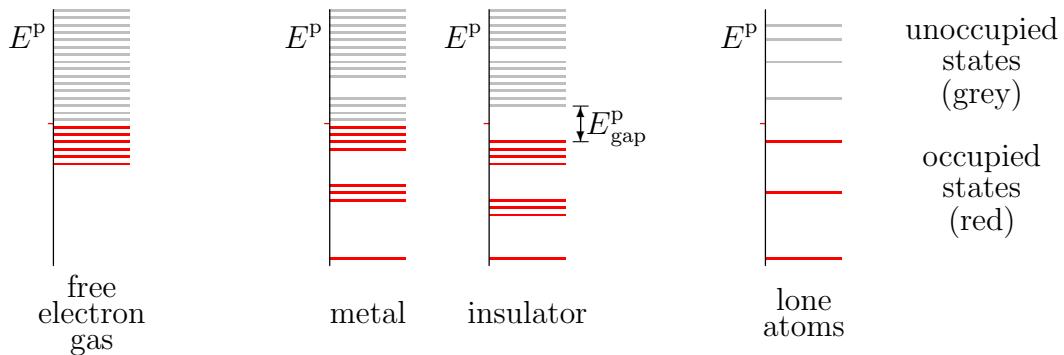


Figure 5.19: Sketch of electron energy spectra in solids at absolute zero temperature. (No attempt has been made to picture a density of states). Far left: the free-electron gas has a continuous band of extremely densely spaced energy levels. Far right: lone atoms have only a few discrete electron energy 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.

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 crystal 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 5.19. One basic example is the hydrogen spectrum shown in figure 3.2. 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 5.19. 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 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 5.19, 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.
-

5.21.2 Typical metals and insulators

The next question is what materials completely fill up their valence band and so are insulators. That is much more tricky.

First, for solids made up of a single element it normally requires that the number of valence electrons per atom is even. That is because the number of spatial states within a band is normally a whole multiple of the number of atoms. (When widely spaced atoms in a periodic box are brought closer together to form a solid, the single-particle states change in shape and energy, but not in number. One spatial state per atom stays one spatial state per atom.) Since each spatial state can be occupied by two electrons, one with spin up and one with spin down, it requires an even number of electrons to fill up a band.

The alkaline metals found below helium in group II of the periodic table figure 4.8 have two valence electrons per atom. However, they do not fill up their valence band and are therefore metals. That happens because the filled band, originating from the atomic s states, merges with an empty band originating from the p states. That results in an unfilled combined band. On the other hand, group IV elements like diamond, silicon, and germanium do fill up their valence band with their four valence electrons. They are insulators at absolute zero temperature.

A solid may of course consist of more than one element. And even for a single element, atoms may combine in groups. In general, crystalline solids consist of basic building blocks called “unit cells” that can involve several atoms. Band theory predicts the solid to be a conductor if the number of electrons per unit cell is odd. If the number of electrons per cell is even, it may be an insulator. That explains how hydrogen and group V elements can fill up energy bands with an odd number of valence electrons *per atom*: their atoms associate in pairs in the crystal structure. With two atoms per cell, there are an even number of valence electrons *per cell*.

Indeed, hydrogen is a nonmetal even though it has only one valence electron per atom just like lithium and the other alkali metals below it in group I. If you think of solid hydrogen as build up of single atoms, it would be hard to understand why there could be a filled band with only one electron in each 1s state. However, if you think of solid hydrogen as build up of hydrogen molecules, as it is, the band gap becomes clear. In the hydrogen molecule, the two atomic 1s states of equal energy are converted into a lower energy molecular state in which the 1s electrons are symmetrically shared, and a higher energy state in which they are antisymmetrically shared, chapter 4.2. The lower energy state

is filled with the two 1s electrons while the higher energy one could also hold two electrons but is empty. In the solid, the hydrogen molecules are barely held together by weak Van der Waals forces. The interactions between the molecules are weak, so the two energy levels broaden only slightly into two thin bands. The gap between the filled symmetric states and the empty antisymmetric ones remains.

(To be precise, the method of sharing two electrons implies a nontrivial interaction between the electrons. That violates the band-theory idea of noninteracting electrons. A lowered-energy state in which two electrons are symmetrically shared is not quite the same as a lowered-energy single-electron state that holds two noninteracting electrons. However, as an approximation it will be assumed that it is. Interactions between electrons would make the analysis extremely difficult. In the current analysis interactions must be roughly represented through ad-hoc corrections to the single-particle energies.)

In any case, it is believed that under extremely high pressures, it is possible to break up the molecules in hydrogen. In that case, the corresponding gap in energy levels would disappear, and hydrogen would 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 should also be noted that diamond, silicon, and germanium pull a similar trick as hydrogen. While their atoms do have four valence electrons, that is still only half of what is required to fill up the four sp^3 -hybrid spatial states available to them. Much like hydrogen, symmetric and antisymmetric bonding states are formed from hybrid states connecting neighboring atoms. This splits the energy band again into two. However, since each atom makes bonds with three neighboring atoms rather than one, the structure does not fall apart into two-atom molecules. In fact, the entire solid can be thought of as one big covalently-bound molecule. Still, while the smallest unit cell of the crystal structure is not unique, it does always contain exactly two atoms just like for hydrogen.

Where hydrogen refuses to be a metal in group I of the periodic table, boron does so in group III. 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. Even more convoluted, under very high pressures boron appears to become *ionic*.

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 band theory would predict that they are conductors.

Key Points

- In the periodic table, the group I, II, and III elements are usually metals.
 - However, hydrogen and helium are nonmetals. Don’t ask about boron.
 - The group IV elements diamond, silicon, and germanium are insulators at absolute zero temperature.
-

5.21.3 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 5.19 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 5.20 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.

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.

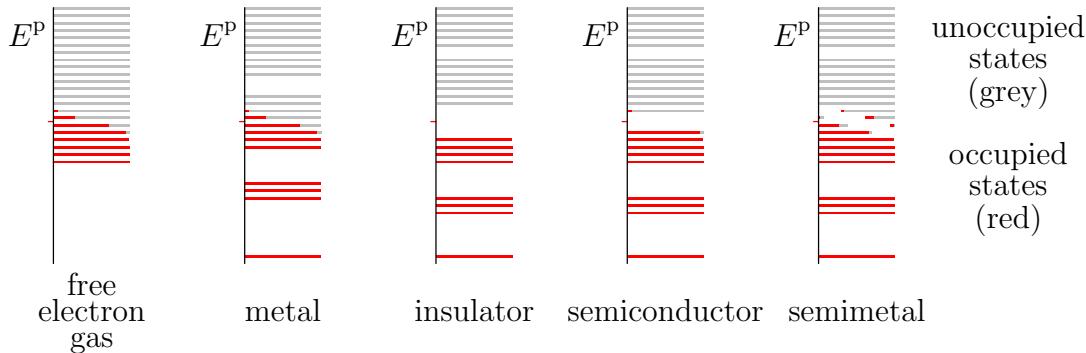


Figure 5.20: Sketch of electron energy spectra in solids at a nonzero temperature.

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.

5.21.4 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 5.19, 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 5.20.

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

- o— Semimetals have properties intermediate between metals and semiconductors.
-

5.21.5 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 because it takes too much energy to excite them. However, atomic vibrations in solids can conduct heat too, so the differences in heat conduction between solids are not by far as large as those in electrical conduction. For example, diamond, an excellent electrical insulator, is also a superb conductor of heat. Since heat conduction is not a monopoly of the electrons, but the atoms can do it too, there are no thermal insulators that are anywhere near as effective as electrical insulators. Practical thermal insulators are highly porous materials whose volume consists largely of voids.

Key Points

- o— Electrons conduct heat very well, but atoms can do it too.

- Practical thermal insulators use voids to reduce atomic heat conduction.
-

5.21.6 Ionic conductivity

It should be mentioned that electrons do not have an absolute monopoly on electrical conduction. In ionic solids a small amount of conduction may be possible due to motion of the charged atoms. This requires defects in the crystal structure, in particular locations with missing atoms, in order to give the atoms some room to move. Obviously, such conduction cannot compete with that in metals, though some ionic solids can compete with ionic liquids.

Key Points

- In some solids, a bit of electrical conduction may occur through the motion of ions instead of electrons.
-

5.22 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.

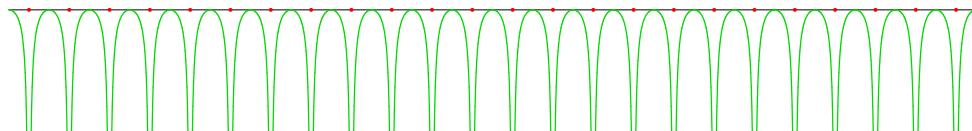


Figure 5.21: Potential energy seen by an electron along a line of nuclei. The potential energy is in green, the nuclei are in red.

The effect of the crystal is to introduce a periodic potential energy for the electrons. For example, figure 5.21 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.

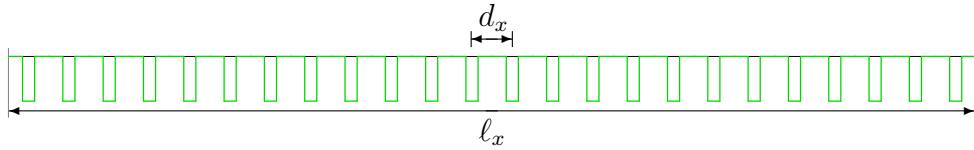


Figure 5.22: Potential energy seen by an electron in the one-dimensional simplified model of Kronig & Penney.

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 energy is piecewise constant, like in figure 5.22. You might think of the regions of low 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 . 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 2.5.8. This section will however limit itself mostly to the one-dimensional case.

5.22.1 Bloch waves

The subsection examines the one-dimensional single-particle quantum states, or energy eigenfunctions, of electrons in solids.

For free electrons, the energy eigenfunctions were given in section 5.18. In one dimension they are:

$$\psi_{n_x}^p(x) = C e^{ik_x x}$$

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 eigenstates do not just have a definite energy $E_x^p = \hbar^2 k_x^2 / 2m_e$, they also have definite linear momentum $p_x = \hbar k_x$. In classical terms, the electron velocity is given by the linear momentum as $v_x^p = p_x / m_e$.

To find the equivalent one-dimensional energy eigenfunctions $\psi_{n_x}^p(x)$ in the presence of a crystal potential $V_x(x)$ is messy. It requires solution of the one-

dimensional Hamiltonian eigenvalue problem

$$-\frac{\hbar^2}{2m_e} \frac{\partial^2 \psi^p}{\partial x^2} + V_x \psi^p = E_x^p \psi^p$$

where E_x^p 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, {A.36}.

However, it can be shown that the eigenfunctions will always be of the form:

$$\boxed{\psi_{n_x}^p(x) = \psi_{p,n_x}^p(x) e^{ik_x x}} \quad (5.32)$$

in which $\psi_{p,n_x}^p(x)$ is a periodic function on the atomic period. Note that as long as $\psi_{p,n_x}^p(x)$ 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 $\psi_{p,n_x}^p(x)$. In that case, all that can be said a priori is that it is periodic on the atomic period.

Energy eigenfunctions of the form (5.32) 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 6.5.5. 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 (5.28). 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^p with respect to the “crystal momentum” $p_{cm,x} = \hbar k_x$. That is the same as for the free-electron gas. If you differentiate the one-dimensional free-electron gas kinetic energy $E_x^p = (\hbar k_x)^2 / 2m_e$ with respect to $p_x = \hbar k_x$, you get the velocity.

Key Points

- In the presence of a crystal potential, the periodic box 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.

5.22.2 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.

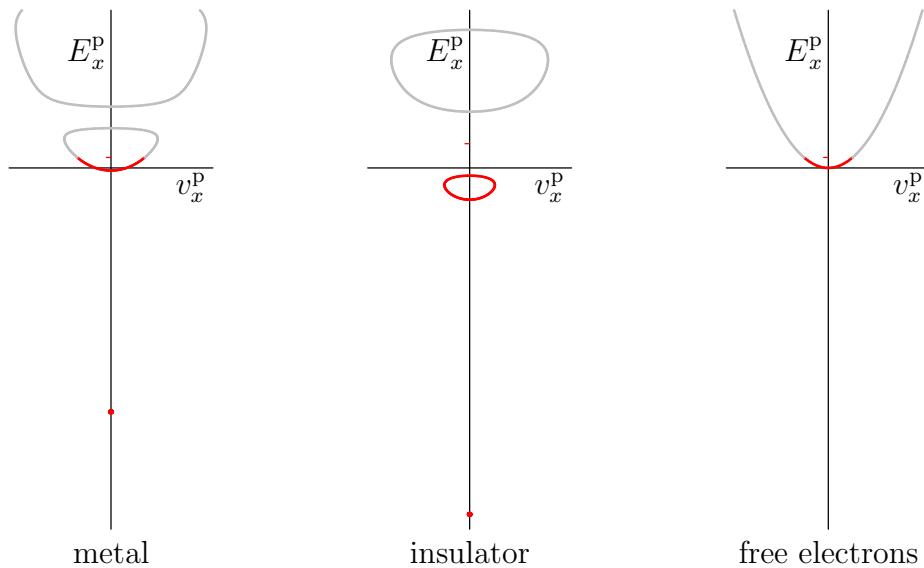


Figure 5.23: Example Kronig & Penney spectra.

Finding the energy eigenvalues is not difficult on a computer, {A.36}. A couple of example spectra are shown in figure 5.23. 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 5.18, but the axes are much more compressed here.) Quantum states occupied by electrons are again in red.

The example to the left in figure 5.23 tries to roughly model a metal like lithium. The depth of the potential drops in figure 5.22 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 5.23. 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 5.18, 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 5.23 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 one dimension.) 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 5.23 are not obvious. Note {A.36} explains by example what happens to the free-electron gas energy eigenfunctions when there is a crystal potential.

Key Points

- A periodic crystal potential produces energy bands.
- Inner atomic electrons would not move even if there were unfilled states.

5.22.3 Effective mass

The spectrum to the right in figure 5.23 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:

$$E_x^p = \frac{1}{2}m_e v_x^{p2}$$

This leads to the parabolic spectrum shown.

It is interesting to compare this spectrum to that of the “metal” to the left in figure 5.23. 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

$$E_x^p - E_{c,x}^p = \frac{1}{2}m_{\text{eff},x} v_x^{p2}$$

where $E_{c,x}^p$ is the energy at the bottom of the conduction band and $m_{\text{eff},x}$ 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 5.23, 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 5.23, 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, [29, 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 5.3, if you substitute in the effective mass.

To see why, note that if the relationship between the energy E_x^p and the velocity v_x^p is the same as that for a free-electron gas, then so is the relationship between the energy E_x^p and the wave number k_x . (At least, it is the same if you measure the wave number k_x from the location of minimum conduction band energy.) That is because the velocity is merely the derivative of E_x^p with respect to $\hbar k_x$. And if the E_x^p versus k_x relation is the same, then so is the density of states, since the energy states are spaced equally in terms of k_x 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 free-electron gas density of states can also be used.

5.22.4 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.

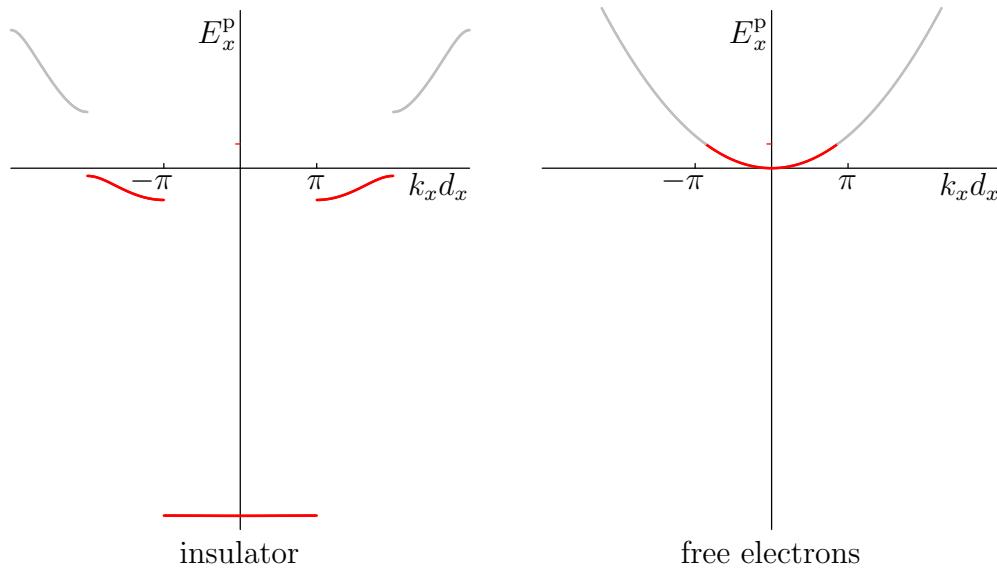


Figure 5.24: Spectrum against wave number in the extended zone scheme.

Figure 5.24 shows the energy plotted against the crystal momentum. Note that this becomes a plot against the wave number k_x since the crystal momentum $p_{cm,x}$ is defined to be $\hbar k_x$. 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,

$$\psi_{n_x}^p(x) = \psi_{p,n_x}^p(x)e^{ik_x x}$$

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

the periodic part ψ_{p,n_x}^P . The periodic part changes, but it remains periodic on the atomic scale.

Therefore there is a problem with how to define a unique value of k_x . There are different solutions to this problem. Figure 5.24 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.

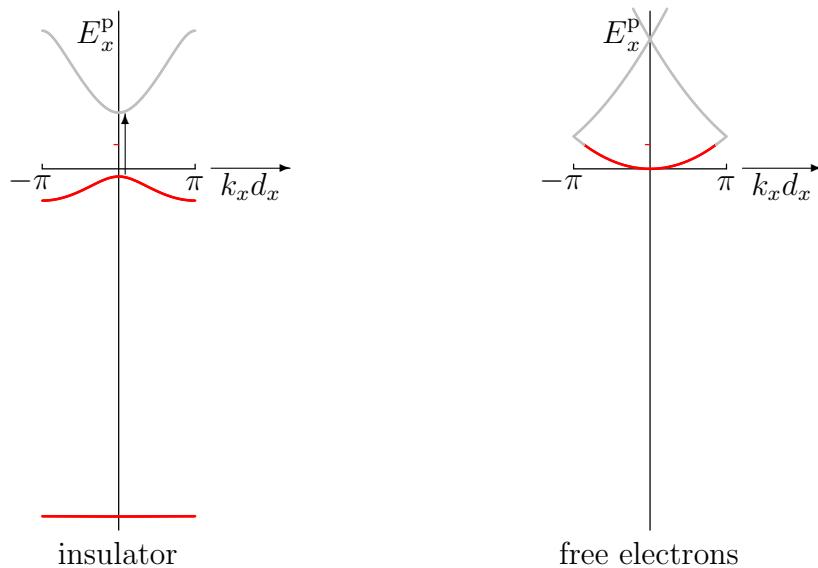


Figure 5.25: Spectrum against wave number in the reduced zone scheme.

A second approach is much more common, though. It uses the indeterminacy in k_x to shift it into the range $-\pi < k_x d_x < \pi$. That range is called the “first Brillouin zone.” Restricting the wave numbers to the first Brillouin zone produces figure 5.25. (Showing the range $0 < k_x d_x < 2\pi$ is a common alternative.) 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 $e^{ik_x x}$.

Second, the reduced zone scheme is important for optical applications of semiconductors. In particular,

Lone photons can only produce electron transitions along the same vertical line in the reduced zone spectrum.

Consider a photon with an energy that is just a bit more than the band gap. The only thing for which such a photon has enough energy is indicated by the vertical arrow in the spectrum to the left in figure 5.25. It can take an electron out of the very top of the valence band and put it in the very bottom of the conduction band. The photon may be able to do so without too much trouble if the band structure is as in figure 5.25. Here the lowest point in the conduction band is straight above the highest point in the valence band. A semiconductor like this is called a “direct gap semiconductor.”

But the two most basic semiconductors, silicon and germanium, have the lowest point in the conduction band to the side of the highest point in the valence band. These are called “indirect gap semiconductors.” A photon that takes an electron from the top of the valence band to the bottom of the conduction band must now change the electron’s crystal momentum $\hbar k_x$. But crystal momentum must be conserved. That is much like ordinary momentum must be preserved for a system of particles in empty space. And the photon itself has negligible crystal momentum. There is therefore a problem: where does the difference in crystal momentum go? A phonon of crystal vibration has to get involved to carry it off (or supply it).

The involvement of the additional phonon makes the entire process much more cumbersome. Silicon and germanium are not good semiconductors for some optical applications like the production of light.

It may be noted that conservation of crystal momentum is often called “conservation of wavevector.” It is the same thing of course, since the crystal momentum $p_{cm,x}$ is simply $\hbar k_x$. 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 even unique, the answer is in the discussion of conservation laws in chapter 6.2. It is not really linear momentum that is conserved, but the product of the single-particle eigenvalues $e^{ik_x d_x}$ 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.)

The so-called “periodic zone scheme” takes the reduced zone scheme and extends it periodically, as in figure 5.26. That makes for very esthetic pictures, especially in three dimensions.

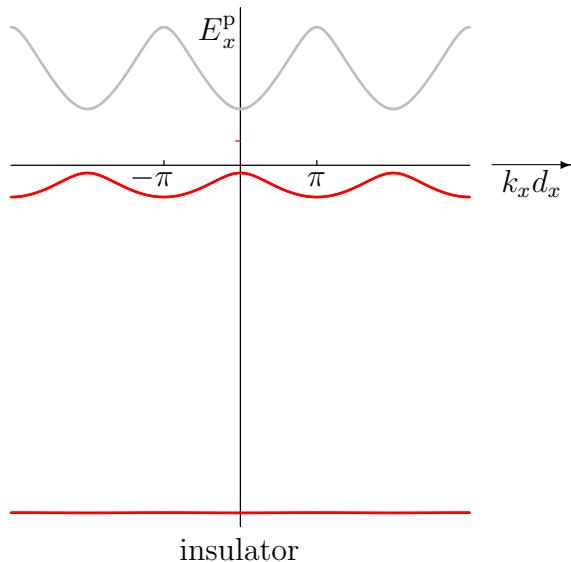


Figure 5.26: Spectrum against wave number in the periodic zone scheme.

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 principal crystal direction to be comprehensive.

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.
- □ 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.

5.23 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.

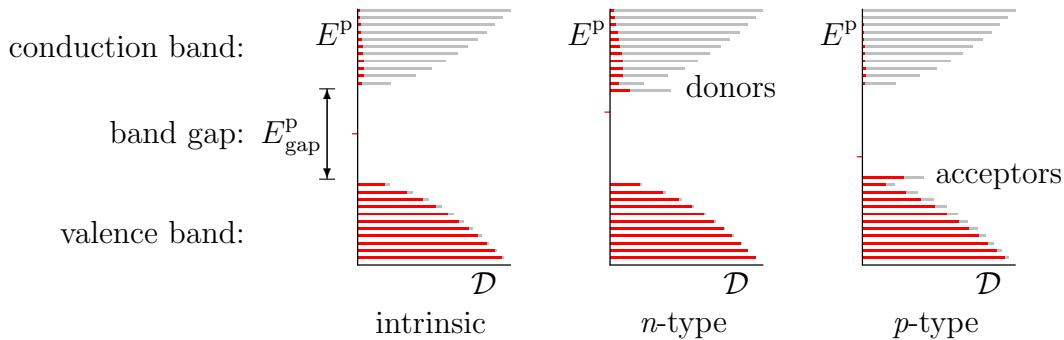


Figure 5.27: Vicinity of the band gap in the spectra of intrinsic and doped semiconductors. The amounts of conduction band electrons and valence band holes have been vastly exaggerated to make them visible.

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 5.27. 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 (5.19). In the conduction band, that may be simplified to the Maxwell-Boltzmann one (5.21) because the number of electrons in the conduction band is small. The average number of electrons per state in the conduction band is then:

$$\boxed{\iota_e = e^{-(E^p - \mu)/k_B T}} \quad (5.33)$$

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 5.27.

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

$$\text{Bottom of the conduction band: } \nu_e = e^{-E_{\text{gap}}^{\text{P}}/2k_B T}$$

At room temperature, $k_B T$ is about 0.025 eV while for silicon, the band gap energy is about 1.12 eV. That makes ν_e about $2 \cdot 10^{-10}$. 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 ν_e of about $3 \cdot 10^{-6}$ 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:

$$\boxed{\nu_h = e^{-(\mu - E^{\text{P}})/k_B T}} \quad (5.34)$$

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 $\nu_h = 1 - \nu_e$. If you plug the Fermi-Dirac distribution into that, you get the expression for ν_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 ν_e and ν_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, ι_e at the bottom of the conduction band is the same as ι_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 came from.

Note that figure 5.27 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 actuality, 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 ι_e and ι_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^p or μ changes just a little bit, $e^{-(E^p - \mu)/k_B T}$ 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, {A.37}:

$$\boxed{i_e = 2 \left(\frac{m_{\text{eff},e} k_B T}{2\pi\hbar^2} \right)^{3/2} e^{-(E_c^p - \mu)/k_B T} \quad i_h = 2 \left(\frac{m_{\text{eff},h} k_B T}{2\pi\hbar^2} \right)^{3/2} e^{-(\mu - E_v^p)/k_B T}} \quad (5.35)$$

Here E_c^p and E_v^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 (5.35) is the value of the number of electrons per state ι_e at the bottom of the conduction band, and the second exponential is the number of holes per state ι_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 5.27 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 the 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 5.27. 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 not move up, then the total number of electrons in the conduction band would not change. And there would be extremely few electrons in the donor states for that Fermi level. That is not possible, because all the other donor electrons cannot just disappear. In fact, the amount of electrons contributed by the donor states is dramatic; that is because there are so extremely few conduction electrons in the intrinsic case. The Fermi level has to move up significantly to explain the increase. An increase in Fermi level μ

increases the number of electrons per quantum state (5.33) in the conduction band. It has to go up far enough that the combined number of electrons in the conduction band and the donor states is just a bit more than the number of donor electrons.

But the Fermi level 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 $k_B T$. 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 states compared to a typical number of donors, the number of electrons in them would dwarf the number of electrons missing from the donor states. And that is not possible since the total number of electrons cannot exceed the number of donor states by any noticeable amount. So the Fermi level has to stay low enough that the number of electrons ν_e stays small in both the donor states and conduction band. That is as sketched in figure 5.27.

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, 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 upward shift in the Fermi level in *n*-type material has another effect. It decimates the already miserably small number of holes in the valence band, (5.34). Therefore the number of conduction band electrons provided by the valence band becomes many times smaller still than it was already for the intrinsic semiconductor. Essentially all conduction band electrons are provided by the donors. Also, almost 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.”

The fact that raising the amount of conduction band electrons lowers the amount of valence band holes may be verified mathematically from (5.35). That equation implies that the product of the electron and hole densities is constant

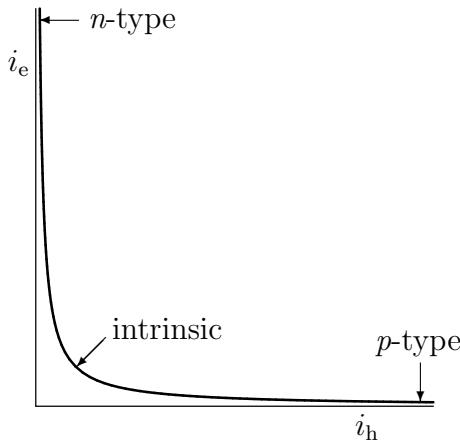


Figure 5.28: Relationship between conduction electron density and hole density. Intrinsic semiconductors have neither much conduction electrons nor holes.

at a given temperature:

$$i_e i_h = 4 \left(\frac{\sqrt{m_{\text{eff},e} m_{\text{eff},h}} k_B T}{2\pi\hbar^2} \right)^3 e^{-E_{\text{gap}}^p/k_B T} \quad (5.36)$$

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 5.28. 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 *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.

The law of mass action can also be understood from more classical arguments. That is useful since band theory has its limits. In thermal equilibrium, the semiconductor is bathed in blackbody radiation. A very small but nonzero fraction of the photons of the 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 a 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 for holes. Equating a creation rate A of electron-hole pairs by photons to a removal rate of the form $B i_e i_h$ shows that the product $i_e i_h$ is constant. The constant A/B will depend primarily on the Maxwell-Boltzmann factor $e^{-E_{\text{gap}}^{\text{P}}/k_B T}$ that limits the number of photons that have sufficient energy to create pairs.

This 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 5.28 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 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.

5.24 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.

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 5.29, 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

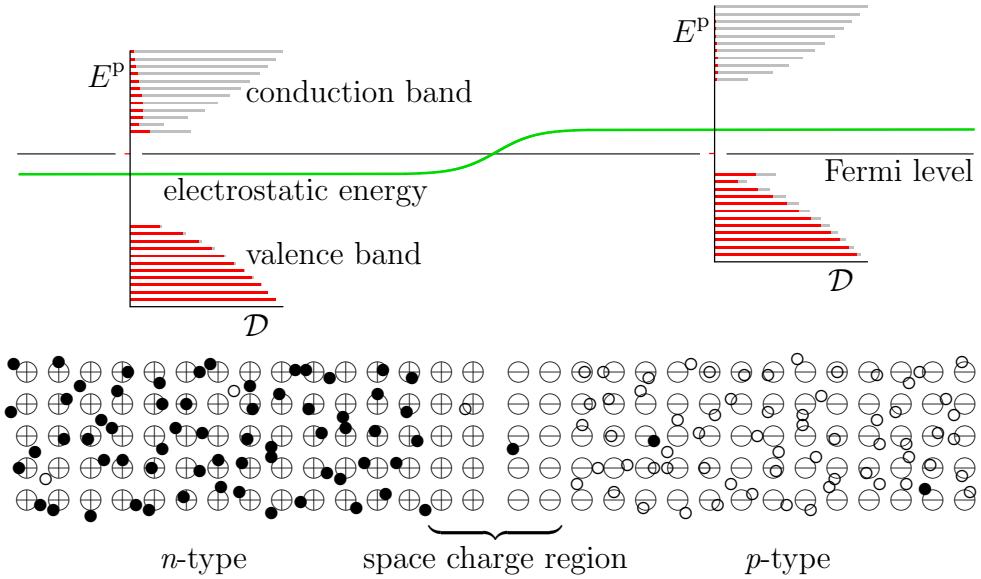


Figure 5.29: The p - n junction in thermal equilibrium. Top: energy spectra. Quantum states with electrons in them are in red. The mean electrostatic energy of the electrons is in green. Below: Physical schematic of the junction. The dots are conduction electrons and the small circles holes. The encircled plus signs are donor atoms, and the encircled minus signs acceptor atoms. (Donors and acceptors are not as regularly distributed as this greatly simplified schematic indicates).

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 electrons that the holes have lost 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, 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.”

Consider first the *p-n* junction in thermal equilibrium, when there is no current. The junction is shown in the bottom of figure 5.29. 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 negative charge of the ionized 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. Both 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 an electrostatic barrier that 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. It is called the “space charge region.” 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 5.28: 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 energy is shown in green in figure 5.29. It raises the electron energy levels in the *p* side relative to the *n* side until the chemical potentials, or Fermi levels, at the two sides become equal. They have to do so; differences in chemical potential produce electron diffusion, section 5.16. For the diffusion to stop, the chemical potential must become everywhere the same.

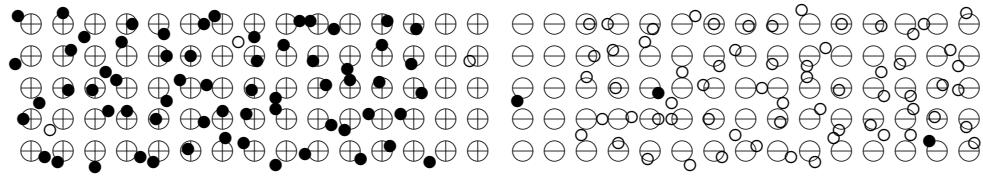
Consider now figure 5.30a. All those *n*-side electrons would love to diffuse into the *p* side, but in equilibrium the electrostatic barrier holds them back. The fraction of *n*-side electrons that do have enough energy to get through the barrier is very small. It is primarily determined by the Maxwell-Boltzmann factor (5.33). If V_j is the electrostatic energy increase over the junction, then there are a factor $e^{-V_j/k_B T}$ less electrons per state with the additional energy V_j than there are at the bottom of the conduction band. The crossings of these lucky few electrons produces a minuscule current through the junction that is indicated as $j_{e,\text{maj}}$ in figure 5.30a. 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 is the direction that the electrons actually move.

The minuscule current of the *n*-side majority electrons is balanced by an equally minuscule but opposite current $j_{e,\text{min}}$ 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 in 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 you apply a “forward-bias” external voltage φ that lowers the barrier by an amount $e\varphi_j$. What happens then is shown in figure 5.30b. 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



a) No voltage applied:



junction crossings by electrons:

$$-j_{e,\min} \leftarrow \rightarrow -j_{e,\text{maj}}$$

junction crossings by holes:

$$j_{h,\text{maj}} \leftarrow \rightarrow j_{h,\min}$$

b) Forward biased:



junction crossings by electrons:

$$-j_{e,\min} \leftarrow \rightarrow -j_{e,\text{maj}}$$

junction crossings by holes:

$$\leftarrow j_{h,\text{maj}} \rightarrow j_{h,\min}$$

net current:

$$\overrightarrow{-j_{e,\text{net}}} + \overrightarrow{-j_{h,\text{net}}} = \overrightarrow{-j_{\text{net}}}$$

c) Reverse biased:



junction crossings by electrons:

$$-j_{e,\min} \leftarrow \rightarrow -j_{e,\text{maj}}$$

junction crossings by holes:

$$j_{h,\text{maj}} \leftarrow \rightarrow j_{h,\min}$$

net current:

$$\overleftarrow{-j_{e,\text{net}}} + \overleftarrow{-j_{h,\text{net}}} = \overleftarrow{-j_{\text{net}}}$$

Figure 5.30: Schematic of the operation of an *p-n* junction.

factor $e^{e\varphi_j/k_B T}$. That is a very large factor if the voltage change is bigger than about 0.025 volt, since $k_B T$ is about 0.025 eV. The currents of majority carriers explode, as sketched in the figure. And therefore, so does the net current.

Figure 5.30c 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. The only thing 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 density through the junction:

$$j = j_0 e^{e\varphi_j/k_B T} - j_0 \quad (5.37)$$

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. Whatever minority carriers reach the junction will cross to the other side. The other term is the net forward current due to the majority carriers. According to the above discussion, it differs from the minority current by the Maxwell-Boltzmann factor of the applied voltage.

For forward bias the exponential in (5.37) explodes, producing significant current. For reverse bias, the exponential is essentially zero and only the small reverse minority current is left.

Equation (5.37) is known as the “Shockley diode equation.” It works well for germanium but not quite that well for semiconductors like silicon with higher band gaps. Recombination of carriers inside the space charge region is an important issue. Also, the creation of additional electron-hole pairs is not as impossible as the above simple story assumes. A fudge factor called the “ideality factor” is often added to the argument of the exponential to improve agreement.

Note that the diode 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. In any case, the equation only applies for a limited range of voltages. The derivation assumed that the forward voltage is small enough that barrier does not invert. And if the reverse voltage is too large, phenomena discussed in section 5.26 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, especially in germanium.
 - At the junction a space-charge region exists. It provides a barrier against the majority carriers. However, it accelerates the minority carriers passing through the junction.
-

5.25 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.

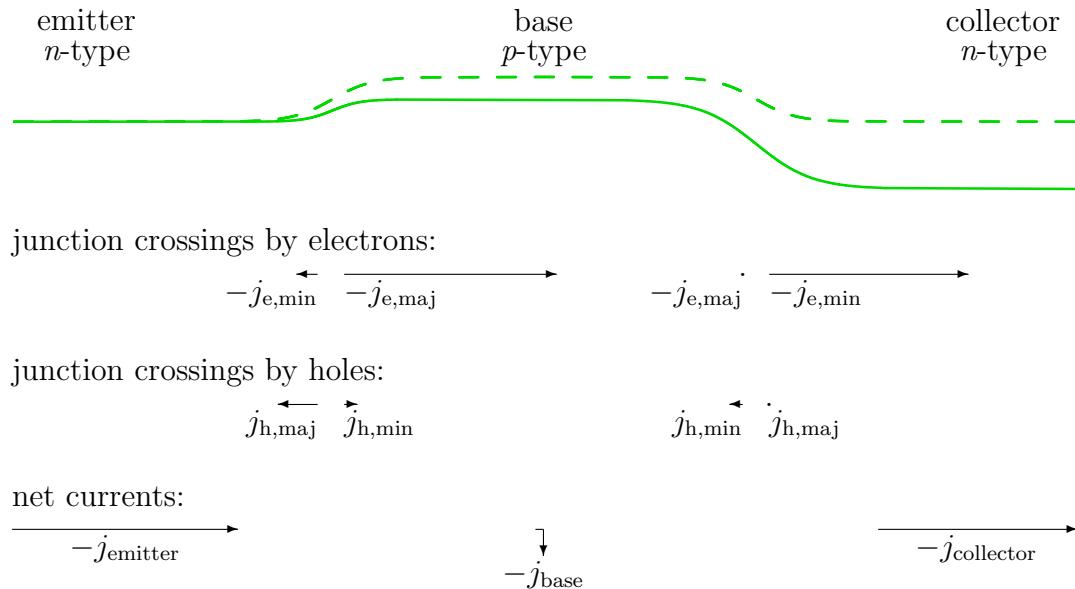


Figure 5.31: Schematic of the operation of an *n-p-n* transistor.

For example, an *n-p-n* transistor allows the flow of electrons through an *n*-type semiconductor to be controlled. A schematic is shown in figure 5.31. Electrons flow through *n*-type semiconductor material from one side of the transistor, called the “emitter” to the other side, called the collector.

To control this current, a very thin 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 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 5.31. 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 minority carriers in the base, they get swept 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 current 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 current, 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

- o A transistor allows current to be regulated.

5.26 Zener and avalanche effects

Section 5.24 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 figure 5.29. 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.

One thing that can happen is that electrons in the valence band on the *p* side in figure 5.29 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 spectrum at one location is raised sufficiently that its valence band reaches the energy level of the conduction band at a nearby 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 5.29, that raises the *p*-side spectra relative to the *n*-side ones. To raise a spectrum significantly relative to one very nearby therefore 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. Zener diodes are therefore used to provide a very stable voltage source. The diode can be 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 because a corresponding change over the Zener would require 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 in the *n* side and conduction electrons in 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. The minority carriers get accelerated by the space charge region that exists at the junction. If the reverse voltage is big enough, the space charge region can accelerate the minority carriers enough 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 passing through the space charge region has picked up enough energy near the end of it to knock an electron out of the valence band. The created hole is swept by the electric field in the opposite direction of the two electrons, 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 electron is swept by the electric field in the opposite direction of the two holes, back into the space charge region...

In short, the tiny current of minority carriers now explodes in an avalanche effect. A diode designed to survive this is properly called an “avalanche diode.” However, avalanche diodes are often loosely referred to as Zener diodes even though the process is different.

Key Points

- o— Unlike the idealized theory suggests, under suitable conditions, significant reverse currents can be made to pass through p - n junctions.
 - o— It allows voltage regulation.
-

5.27 Optical applications

This section gives an overview of optical physics ranging from the x-ray spectrum of solids to semiconductor devices such as solar cells and light-emitting diodes.

5.27.1 Atomic spectra

As sketched earlier in the spectra of figure 5.19, lone atoms and molecules have discrete energy levels. Electrons that transition between these levels emit and absorb electromagnetic energy at corresponding discrete frequencies. A basic example is the red line of the E_3 to E_2 Balmer transition of the hydrogen atom, chapter 3.2. In general, when the light emitted by lone atoms is send through a prism, it separates into a few discrete thin beams of very specific colors.

5.27.2 Spectra of solids

In solids, and especially metals, the electrons can transition between energy levels with a broad range of energies. Therefore a solid can emit visible light at a continuum of frequencies. Send through a prism, such light will spread out into a band of gradually changing color. That is called “broadband” radiation.

To be sure, transitions involving the inner electrons in solids still produce radiation at discrete frequencies. The reason is that the energies of the inner electrons do not broaden significantly from the discrete atomic values. 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.

5.27.3 Band gap effects

Further, it is not true that solids can absorb and emit all visible light. 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 5.32. A photon of light with enough energy can take an electron out of the valence band and put it into the conduction band. That leaves a hole behind in the valence band. 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 (red) to 3.2 eV (violet). 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.

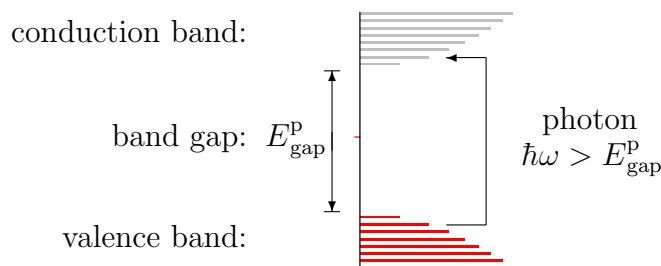


Figure 5.32: Vicinity of the band gap in the spectrum of an insulator. A photon of light with an energy greater than the band gap can take an electron from the valence band to the conduction band. The photon is absorbed in the process.

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 5.32 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. Ideally, the energy levels of lone excitons should be discrete like those of the hydrogen atom. However, they broaden considerably in the less than perfect 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.

5.27.4 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. This 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.

5.27.5 Photoconductivity

When photons of light do have enough energy to take electrons to the conduction band, both the created conduction electrons and the holes 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 produce 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.

5.27.6 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 the photons is sunlight.

To understand how they work, consider the schematic of a *p-n* junction in figure 5.29. 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 5.29. 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 current 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. Whatever excess energy 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 frequencies of the incoming light can significantly improve efficiency.

5.27.7 Light-emitting diodes

The defining feature of the photovoltaic effect is that light hitting a *p-n* junction 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 5.29. 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.

Silicon or germanium are not suitable. That has to do with conservation of “crystal momentum,” as defined in section 5.22.4. Of course, you cannot just apply momentum conservation to electrons in solids as if they were in free space. The crystal structure exerts big forces on the electrons. However, a degenerate form of momentum called crystal momentum is still conserved. At least it is to the approximation to which the crystal exerts perfectly periodic forces on the electrons on an atomic scale,

Now for silicon and germanium, the lowest energy states of the conduction band, where most conduction electrons are, have a different crystal momentum than the highest energy states of the valence band, where most holes are. For that reason the band gap of silicon and germanium is called “indirect.” It implies that the crystal momentum of an electron changes when it drops into a hole. Since crystal momentum is conserved, something else has to absorb the difference. And photons have negligible crystal momentum. It is a phonon of crystal vibration that must absorb the momentum difference. The required presence of the phonon makes the entire process cumbersome. Therefore the recombination is likely to proceed according to some different mechanism that releases the energy as heat rather than light.

LEDs use “direct” band gap semiconductors. For these, the lowest states in the conduction band and the highest states in the valence band have the same crystal momentum. Then no phonon is required and the creation of light instead of heat is much more likely. The classical direct gap material is gallium arsenide, which produced the first red LEDs.

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 there are enough available conduction electrons and holes, and if the stimulated emission process is dominant enough. Under these conditions a photon may produce another photon through stimulated emission, then the two of them go on to produce two more photons for a total of four, 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 6.3.2. However, the requirement of a population inversion in that discussion does not apply to semiconductor lasers, as they are not collections of independent two-state systems.

Key Points

- Solids emit and absorb electromagnetic radiation in continuous energy bands.
 - The X-ray range of the inner electrons is still discrete.
 - A perfect crystal of a solid with a large enough band gap will be transparent.
 - The colors of most nonmetals are caused by crystal imperfections.
 - An exciton is a bound state of an electron and a hole.
 - An electron bound to a vacancy in a ionic crystal is a called an *F*-center.
 - Photoconductivity is the increase in conductivity of nonmetals when photons of light create additional charge carriers.
 - Photovoltaics is the creation of electricity by photons. Solar cells are an important example.
 - 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.
-

5.28 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 5.23. 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.

5.28.1 Peltier effect

Thermoelectric cooling can be achieved through what is called the “Peltier effect.” The top part of figure 5.33 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 5.33 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. (Note that since electrons are negatively charged, their current is in the direction opposite to their motion.) The same current that enters a metal strip through one block leaves the strip again at the other block.

Consider now a metal strip at the top of the device in figure 5.33. 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 spectra at the bottom of figure 5.33, 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 5.33 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

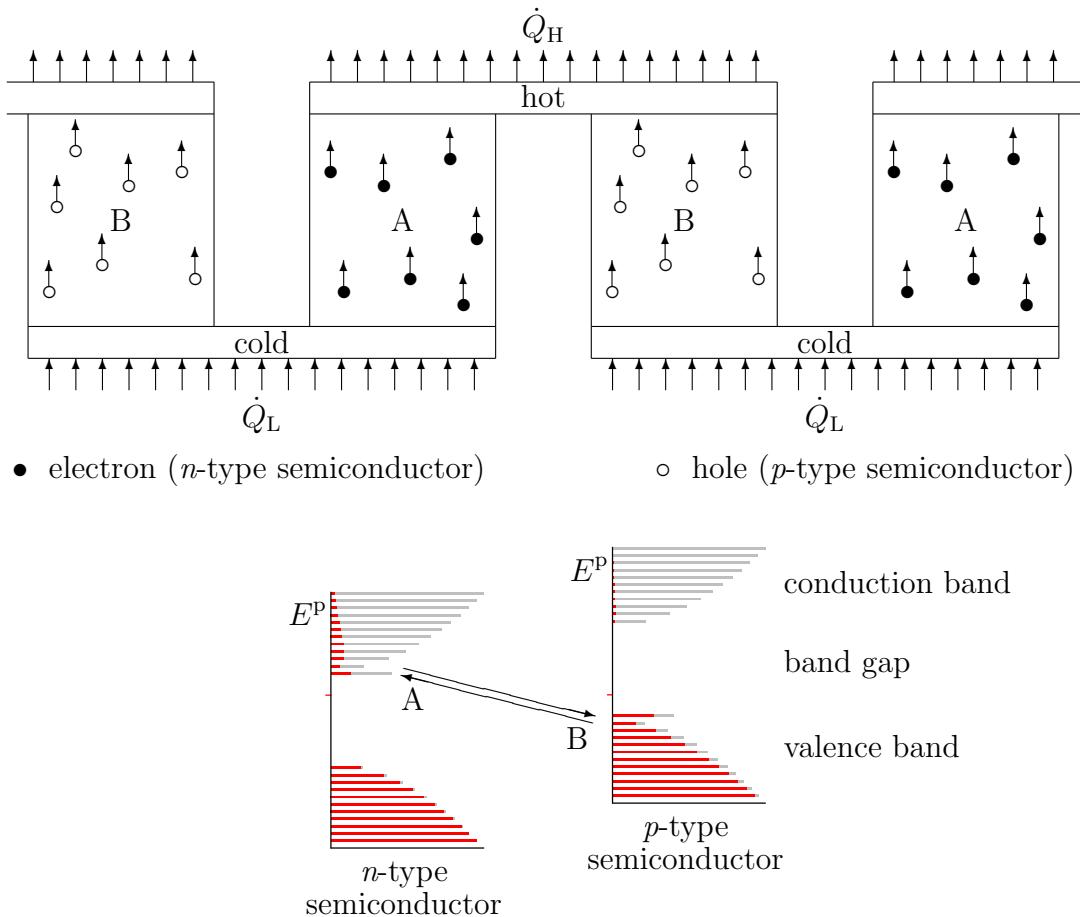


Figure 5.33: Peltier cooling. Top: physical device. Bottom: Electron energy spectra of the semiconductor materials. Quantum states filled with electrons are shown in red.

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” Π of a material is defined as the heat flow produced by an electric current, taken per unit current.

$$\boxed{\Pi \equiv \frac{\dot{Q}}{I}} \quad (5.38)$$

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 5.33. 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 $\dot{Q}_A = \Pi_A I_{AB}$ from material A that flows into the strip. But a different heat flux $\dot{Q}_B = \Pi_B I_{AB}$ 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:

$$\boxed{\dot{Q} = -(\Pi_B - \Pi_A) I_{AB}} \quad (5.39)$$

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 5.33, I_{AB} is negative. Also, as discussed below, the *n*-type Π_A will be negative and the *p*-type Π_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.38}. 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.

It is instructive to write down some ballparks for the Peltier coefficients. In terms of the spectra in figure 5.33, in the *n*-type semiconductor, each conduction electron has an energy per unit charge of about

$$\Pi_{n \text{ type}} \sim \frac{E^p}{-e} = \frac{E_c^p - \mu + \frac{3}{2}k_B T}{-e}$$

Here $-e$ in the denominator is the charge of the electron, while E_c^p in the numerator is the energy at the bottom of the conduction band. The thermal kinetic energy above E_c^p has been assumed to be the classical value of $\frac{3}{2}k_B T$. Further the chemical potential, or Fermi level, μ has been taken as the zero level of energy.

The reason for doing this has to do with the fact that in thermal equilibrium, all solids in contact have the same chemical potential. That is achieved by charge double layers at the contact surfaces between the solids. These charge layers affect the energy of the electrons when they move from one material to the next. In the separate materials, electrons might well have an advantage in intrinsic chemical potential in one material compared to the next. But they have to hand

in that advantage when they pass through the charge layers. These convert it into electrostatic energy. Therefore only the difference in electron energy relative to the chemical potential brings in actual thermal energy. (Accordingly, the flow of thermal energy through a material is defined as the total flow of energy through the material minus the total flow of the chemical potential that the electrons carry along, {A.38}.)

Another way of seeing the point is from the energy spectra in figure 5.33. These spectra show the chemical potential as a red tick mark. Consider the energy yield in transferring electrons between 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.

As the figure 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

$$\Pi_{p \text{ type}} \sim \frac{E^p}{e} = \frac{\mu - E_v^p + \frac{3}{2}k_B T}{e}$$

where E_v^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 $\frac{3}{2}k_B T$. 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.38}. Such issues are far beyond the scope of this book.

It is however interesting to compare the above semiconductor ballparks to one for metals:

$$\Pi_{\text{metal}} \sim -\frac{\pi^2}{2} \frac{k_B T}{E_F^p} \frac{k_B T}{e}$$

This ballpark comes from assuming the spectrum of a free-electron gas, {A.38}. The final ratio is easily understood as the classical thermal energy $k_B T$ per unit charge e . The ratio in front of it is the thermal energy $k_B T$ divided by the Fermi energy E_F^p . As discussed in section 5.10, this fraction is much less than one. Its presence can be understood from the exclusion principle: as illustrated in figure 5.15, 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.38}:

$$\frac{I^2\sigma}{T\kappa}$$

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.

5.28.2 Seebeck effect

Thermoelectric temperature sensing and power generation can be achieved by what is known as the “Seebeck effect.” It is the opposite of the Peltier effect of the previous section.

Consider the configuration shown in figure 5.34. Blocks of n -type and p -type doped semiconductor are electrically connected at their tops using a copper strip. Copper terminals are also attached to the bottoms of the semiconductor blocks, but these are electrically not in contact. No external voltage source is attached, so there is no current through the device. It is what is called an open-circuit configuration.

However, in this case 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 5.34. The device itself turns into a electric power supply. A voltage difference develops spontaneously between the bottom two terminals.

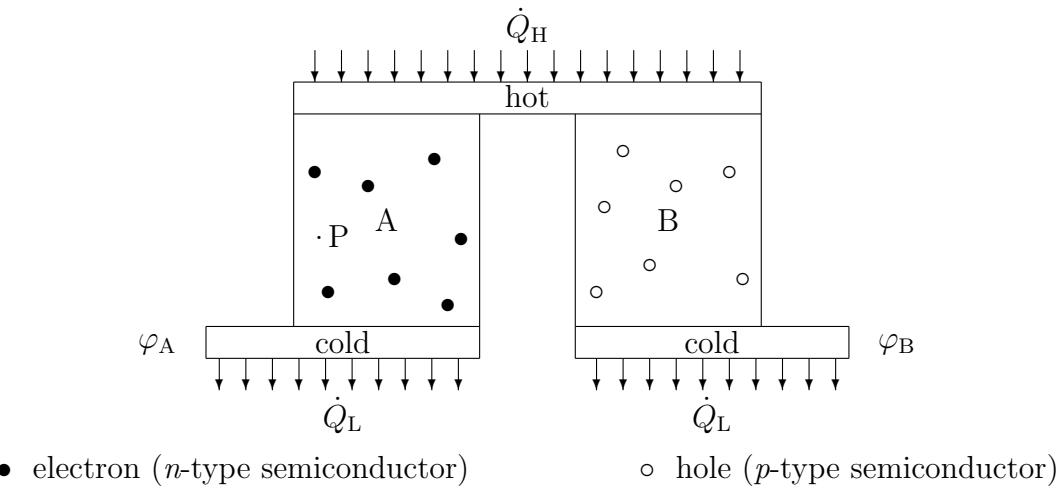


Figure 5.34: Seebeck voltage generator.

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.”

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 5.34. Imagine yourself standing at this point, shrunk down to microscopic dimensions. Due to random heat motion, 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, all else being 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.38}.

In the Seebeck effect, an incremental temperature change dT in a material causes a corresponding change in voltage $d\varphi$ given by:

$$d\varphi_\mu = -\Sigma dT$$

The subscript on φ_μ indicates that the intrinsic chemical potential of the material is included in addition to the electrostatic potential φ . In other words, φ_μ is the total chemical potential per unit electron charge. Further Σ is a 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 difference between the integrated voltage changes over the lengths of the two materials:

$$\varphi_B - \varphi_A = \int_{T_L}^{T_H} (\Sigma_B - \Sigma_A) dT \quad (5.40)$$

Since the bottom strips are both copper and at the same temperature, their intrinsic potentials are the same. Therefore the difference in φ_μ is the same as the difference in electrostatic potential φ . And it is that difference that will show up on a voltmeter connected between the bottom strips.

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_L of the bottom strips, the voltage only depends on the temperature T_H of the top strip. In terms of figure 5.34, 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.)

It is sometimes suggested, even by some that surely know better like [15, p. 14-9], that the Seebeck potential is due to jumps in potential at the contact surfaces. To explain the idea, consider figure 5.35. 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

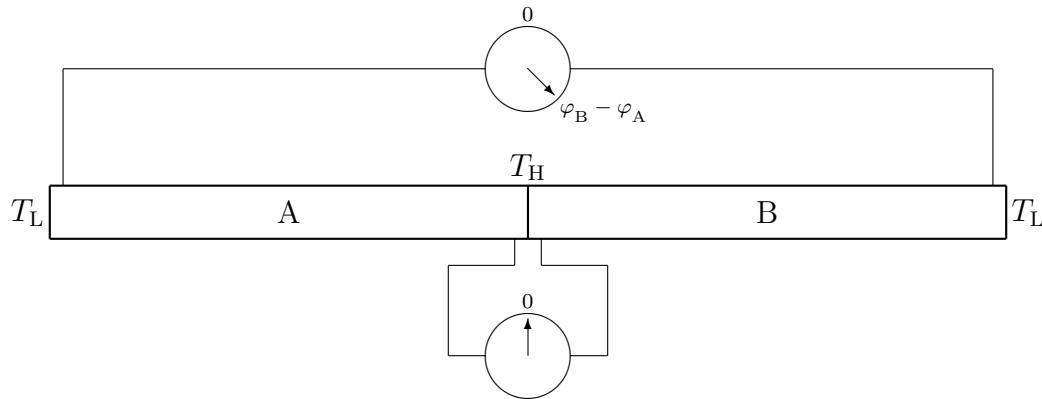


Figure 5.35: The Galvani potential jump over the contact surface does not produce a usable voltage.

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 5.35, 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 5.35. 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_L than the temperature T_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, {A.85}, that change is only one third of the potential change produced by the Seebeck coefficient given in note {A.38}. Galvani potential changes then sum to only a third of the measured potential. No, there is no partial credit.

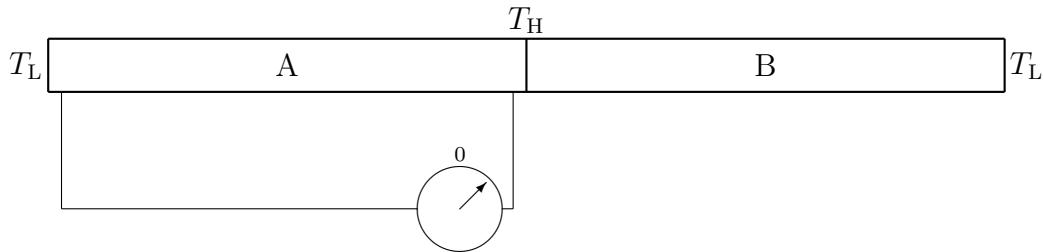


Figure 5.36: The Seebeck effect is not directly measurable.

It should also be pointed out that the Seebeck effect of a material is not directly measurable. Figure 5.36 illustrates an attempt to directly measure the Seebeck effect of material A. Unfortunately, the only thing that changes compared to figure 5.35 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 (or rather Onsager) showed that you can find the Seebeck coefficient Σ from the Peltier coefficient Π simply by dividing by the 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 eliminate 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.38}. 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.

- 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.
-

5.28.3 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 material with a current through it can be written as:

$$\dot{e} = \frac{d}{dx} \left(\kappa \frac{dT}{dx} \right) + \frac{j^2}{\sigma} - \mathcal{K} j \frac{dT}{dx} \quad (5.41)$$

Here 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 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 note {A.38}.

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.38}:

$$\frac{d\Sigma}{d \ln T} = \mathcal{K} \quad \Pi = \Sigma T \quad (5.42)$$

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

- o— 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.
 - o— Unlike the Peltier and Seebeck coefficients, the Kelvin (Thomson) coefficient can be measured without involving a second material.
 - o— The Kelvin (Thomson) relations allow you to compute the Peltier and Seebeck coefficients from the Kelvin one.
-

Chapter 6

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 atomic and nuclear decays.

The first section introduces the Schrödinger equation, which is as important for quantum mechanics as Newton's second law is for classical mechanics. For one, the Schrödinger equation explains why quantum mechanics becomes classical mechanics for macroscopic objects. It also implies that energy eigenstates are stationary. Therefore, nontrivial evolution of a system requires uncertainty in energy.

The second section explains where 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.

The third section examines the evolution of two-state systems when they are perturbed by an unsteady effect. The emphasis is on the absorption and emission of radiation by atoms. The operating principle of lasers is explained. The spontaneous decay of excited atoms is explained in terms of a background radiation level that is always present.

The final sections discuss examples of the nontrivial evolution of simple quantum systems in time. 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 nontrivial localization in both space and time can be identified as “packets” of waves.

The motion of such wave packets is then examined. If the forces change slowly on quantum scales, wave packets move just like classical particles do. Under such conditions, a simple theory called the WKB approximation applies. These ideas can be generalized to the motion of conduction electrons in crystals.

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. Normally, a tunneling wave packet will be partially transmitted and partially reflected by a 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.

6.1 The Schrödinger Equation

Section Abstract

The Schrödinger equation describes the time-evolution of the wave function. It takes a similar place in quantum mechanics as Newton's second law does in classical mechanics. Assuming that the energy eigenfunctions are known, the Schrödinger equation is readily solved by taking the coefficients of the eigenfunctions to be appropriate functions of time.

From that a number of immediate conclusions can be drawn. For example, it follows that energy eigenstates are stationary. To get a nontrivial time evolution requires uncertainty in energy.

The effect of uncertainty in energy is examined for a two state system like the electron in the hydrogen molecular ion. There is uncertainty in energy if the electron is initially around a single nucleus rather than being symmetrically or antisymmetrically shared. It turns out that if the electron is around a single nucleus, it will after some time jump over to the other nucleus. Phrased differently, the nuclei "play catch" with the electron. In those terms, the bound molecular ion state can be thought of as a linear combination of two "play catch" solutions. This is in fact how relativistic quantum mechanics, (not covered in this book), thinks about forces. For example, quantum electrodynamics says that the electromagnetic force between charged particles is not due to a Coulomb potential, but mostly due to the fact that the particles play catch with a photon.

The two-state system is also used to derive a version of an “energy-time” uncertainty relation. Such a relation is similar to Heisenberg’s momentum-position uncertainty relation, but now involving energy and time.

Next, the Schrödinger equation is used to derive an evolution equation for expectation values. That equation is then used to derive the equations of classical physics, including Newton’s second law. Quantum mechanics shows that Newton’s second law only applies to expectation values.

Two additional topics are also briefly discussed. The adiabatic theorem is concerned with approximate solutions that assume that the parameters of the quantum system vary only slowly. The Heisenberg picture is an alternative formulation of the Schrödinger equation that has some advantages in the relativistic case.

In Newtonian mechanics, Newton’s second law states that the linear momentum changes in time proportional to the applied force; $d\vec{m}/dt = \vec{m} = \vec{F}$. 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.

6.1.1 Intro to the equation

The Schrödinger equation says that the time derivative of the wave function is obtained by applying the Hamiltonian on it. More precisely:

$$\boxed{i\hbar \frac{\partial \Psi}{\partial t} = H\Psi} \quad (6.1)$$

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 the energy eigenfunctions $\psi_{\vec{n}}$ of the Hamiltonian,

$$\Psi = c_{\vec{n}_1}(t)\psi_{\vec{n}_1} + c_{\vec{n}_2}(t)\psi_{\vec{n}_2} + \dots = \sum_{\vec{n}} c_{\vec{n}}(t)\psi_{\vec{n}} \quad (6.2)$$

Then the coefficients $c_{\vec{n}}$ must evolve in time as complex exponentials:

$$\boxed{c_{\vec{n}}(t) = c_{\vec{n}}(0)e^{-iE_{\vec{n}}t/\hbar}} \quad (6.3)$$

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 $e^{-iE_{\vec{n}}t/\hbar}$. 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.

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

- □ The Schrödinger equation describes the time evolution of the wave function.
- □ The time derivative is proportional to the Hamiltonian.
- □ The coefficients of the energy eigenfunctions must be proportional to $e^{-iE_{\vec{n}}t/\hbar}$.

6.1.2 Some examples

As a simple example of the application of the Schrödinger equation, consider the particle stuck in a pipe discussed much earlier in chapter 2.5. It has three quantum numbers, one for each Cartesian coordinate. In the ground state ψ_{111} all three quantum numbers are one. Now assume that the wave function Ψ is exactly the ground state ψ_{111} at time zero. Then according to the Schrödinger equation, at later times, the wave function is

$$\Psi = e^{-iE_{111}t/\hbar}\psi_{111}$$

for any time t .

So how does it look? Well the probability of finding the particle is given by the square magnitude of Ψ . But the square magnitude of the exponential is always one, whatever the time t . So the probability of finding the particle never changes. At all times it looks like:



Figure 6.1: The ground state wave function looks the same at all times.

The same happens if the initial wave function is exactly the first excited state ψ_{211} . Then for later times, the wave function is

$$\Psi = e^{-iE_{211}t/\hbar}\psi_{211}$$

Once again, the exponential makes no difference for the probability of finding the particle. At all times, this solution looks like:



Figure 6.2: The first excited state at all times.

You see why wave functions that consist of a single energy eigenfunction are called “stationary states.”

Things only become somewhat interesting if the initial wave function is a combination of energy eigenfunctions with different energy. For example, assume that the initial wave function is the following combination of the ground state and the first excited state:

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

Then at later times the wave function is, shoving in the exponentials,

$$\Psi = \sqrt{\frac{4}{5}}e^{-iE_{111}t/\hbar}\psi_{111} + \sqrt{\frac{1}{5}}e^{-iE_{211}t/\hbar}\psi_{211}$$

The square magnitude of that multiplies out to

$$|\Psi|^2 = \Psi^*\Psi = \frac{4}{5}|\psi_{111}|^2 + \frac{4}{5}\cos((E_{111} - E_{211})t/\hbar)\psi_{111}\psi_{211} + \frac{1}{5}|\psi_{211}|^2$$

Note that now the exponentials no longer completely drop out of the square magnitude. That means that the probability of finding the particle is now different at different times. In fact, at four representative times it looks like:

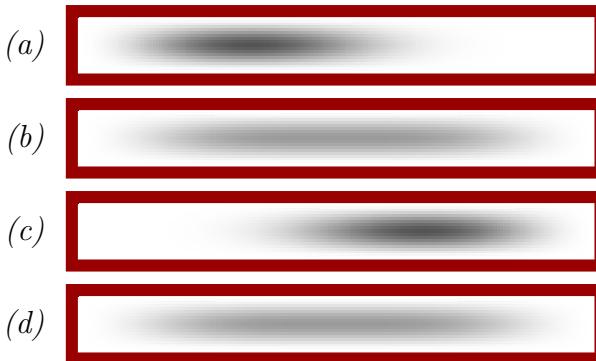


Figure 6.3: A combination of ψ_{111} and ψ_{211} seen at some typical times.

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 would bounce back and forth between the ends of the pipe.

Note that this nontrivial evolution is associated with uncertainty in energy. In particular, it is seen from the wave function

$$\Psi = \sqrt{\frac{4}{5}}e^{-iE_{111}t/\hbar}\psi_{111} + \sqrt{\frac{1}{5}}e^{-iE_{211}t/\hbar}\psi_{211}$$

that at all times there is a 4/5 chance, (the square magnitude of the coefficient of ψ_{111}), for the energy to be E_{111} , and a 1/5 chance, (the square magnitude of the coefficient of ψ_{211}), for the energy to be E_{211} . The exponentials drop out of the square magnitudes.

As another example, the energy eigenfunctions $\psi_{nlm\uparrow}$ and $\psi_{nlm\downarrow}$ of the hydrogen atom are characterized by the set of quantum numbers n , l , m , and m_s , where $m_s = \frac{1}{2}$ indicates spin up and $m_s = -\frac{1}{2}$ spin down. Any hydrogen wave function can always be written as a linear combination of all these individual states:

$$\Psi = \sum_{n=1}^{\infty} \sum_{l=0}^{n-1} \sum_{m=-l}^l c_{nlm+} \psi_{nlm} + c_{nlm-} \psi_{nlm}$$

To make this is a valid solution of the Schrödinger equation, simply ensure that the coefficients have the correct exponential dependence on time. Written out in full, the wave function is:

$$\Psi(r, \theta, \phi, t) = \sum_{n=1}^{\infty} \sum_{l=0}^{n-1} \sum_{m=-l}^l c_{nlm+}(0)e^{-iE_nt/\hbar}\psi_{nlm}(r, \theta, \phi)\uparrow + c_{nlm-}(0)e^{-iE_nt/\hbar}\psi_{nlm}(r, \theta, \phi)\downarrow$$

(This solution ignores any external disturbances such as the ambient electromagnetic field and small errors due to spin and relativity.) The values of the initial coefficients $c_{nlm+}(0)$ and $c_{nlm-}(0)$ need to be found from whatever is given for the initial wave function.

Key Points

- o Some example solutions of the Schrödinger equation were given.

6.1.2 Review Questions

- 1 The energy of a photon is $\hbar\omega$ where ω is the classical frequency of the electromagnetic field given by the photon. So what is $e^{-iE_{\vec{n}}t/\hbar}$ for a photon? Are you surprised by the result?

- 2** For the one-dimensional harmonic oscillator, the energy eigenvalues are

$$E_n = \frac{2n+1}{2}\omega$$

Evaluate the coefficients $c_n(0)e^{-iE_n t/\hbar}$.

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.

- 3** Write the full wave function for a one-dimensional harmonic oscillator. Formulae are in chapter 2.6.2.
-

6.1.3 Energy conservation [Descriptive]

Assuming that there are no external influences, the Schrödinger equation implies that the energy of a system is conserved. To see why, remember that the square magnitudes $|c_{\vec{n}}|^2$ of the coefficients of the energy eigenfunctions give the probability for the corresponding energy. While according to the Schrödinger equation the coefficients vary with time, their square magnitudes do not:

$$|c_{\vec{n}}(t)|^2 \equiv c_{\vec{n}}^*(t)c_{\vec{n}}(t) = c_{\vec{n}}^*(0)e^{iE_{\vec{n}}t/\hbar}c_{\vec{n}}(0)e^{-iE_{\vec{n}}t/\hbar} = |c_{\vec{n}}(0)|^2$$

So according to the orthodox interpretation, the probability of measuring a given energy level does not vary with time either. For example, the wave function for a hydrogen atom at the excited energy level E_2 might be of the form:

$$\Psi = e^{-iE_2 t/\hbar} \psi_{210\uparrow}$$

(This corresponds to an assumed initial condition in which all $c_{nlm\pm}$ are zero except $c_{210+} = 1$.) 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 is conserved.

This also illustrates that left to itself, an excited atom will maintain its energy indefinitely. It will *not* emit a photon and drop back to the ground state energy E_1 . Excited atoms emit radiation because they are perturbed by an electromagnetic field. This is true even in vacuum at absolute zero. For one thing, even in vacuum, the electromagnetic field has a nonzero ground state energy. However, the radiation that the atom interacts with is its own, through a weird twilight effect, chapter 12.2.4. In any case, eventually, at a time that is observed to be random for reasons discussed in chapter 13.5, the perturbation

causes the excited atom to drop back to the lower energy state and emit the photon.

Returning to the unperturbed atom, it should also be noted that even if the energy is uncertain, still the probabilities of measuring the various energy levels do not change with time. As an arbitrary example, the following wave function describes a case of an undisturbed hydrogen atom where the energy has a 50/50 chance of being measured as E_1 , (-13.6 eV), or as E_2 , (-3.4 eV):

$$\Psi = \frac{1}{\sqrt{2}}e^{-iE_1 t/\hbar}\psi_{100}\downarrow + \frac{1}{\sqrt{2}}e^{-iE_2 t/\hbar}\psi_{210}\uparrow$$

The 50/50 probability applies regardless how long the wait is before the measurement is done.

How about the other conservation laws, such as conservation of linear or angular momentum for a system that is left alone? Surprisingly, it turns out that these conservation laws arise from the symmetries of physics. That is discussed in section 6.2.

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.

6.1.4 Stationary states [Descriptive]

The previous subsection examined the time variation of energy, but the Schrödinger equation also determines how the other physical properties, such as positions and momenta, of a given system vary with time.

The simplest case is that in which the energy is certain, in other words, states in which the wave function is a *single* energy eigenfunction:

$$\Psi = c_{\vec{n}}(0)e^{-iE_{\vec{n}}t/\hbar}\psi_{\vec{n}}$$

It turns out, {A.39}, that none of the physical properties of such a state changes with time. The physical properties may be uncertain, but the probabilities for their possible values will remain the same. For that reason, states of definite energy are called “stationary states.”

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 will be no change in the probability of finding the particle at any given location regardless of the time you look, so how could it possibly resemble a classical particle that is at different positions at different times?

Similarly, while classically the linear momentum of a particle that experiences forces will change with time, in energy eigenstates the chances of measuring a given momentum do not change with time.

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

- If an isolated system has definite energy, the statistics of every physical variable are independent of time.
- Therefore, to get nontrivial time variation of a system requires uncertainty in energy.

6.1.5 Particle exchange [Descriptive]

The simplest case of a physical system that can have a nontrivial dependence on time is a system described by two different states. Some examples of such systems were given in chapter 4.3. One was the hydrogen molecular ion, consisting of two protons and one electron. In that case, there was a state ψ_1 in which the electron was in the ground state around one proton, and a state ψ_2 in which it was around the other proton. Another example was the ammonia molecule, where the nitrogen atom was at one side of its ring of hydrogens in state ψ_1 , and at the other side in state ψ_2 . This section examines the time variation of such systems.

It will be assumed that the states ψ_1 and ψ_2 are physically equivalent, like the mentioned examples. In that case, according to chapter 4.3 the ground state of lowest energy, call it E_L , is an equal combination of the two states ψ_1 and ψ_2 . The state of highest energy E_H is also an equal combination, but with the opposite sign. The solution of the Schrödinger equation is in terms of these two combinations of states, {A.40}:

$$\Psi = c_L e^{-iE_L t/\hbar} \frac{\psi_1 + \psi_2}{\sqrt{2}} + c_H e^{-iE_H t/\hbar} \frac{\psi_1 - \psi_2}{\sqrt{2}}$$

Consider now the case of the hydrogen molecular ion, and assume that the electron is around the first proton, so in state ψ_1 , at time $t = 0$. The wave function must then be:

$$\Psi = c_L e^{-iE_L t/\hbar} \left[\frac{\psi_1 + \psi_2}{\sqrt{2}} + e^{-i(E_H - E_L)t/\hbar} \frac{\psi_1 - \psi_2}{\sqrt{2}} \right]$$

At time zero, this produces indeed state ψ_1 , but when the exponential in the last term becomes -1, the system converts into state ψ_2 . The electron has jumped over to the other proton.

The time that this takes can be found from setting the argument of the exponential equal to $-i\pi$, since $e^{-i\pi} = -1$, (1.5). The time interval for the electron to jump over is then found to be:

$$\Delta t = \frac{\pi\hbar}{E_H - E_L}$$

After another time interval Δt the electron will be back in state ψ_1 around the first proton, and so on. In a sense, the protons play catch with the electron.

The time interval for the two protons to exchange the electron is inversely proportional to the energy difference $E_H - E_L$. In chapter 4.3 this energy difference appeared in another context: it is twice the molecular binding energy produced by the “twilight terms” when the electron is shared. It is interesting now to see that this binding energy also determines the time it takes for the electron to be exchanged if it is not shared. The more readily the protons exchange the non-shared electron, the more the binding energy of the shared state will be. It may be one reason that the twilight terms are commonly referred to as “exchange terms.”

The mathematics for the time evolution of the nitrogen atom in ammonia is similar. If measurements locate the nitrogen atom at one side of the hydrogen ring, then after a certain time, it will pop over to the other side.

The “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.

That has some similarity with the force that keeps the protons together in the hydrogen molecular ion. In particular, the bound molecular state is an equal combination of the electron-catch solution given above in which the electron is initially around the first proton, and the similar electron-catch solution in which it is initially around the second proton. Note however that the electron-catch solution was based on the Coulomb potential. This potential implies instantaneous interaction at a distance, and relativity does not allow that. In a relativistic description, the interaction between charged particles and photons is local.

The other three fundamental forces of nature arise similarly as electromagnetism, but with different exchanged particles than photons. In the “color force” between “quarks,” “gluons” are exchanged. The “weak force” is mediated by

massive particles called “intermediate vector bosons” and gravity supposedly by particles called “gravitons.”

Key Points

- The fundamental forces are due to the exchange of particles.
- The particles are photons for electromagnetism, gluons for the color force, intermediate vector bosons for the weak force, and presumably gravitons for gravity.

6.1.6 Energy-time uncertainty relation [Descriptive]

The evolution of a two state system as given in the previous subsection is a simple model for the decay of a system, like maybe the decay of an excited state of an hydrogen atom.

The initial excited atomic state ψ_1 seems to be an energy eigenstate. That would make it a stationary state, 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 background electromagnetic radiation. The state ψ_1 has therefore some uncertainty in energy. The state ψ_2 consists of the ground state atom 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. The true energy eigenstates are ψ_L , a symmetric combination of ψ_1 and ψ_2 , and ψ_H , an antisymmetric combination of the two.

In that case, according to the previous section the state ψ_1 , the excited atom, will after some time turn into the state ψ_2 , the ground state atom plus emitted photon. The time Δt that that takes is of the order $\hbar/(E_H - E_L)$. Also, since the two states are combinations of ψ_L and ψ_H , they have an uncertainty in energy ΔE equal to half the energy difference $E_H - E_L$. Therefore:

$$\boxed{\Delta E \Delta t \sim \frac{1}{2} \hbar} \quad (6.4)$$

A relation of this type is called an “energy-time uncertainty relation.” It is similar to the Heisenberg relation $\Delta p \Delta x \geq \frac{1}{2} \hbar$, but for energy and time instead of momentum and spatial position. Such relations should be expected to exist because of relativity. Relativity treats energy much like a fourth component of momentum and time as a fourth position coordinate, {A.4}.

Returning to the decay of the excited atom, the two-state model would eventually have the decayed state ψ_2 reverting back to the excited state ψ_1 . That means that the emitted photon will eventually again excite the atom and disappear itself. To prevent this from happening, it must be assumed the

state of the system is “measured.” 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 given 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. (See chapter 13.5 for a further discussion of this process.)

Key Points

- An energy-time uncertainty relation is of the form $\Delta E \Delta t \sim \frac{1}{2}\hbar$.
- It resembles the Heisenberg momentum-position uncertainty relation.
- The two-state system provides a model for the decay of physical states.
- Interaction with the environment is needed to make the decay permanent.

6.1.7 Time variation of expectation values [Descriptive]

The time evolution of more complex systems can be described in terms of the energy eigenfunctions of the system, just like for the two state systems of the previous subsections. However, finding all the eigenfunctions may not be easy.

Fortunately, it is possible to find the evolution of the expectation value of physical quantities without solving the energy eigenvalue problem. The expectation value, defined in chapter 3.3, gives the average of the possible values of the physical quantity.

The Schrödinger equation requires that the expectation value $\langle a \rangle$ of any physical quantity a with associated operator A evolves in time as:

$$\frac{d\langle a \rangle}{dt} = \frac{i}{\hbar} \langle [H, A] \rangle + \left\langle \frac{\partial A}{\partial t} \right\rangle \quad (6.5)$$

The derivation is in note {A.41}. The commutator $[H, A]$ of A with the Hamiltonian was defined in chapter 3.4 as $HA - AH$. The final term in (6.5) 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. Its main application is to relate quantum mechanics to Newtonian mechanics, as in the next section. (Some minor applications that will be left to the notes for the interested are the “virial theorem” {A.42} relating kinetic and potential energy and the Mandelshtam-Tamm version of the “energy-time uncertainty principle” $\Delta E \Delta t \geq \frac{1}{2}\hbar$ {A.43}.)

Note that if A commutes with the Hamiltonian, i.e. $[H, A] = 0$, then the expectation value of the corresponding quantity a will not vary with time. Such a quantity has eigenfunctions that are also energy eigenfunctions, so it has the same time-conserved statistics as energy. Equation (6.5) demonstrates this for the expectation value, but the standard deviation, etcetera, would not change with time either.

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.

6.1.8 Newtonian motion [Descriptive]

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 $\sigma_x \sigma_{p_x} \geq \frac{1}{2}\hbar$, 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 (6.5) 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:

$$H = \frac{\hat{p}_x^2}{2m} + V(x)$$

where \hat{p}_x is the linear momentum operator and m the mass of the particle.

Now according to evolution equation (6.5), the expectation position $\langle x \rangle$ changes at a rate:

$$\frac{d\langle x \rangle}{dt} = \left\langle \frac{i}{\hbar} [H, \hat{x}] \right\rangle = \left\langle \frac{i}{\hbar} \left[\frac{\hat{p}_x^2}{2m} + V(x), \hat{x} \right] \right\rangle \quad (6.6)$$

Recalling the properties of the commutator from chapter 3.4, $[V(x), \hat{x}] = 0$, since multiplication commutes. Further, according to the rules for manipulation of products and the canonical commutator

$$[\hat{p}_x^2, \hat{x}] = \hat{p}_x[\hat{p}_x, \hat{x}] + [\hat{p}_x, \hat{x}]\hat{p}_x = -\hat{p}_x[\hat{x}, \hat{p}_x] - [\hat{x}, \hat{p}_x]\hat{p}_x = -2i\hbar\hat{p}_x$$

So the rate of change of expectation position becomes:

$$\frac{d\langle x \rangle}{dt} = \left\langle \frac{p_x}{m} \right\rangle \quad (6.7)$$

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 $[H, \hat{p}_x]$ is needed. Now \hat{p}_x commutes, of course, with itself, but just like it does not commute with \hat{x} , it does not commute with the potential energy $V(x)$; the generalized canonical commutator (3.44) says that $[V, \hat{p}_x]$ equals $-\hbar\partial V/i\partial x$.

As a result, the rate of change of the expectation value of linear momentum becomes:

$$\frac{d\langle p_x \rangle}{dt} = \left\langle -\frac{\partial V}{\partial x} \right\rangle \quad (6.8)$$

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 classical equations is known as "Ehrenfest's theorem."

(For a quantum 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.)

Key Points

- The equations of classical physics apply to expectation values.

6.1.9 The adiabatic approximation [Descriptive]

The adiabatic approximation describes the evolution of systems for which the Hamiltonian changes nontrivially in time, but slowly. (Note that this use of the

word “adiabatic” is not to be confused with adiabatic in thermodynamics, which normally describes processes that occur sufficiently *quickly* that heat transfer with the surroundings can be ignored. The term “quasi-steady” for slowly evolving quantum systems instead of adiabatic would be understandable, so physicists could not use that one.)

As a simple example, assume that you have a particle in the ground state in a box, and 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, the particle will indeed remain in the ground state, even when that state slowly changes into a completely different one.

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 remains at all times non-degenerate, so that the energy state is unambiguous. More sophisticated versions of the analysis exist to deal with degeneracy and continuous spectra.

Note A.44 gives a derivation of the theorem and some additional implications. The most important practical application of the adiabatic theorem is without doubt the Born-Oppenheimer approximation, which is discussed separately in chapter 7.2.

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.”
-

6.1.10 Heisenberg picture [Descriptive]

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 wave function. According to the Schrödinger equation, it can be written as

$$\Psi(\dots; t) = e^{-iHt/\hbar} \Psi(\dots; 0) \quad (6.9)$$

where the exponential of an operator is defined through its Taylor series:

$$e^{-iHt/\hbar} = 1 - i\frac{t}{\hbar}H - \frac{t^2}{2!\hbar^2}H^2 + \dots \quad (6.10)$$

This exponential form applies assuming that the Hamiltonian is independent of time. If it is not, the transformation from the initial wave function $\Psi(\dots; 0)$ to a later time one $\Psi(\dots; t)$ still remains a “unitary” one; one that keeps the wave function normalized.

Now consider an arbitrary Schrödinger operator \hat{A} . The physical effects of the operator can be characterized by inner products, as in

$$\langle \Psi_1(\dots; t) | \hat{A} \Psi_2(\dots; t) \rangle \quad (6.11)$$

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, that operator becomes

$$\tilde{A} \equiv e^{iHt/\hbar} \hat{A} e^{-iHt/\hbar} \quad (6.12)$$

where the argument of the first exponential changed sign due to being taken to the other side of the inner product.

The operator \tilde{A} depends on time. To see how it evolves, differentiate the product with respect to time:

$$\frac{d\tilde{A}}{dt} = \frac{i}{\hbar} H e^{iHt/\hbar} \hat{A} e^{-iHt/\hbar} + e^{iHt/\hbar} \frac{\partial \hat{A}}{\partial t} e^{-iHt/\hbar} - e^{iHt/\hbar} \hat{A} e^{-iHt/\hbar} \frac{i}{\hbar} H$$

The first and third terms can be recognized as the commutator of H and \tilde{A} , while the middle term is the Heisenberg version of the time derivative of \hat{A} , in case \hat{A} does depend on time. So the evolution equation for the Heisenberg operator becomes

$$\frac{d\tilde{A}}{dt} = \frac{i}{\hbar} [H, \tilde{A}] + \frac{\widetilde{\partial \hat{A}}}{\partial t} \quad [H, \tilde{A}] = e^{iHt/\hbar} [H, \hat{A}] e^{-iHt/\hbar} \quad (6.13)$$

(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 x position and linear momentum operators of a particle. These do not depend on time, and using the commutators as figured out in the previous subsection, the Heisenberg operators evolve as:

$$\frac{d\tilde{x}}{dt} = \frac{1}{m}\tilde{p}_x \quad \frac{d\tilde{p}_x}{dt} = -\frac{\partial \widetilde{V}}{\partial x}$$

Those have the same form as the equations for the classical position and momentum. It is the Ehrenfest theorem on steroids.

In fact, the equivalent of the general equation (6.13) is also found in classical physics: it is derived in advanced mechanics in this form, 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 is some function of the position and linear momentum of the particle has a time derivative given by

$$\frac{da}{dt} = \frac{\partial a}{\partial x} \frac{dx}{dt} - \frac{\partial a}{\partial p_x} \frac{dp_x}{dt}$$

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”

$$\frac{dx}{dt} = \frac{\partial H}{\partial p_x} \quad \frac{dp_x}{dt} = -\frac{\partial H}{\partial x}$$

Substituting this into the time derivative of a gives

$$\frac{da}{dt} = \frac{\partial a}{\partial x} \frac{\partial H}{\partial p_x} - \frac{\partial a}{\partial p_x} \frac{\partial H}{\partial x}$$

The negative of the right hand side is by definition the Poisson bracket (H, a) . Note that it, like the commutator, is antisymmetric under exchange of H and a .

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 note {A.3}, 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

- o In the Heisenberg picture, operators evolve in time just like their physical variables do in classical physics.

6.2 Conservation Laws and Symmetries

This section explains where conservation laws such as conservation of linear and angular momentum come from. They are shown to be consequences of symmetry properties of nature. This is of particular interest for applications like nuclear physics in which much is still not understood; symmetry can be used regardless of the precise details of the nuclear forces.

Pretend for now that you have never heard of angular momentum, nor that it would be conserved, nor what its operator would be. However, there is at least one operation you do know without being told about: rotating a system over an angle.

Consider the effect of this operation on a *complete system* in otherwise empty space. Since empty space by itself has no preferred directions, it does not make a difference under what angle you initially position the system. Identical systems placed in different initial angular orientations will evolve the same, just seen from a different angle.

This “invariance” with respect to angular orientation has consequences when phrased in terms of operators and the Schrödinger equation. In particular, let a system of particles 1, 2, . . . , be described in spherical coordinates by a wave function:

$$\Psi(r_1, \theta_1, \phi_1, S_{z1}, r_2, \theta_2, \phi_2, S_{z2}, \dots; t)$$

and let R_φ be the operator that rotates this entire system over a given angle φ around the z -axis:

$$\begin{aligned} R_\varphi \Psi(r_1, \theta_1, \phi_1, S_{z1}, r_2, \theta_2, \phi_2, S_{z2}, \dots; t) \\ = \Psi(r_1, \theta_1, \phi_1 + \varphi, S_{z1}, r_2, \theta_2, \phi_2 + \varphi, S_{z2}, \dots; t) \end{aligned}$$

(For the formula as shown, the rotation of the system φ is in the direction of decreasing ϕ . Or if you want, it corresponds to an observer or axis system rotated in the direction of increasing ϕ ; in empty space, who is going to see the difference? The first viewpoint is called the active one, the second the passive one.)

Now the key point is that if space has no preferred direction, the operator R_φ must commute with the Hamiltonian:

$$HR_\varphi = R_\varphi H$$

After all, it should not make any difference at what angle compared to empty space the Hamiltonian is applied: if you first rotate the system and then apply the Hamiltonian, or first apply the Hamiltonian and then rotate the system, the result should be the same. For that reason, an operator such as R_φ , which commutes with the Hamiltonian of the considered system, is called a physical *symmetry* of the system.

The fact that R_φ and H commute has a mathematical consequence, {A.19}: it means that R_φ must have a complete set of eigenfunctions that are also energy eigenfunctions. And for energy eigenfunctions the Schrödinger equation gives the evolution. In particular, the Schrödinger equation says that if the system starts out in a single energy eigenfunction, it stays in that eigenfunction. That means that there is only one value of the energy that has a nonzero probability; the eigenvalue of that single state. Therefore the energy does not change with time, it is conserved.

But this is also an eigenfunction of R_φ ! So there is at all times also only a single eigenvalue of R_φ that has a nonzero probability. Therefore, whatever the physical quantity given by the eigenvalues of R_φ might turn out to be, it is conserved just like energy is. If a system starts out with a definite value for the quantity, it conserves that value. (See note {A.45} to verify this for a more general initial condition. Also, even if there is uncertainty in the value, still the probabilities of the individual eigenvalues do not change with time. That follows because the Schrödinger equation is linear, so you can add solutions.)

The next step is to figure out exactly what this conserved quantity might correspond to physically. First of all, the magnitude of any eigenvalue of R_φ must be one: if it was not, the square integral of Ψ could increase by that factor during the rotation, but of course it must stay the same. Since the magnitude is one, the eigenvalue can be written in the form e^{ia} where a is some ordinary real number. The eigenvalue has been narrowed down a bit already.

But the eigenvalue must more specifically be of the form $e^{im\varphi}$, where m is some real number independent of the amount of rotation. The reasons are that there must be no change in Ψ when the angle of rotation is zero, and a single rotation over φ must be the same as two rotations over an angle $\frac{1}{2}\varphi$. Those requirements imply that the eigenvalue is of the form $e^{im\varphi}$.

So $e^{im\varphi}$ is a conserved quantity if the system starts out as the corresponding eigenfunction of R_φ . You can simplify that statement to say that m by itself is conserved; if m varied in time, $e^{im\varphi}$ would too. Also, you might scale m by some constant, call it \hbar , so that you can conform to the dimensional units others, such as classical physicists, might turn out to be using for this conserved quantity.

You can just give a fancy name to this conserved quantity $m\hbar$. You can call it “net angular momentum around the z -axis.” That sounds less nerdy at parties than “scaled logarithm of the conserved eigenvalue of R_φ .” You might think of even better names, but whatever the name, it is conserved.

Next, you would probably like to define a nicer operator for this “angular momentum” than the rotation operators R_φ . The problem is that there are infinitely many of them, one for every angle φ , and they are all related, a rotation over an angle 2φ being the same as two rotations over an angle φ . If you define a rotation operator over a very small angle, call it angle ε , then you can approximate all the other operators R_φ by just applying R_ε sufficiently many times. To make these approximations exact, you need to make ε infinitesimally small. But when ε becomes zero, R_ε would become just 1. You have lost the 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 :

$$\lim_{\varepsilon \rightarrow 0} \frac{R_\varepsilon - 1}{\varepsilon}$$

is the operator you want.

Now consider what this operator really means for a single particle with no spin:

$$\lim_{\varepsilon \rightarrow 0} \frac{R_\varepsilon - 1}{\varepsilon} \Psi(r, \theta, \phi) = \lim_{\varepsilon \rightarrow 0} \frac{\Psi(r, \theta, \phi + \varepsilon) - \Psi(r, \theta, \phi)}{\varepsilon}$$

By definition, the final term is the partial derivative of Ψ with respect to ϕ . So the operator you just defined is just the operator $\partial/\partial\phi$!

You can go one better still, because the eigenvalues of the operator just defined are

$$\lim_{\varepsilon \rightarrow 0} \frac{e^{im\varepsilon} - 1}{\varepsilon} = im$$

If you add a factor \hbar/i to the operator, the eigenvalues of the operator are going to be $m\hbar$, the quantity you defined to be the angular momentum. So you are led to define the angular momentum operator as:

$$\hat{L}_z \equiv \frac{\hbar}{i} \frac{\partial}{\partial\phi}$$

This agrees perfectly with what you got much earlier in chapter 3.1.2 from guessing that the relationship between angular and linear momentum is the

same in quantum mechanics as in classical mechanics. Now you derived it from the fundamental rotational symmetry property of nature, instead of from guessing.

How about the angular momentum of a system of multiple, but still spinless, particles? It is easy to see that the operator

$$\frac{\hbar}{i} \lim_{\varepsilon \rightarrow 0} \frac{R_\varepsilon - 1}{\varepsilon}$$

now acts as a total derivative, equivalent to the sum of the partial derivatives of the individual particles. So the orbital angular momenta of the individual particles just add, as they do in classical physics.

How about spin? Well, take a hint from nature. If a particle in a given spin state has an inherent angular momentum in the z -direction $m\hbar$, then apparently the wave function of that particle changes by $e^{im\varphi}$ when you rotate the particle over an angle φ . A surprising consequence is that if the system is rotated over an angle 2π , half integer spin states do not return to the same value; they change sign. Since only the magnitude of the wave function is physically observable, this change of sign does not affect the physical symmetry.

With angular momentum defined, the rotation operator R_φ can be explicitly identified if you are curious. It is

$$R_\varphi = \exp(\varphi i \hat{L}_z / \hbar)$$

where the exponential of an operator is found by writing the exponential as a Taylor series. R_φ is called the “generator of rotations around the z -axis.” To check that it does indeed take the form above, expand the exponential in a Taylor series and multiply by a state with angular momentum $L_z = m\hbar$. The effect is seen to be to multiply the state by the Taylor series of $e^{im\varphi}$ as it should. So R_φ gets all eigenstates and eigenvalues correct, and must therefore be right since the eigenstates are complete. As an additional check, R_φ can also be verified explicitly for purely orbital momentum states; for example, it turns the wave function $\Psi(r, \theta, \phi)$ for a single particle into

$$\exp\left(\varphi \frac{i}{\hbar} \hat{L}_z\right) \Psi(r, \theta, \phi) = \exp\left(\varphi \frac{\partial}{\partial \phi}\right) \Psi(r, \theta, \phi)$$

and expanding the exponential in a Taylor series produces the Taylor series for $\Psi(r, \theta, \phi + \varphi)$, the correct expression for the wave function in the rotated coordinate system.

There are other symmetries of nature, and they give rise to other conservation laws and their operators. For example, nature is symmetric with respect to translations: it does not make a difference where in empty space you place your system. This symmetry gives rise to linear momentum conservation in the

same way that rotational symmetry led to angular momentum conservation. Symmetry with respect to time delay gives rise to energy conservation.

Initially, it was also believed that nature was symmetric with respect to mirroring it (looking at the physics in a mirror). That gave rise to a law of conservation of “parity”. Parity is called “even” if the wave function remains the same when you replace \vec{r} by $-\vec{r}$ and “odd” if it changes sign. (Replacing \vec{r} by $-\vec{r}$ is called inversion. It is a mathematically neat way of mirroring, even though really it also involves a 180° rotation around the axis normal to the mirror. And inversion does not work in two dimensions, because the rotation has to be a three-dimensional one.)

The parity of a complete system was believed to be conserved for a long time. However, it turned out that the weak nuclear force does not stay the same under mirroring, so that parity is not conserved when weak interactions play a role. Nowadays, most physicists believe that in order to get an equivalent system under those conditions, in addition to the mirroring, you also need to replace the particles by their antiparticles, having opposite charge, and reverse the direction of time.

Invariances of systems such as the ones above are called “group properties.” There is an entire branch of mathematics devoted to how they relate to the solutions of the systems, called “group theory.” It is essential to advanced quantum mechanics, but beyond the scope of this book.

Key Points

- □ Symmetries of nature give rise to conserved numerical quantities.
- □ They are of particular interest in obtaining an understanding of complicated systems.
- □ For transformations like rotations and translations, you get the operator of the conserved quantity from infinitesimal rotations or translations.
- □ The parity transformation, inversion, replaces every \vec{r} by $-\vec{r}$.
- □ Parity is not conserved if the weak nuclear force plays a role.

6.3 Unsteady Perturbations of Systems

Section Abstract

This section takes a general look at what happens to a system, say an hydrogen atom, that can be in two different relevant energy eigenstates

and you poke at it with a perturbation, say an electromagnetic field. A typical application is the emission and absorption of radiation by atoms.

First an equation governing the evolution of the system is derived. It is then observed that perturbations can act either way: if a perturbation can excite a system, then that same perturbation can also de-excite the system. The latter allows photons to de-excite excited atoms, producing more photons. That leads to a run-away process that is the basic principle behind lasers.

The section continues with a fairly detailed analysis of exactly how radiation interacts with atoms. It is found that restrictions due to the demands of conservation laws can greatly slow down radiative processes. The analysis is of interest to understand the emission and absorption of light. Many of the ideas are also of importance for understanding nuclear decay, even though that is not explicitly covered. Readers with little interest in such areas might consider skipping the final sections.

The purpose of this section is to examine two-state systems that are perturbed. To analyze such systems, the energy eigenstates of the *unperturbed* system will be used to describe the system both with and without the perturbation. So, let ψ_L be the unperturbed lowest energy state and ψ_H be the unperturbed highest energy, or “excited”, state. In principle, ψ_L and ψ_H can be any two energy eigenstates of a system, but to be concrete, think of ψ_L as the ψ_{100} ground state of an hydrogen atom, and of ψ_H as an excited state like ψ_{210} , with energy $E_2 > E_1$.

Key Points

- The problem involves a low energy state ψ_L , an excited state ψ_H , and a perturbation.
-

6.3.1 Schrödinger equation for a two-state system

For a perturbed two-state system, by assumption the wave function can be approximated as a combination of the two unperturbed eigenstates:

$$\Psi = a\psi_L + b\psi_H \quad |a|^2 + |b|^2 = 1 \quad (6.14)$$

where $|a|^2$ is the probability that the system can be found at the lower energy E_L , and $|b|^2$ the probability that it can be found at the higher energy E_H . The sum of the two probabilities must be one; the two-state system must be found in either state, {A.46}.

The time evolution of a quantum system is given by the Schrödinger equation $i\hbar\dot{\Psi} = H\Psi$, where the dot indicates a time derivative. Here that becomes:

$$i\hbar(\dot{a}\psi_L + \dot{b}\psi_H) = H(a\psi_L + b\psi_H)$$

The still unspecified Hamiltonian H includes the perturbation.

Separate equations for \dot{a} and \dot{b} can be obtained by taking inner products with $\langle\psi_L|$, respectively $\langle\psi_H|$ and using orthonormality in the left hand side:

$$\boxed{i\hbar\dot{a} = H_{LL}a + H_{LH}b \quad i\hbar\dot{b} = H_{HL}a + H_{HH}b} \quad (6.15)$$

which involves the following “Hamiltonian coefficients”

$$\boxed{\begin{aligned} H_{LL} &= \langle\psi_L|H\psi_L\rangle & H_{LH} &= \langle\psi_L|H\psi_H\rangle \\ H_{HL} &= \langle\psi_H|H\psi_L\rangle & H_{HH} &= \langle\psi_H|H\psi_H\rangle \end{aligned}} \quad (6.16)$$

Note that H_{LL} and H_{HH} are real, (1.16) and that H_{LH} and H_{HL} are complex conjugates, $H_{HL} = H_{LH}^*$.

A general analytical solution to the system (6.15) cannot be given, but you can get rid of half the terms in the right hand sides. The trick is to define new coefficients \bar{a} and \bar{b} by

$$\bar{a} = ae^{i \int H_{LL} dt / \hbar} \quad \bar{b} = be^{i \int H_{HH} dt / \hbar} \quad (6.17)$$

The new coefficients \bar{a} and \bar{b} are physically just as good as a and b : the probabilities are given by the square magnitudes of the coefficients, and the square magnitudes of \bar{a} and \bar{b} are exactly the same as those of a and b . That is because the exponentials are of magnitude one. Also, the initial conditions, call them a_0 and b_0 , are unchanged assuming you choose the integration constants appropriately.

The equations for \bar{a} and \bar{b} are a lot simpler; substituting the definitions into (6.15) and simplifying:

$$\boxed{i\hbar\dot{\bar{a}} = \bar{H}_{LH}\bar{b} \quad i\hbar\dot{\bar{b}} = \bar{H}_{HL}\bar{a}} \quad (6.18)$$

where

$$\boxed{\bar{H}_{LH} = \bar{H}_{HL}^* = H_{LH}e^{-i \int (H_{HH} - H_{LL}) dt / \hbar}} \quad (6.19)$$

Key Points

- Equations for a perturbed two state system were derived.
- The amplitudes \bar{a} and \bar{b} of the lower respectively higher energy states evolve according to (6.18).

- The perturbation enters these equations in the disguised form of a reduced Hamiltonian coefficient \bar{H}_{LH} .
 - This reduced Hamiltonian coefficient follows from (6.19), which in turn is in terms of the original Hamiltonian coefficients (6.16).
-

6.3.2 Spontaneous and stimulated emission

For a two state system, the amplitudes \bar{a} and \bar{b} of the lower and higher energy states evolve according to (6.18) of the previous subsection. It can be seen that these equations have a remarkable property: for every solution \bar{a}, \bar{b} there is a second solution $\bar{a}_2 = \bar{b}^*, \bar{b}_2 = -\bar{a}^*$ that has the probabilities of the low and high energy states exactly reversed. It means that

A perturbation that lifts a system out of the ground state will equally take that system out of the excited state.

It is a consequence of the Hermitian nature of the Hamiltonian; it would not apply if \bar{H}_{LH} was not equal to \bar{H}_{HL}^* .

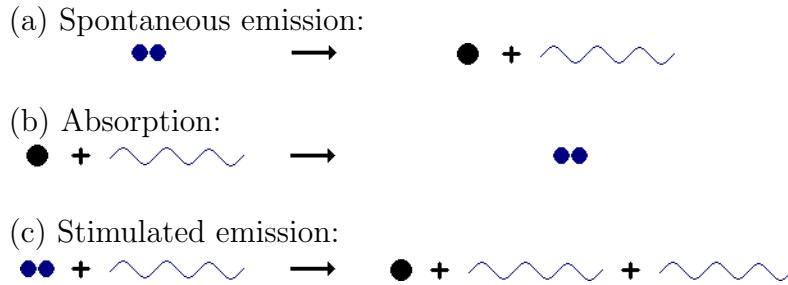


Figure 6.4: Emission and absorption of radiation by an atom.

Consider again the example of the atom. If the atom is in the excited state, it can spontaneously emit a photon, a quantum of electromagnetic energy, transitioning back to the ground state, as sketched in figure 6.4(a). That is spontaneous emission; the emitted photon will have an energy E_{photon} and an electromagnetic frequency ω_p given by:

$$E_{\text{photon}} \equiv \hbar\omega_p = E_{\text{H}} - E_{\text{L}} \quad (6.20)$$

The inverse of this process is where you perturb the ground state atom with an electromagnetic wave of frequency ω_p and the atom absorbs one photon of energy from that wave, entering the excited state. That is absorption, as sketched in figure 6.4(b). But according to the reversed solution above, there must then

also be a corresponding process where the same perturbing photon takes the system out of the *excited* state back to the *ground* state, figure 6.4(c). Because of energy conservation, this process, called “stimulated emission”, will produce a *second* photon.

It is the operating principle of the laser: if you have a collection of atoms all in the excited state, you can create a runaway process where a single photon stimulates an atom to produce a second photon, and then those two photons go on to produce two more, and so on. The result will be monochromatic, coherent light, since all its photons originate from the same source.

Note that you must initially have a “population inversion,” you must have more excited atoms than ground state ones, because absorption competes with stimulated emission for photons. Indeed, if you have a 50/50 mixture of ground state and excited atoms, then the processes of figures 6.4(b) and 6.4(c) exactly cancel each other’s effects.

Going back to spontaneous emission; as has been mentioned in section 6.1.3, there is really no such thing. The Schrödinger equation shows that an excited atom will maintain its energy indefinitely if not perturbed. Spontaneous emission, figure 6.4(a), is really stimulated emission, figure 6.4(c), in which, loosely speaking, the triggering photon jumps into and out of existence due to the quantum fluctuations of the electromagnetic field. Details of that process are in chapter 12.2.4.

Key Points

- □ Perturbations that will excite a two-state system will equally de-excite it.
- □ In lasers, photon perturbations cause excited atoms to release additional photons.

6.3.3 Effect of a single wave

This subsection will derive the basic equations for the interaction between an atom and a single electromagnetic wave. It will be assumed that the atom is hydrogen, or at least that interactions between the electrons can be ignored.

The wave

The perturbing electromagnetic field will be assumed to be a monochromatic wave that is propagating along the y -axis and is polarized in the z -direction. Such a wave takes the form, (10.28):

$$\vec{E} = \hat{k}E_0 \cos(\omega(t - y/c) - \phi) \quad \vec{B} = \hat{i}\frac{1}{c}E_0 \cos(\omega(t - y/c) - \phi).$$

where \vec{E} is the electric field strength, \vec{B} the magnetic field strength, the constant E_0 is the amplitude of the electric field, $\omega > 0$ the angular frequency of the wave, c the speed of light, and ϕ is some unimportant phase angle.

But this can be greatly simplified. At non relativistic velocities, the charged electron primarily reacts to the electric field, so the magnetic field can be ignored. Moreover, the atom, supposed to be at the origin, is so small compared to the typical wave length of an electromagnetic wave, (assuming it is light and not an X-ray,) that y can be put to zero. Then the electromagnetic field simplifies to a spatially uniform electric field:

$$\vec{E} = \hat{k}E_0 \cos(\omega t - \phi) \quad (6.21)$$

The Lyman-transition wave lengths are of the order of a thousand Å, and the atom about one Å, so this approximation seems reasonable enough.

Key Points

- The electromagnetic perturbation has been identified.
- It is by approximation a spatially uniform electric field.

The Hamiltonian coefficients

The previous subsection approximated the electromagnetic field of a typical single electromagnetic wave as a spatially uniform electric field. The question now is what effect this field has on the atom. That must be answered by finding the Hamiltonian due to the perturbation and then the Hamiltonian coefficients.

If the electric field would not vary with time, the energy of the charged electron in the electric field would be exactly given by

$$H_1 = eE_0 \cos(\omega t - \phi)z \quad (6.22)$$

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 $E_0 \cos(\omega t - \phi)$ that of the gravity strength g , and z that of the height h .

Of course, the cosine is not a constant, but varies with time. However, the time variation is typically slow compared to the relevant time scale of the electron. Therefore the above expression for the perturbation Hamiltonian H_1 can be used as an approximation.

To it the unperturbed Hamiltonian of the hydrogen atom must be added; that one was written down in chapter 3.2.1, but its form is not important for the effect of the perturbation, and it will just be referred to as H_0 . So the total Hamiltonian is

$$H = H_0 + H_1$$

with H_1 as above.

Now the Hamiltonian matrix coefficients (6.16) are needed. The first one is H_{LL} :

$$H_{LL} = \langle \psi_L | H_0 + H_1 | \psi_L \rangle = E_L + eE_0 \cos(\omega t - \phi) \langle \psi_L | z | \psi_L \rangle$$

The inner product $\langle \psi_L | H_0 | \psi_L \rangle$ produces the lower atom energy E_L because H_0 is the atomic Hamiltonian. As far as the final inner product is concerned, that is zero. That can be seen from the symmetry properties of the eigenfunctions of the hydrogen atom as given in chapter 3.2.4. The square magnitude of an atomic wave function is the same at an arbitrary position \vec{r} as it is at the opposite position $-\vec{r}$. However, z is of opposite sign at $-\vec{r}$. Therefore, negative z values integrate away against positive ones.

It follows that H_{LL} is just the lower atom energy E_L . Similarly, H_{HH} is just the higher atom energy E_H .

For H_{LH} , the inner product with H_0 is zero, since ψ_L and ψ_H are orthogonal eigenfunctions of H_0 , and the inner product with H_1 gives:

$$H_{LH} = eE_0 \cos(\omega t - \phi) \langle \psi_L | z | \psi_H \rangle$$

To get the Hamiltonian coefficient of the simplified evolution equations (6.18), according to (6.19) the coefficient H_{LH} needs to be multiplied with

$$e^{-i \int (H_{HH} - H_{LL}) dt / \hbar} = e^{-i \int (E_H - E_L) dt / \hbar}$$

By definition

$$E_H - E_L = \hbar\omega_p$$

with ω_p the frequency of the photon released when the atom transitions from the high energy state to the low one. So the factor to multiply H_{LH} with is simply $e^{-i\omega_p t}$.

The Hamiltonian coefficient of the simplified system becomes

$$\overline{H}_{LH} = eE_0 \cos(\omega t - \phi) e^{-i\omega_p t} \langle \psi_L | z | \psi_H \rangle \quad (6.23)$$

It governs the evolution of an atom under an electromagnetic wave that is polarized in the z -direction. The amplitude of the wave is E_0 and its frequency is ω , while ω_p is the frequency of the photon emitted in a transition from the higher atom energy level to the lower.

Key Points

- □ The evolution of an atom under an electromagnetic wave that is polarized in the z -direction is governed by a reduced Hamiltonian coefficient.
- □ By approximation, the value of this coefficient is given by (6.23).

6.3.4 Forbidden transitions

According to the final result of the previous subsection, the effect of an electromagnetic wave on an atomic transition between two states ψ_L and ψ_H is proportional to the inner product $\langle \psi_L | z | \psi_H \rangle$. If that inner product is zero, the electromagnetic wave cannot cause such a transition, to the approximations made.

However, the assumed wave had its electric field in the z -direction. A more general electromagnetic field has electric field components in all three axial directions. Such a field can cause transitions as long as at least one of the three inner products of the form $\langle \psi_L | r_i | \psi_H \rangle$, with r_i equal to x , y , or z , is nonzero. Because these inner products vaguely resemble integrals that produce the so-called electric dipole strength of a charge distribution, chapter 10.5, transitions caused by them are called “electric dipole transitions.” They could not be called “uniform electric field transitions,” because that would have been meaningful and understandable.

If all three inner products $\langle \psi_L | r_i | \psi_H \rangle$ are zero, then the transition cannot occur to the approximations made. Such transitions are called “forbidden transitions.” An example is the hydrogen 2s to 1s transition, i.e. $\psi_H = \psi_{200}$ and $\psi_L = \psi_{100}$. Both of these states are spherically symmetric, making the inner product $\langle \psi_L | r_i | \psi_H \rangle$ zero by symmetry: the negative values of r_i integrate away against the positive values. So, with no perturbation effect left, the prediction must then unavoidably be that the excited 2s state does not decay!

The term is misleading, however: forbidden transitions often take place just fine, even if the electric dipole approximation says they cannot. Reasons for them to occur anyway can be the ignored spatial variations in the electric field, leading to so-called electric multipole transitions, or the also ignored interaction with the magnetic field, leading to magnetic dipole or multipole transitions. However, since these are very small effects, “forbidden” transitions do take much longer to occur than normal electric dipole ones by orders of magnitude.

Key Points

- The approximate problem identified for the interaction of an atom with an electromagnetic wave is called the electric dipole approximation.
- It is possible that the electric dipole approximation predicts that electromagnetic waves have no effect at all on transitions between the two considered atom states. If so, the transition is called forbidden.
- However, because of the approximations made, often forbidden transitions take place just fine.
- Forbidden transitions could be electric multipole, magnetic dipole, or magnetic multipole ones.

- Forbidden transitions are slower than electric dipole ones.
-

6.3.5 Selection rules

Electric dipole transitions between hydrogen atom states cannot occur unless the quantum numbers of the states involved satisfy certain conditions. These conditions are called “selection rules.” The quantum numbers appearing in the rules are the orbital azimuthal quantum number l , giving the square orbital angular momentum as $l(l + 1)\hbar^2$, the orbital magnetic quantum number m , giving the orbital angular momentum in the chosen z -direction as $m\hbar$, and the spin magnetic quantum number $m_s = \pm \frac{1}{2}$, called spin up respectively down, giving the spin angular momentum in the z -direction as $m_s\hbar$. For electric dipole transitions to occur, these quantum numbers must satisfy, {A.47},

$$l_H = l_L \pm 1 \quad m_H = m_L \text{ or } m_L \pm 1 \quad m_{s,H} = m_{s,L} \quad (6.24)$$

It should be noted that in a more sophisticated analysis of the hydrogen atom, there is a slight interaction between orbital and spin angular momentum in the atom, chapter 12.1.6. As a result the correct energy eigenfunctions no longer have a definite z -component of orbital angular momentum nor of spin, and the above rules are no longer really right. The energy eigenfunctions do have definite values for the square magnitude $J^2 = j(j + 1)\hbar^2$ of the combined angular momentum $\hat{J} = \hat{L} + \hat{S}$ and its z -component $\hat{J}_z = m_j\hbar$. In those terms the appropriate selection rules become

$$l_H = l_L \pm 1 \quad j_H = j_L \text{ or } j_L \pm 1 \quad m_{j,H} = m_{j,L} \text{ or } m_{j,L} \pm 1 \quad (6.25)$$

If these selection rules are not satisfied, the transition is called forbidden. However, the transition may still occur through a different mechanism. One possibility is a “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 10.6. The selection rules in this case are, {A.47},

$$l_H = l_L \quad m_H = m_L \text{ or } m_L \pm 1 \quad m_{s,H} = m_{s,L} \text{ or } m_{s,L} \pm 1 \quad (6.26)$$

In addition either the orbital or the spin magnetic quantum numbers must be unequal. In the presence of spin-orbit coupling, that becomes

$$l_H = l_L \quad j_H = j_L \text{ or } j_L \pm 1 \quad m_{j,H} = m_{j,L} \text{ or } m_{j,L} \pm 1 \quad (6.27)$$

Transitions may also occur through spatial variations in the electric or magnetic fields. The strongest of these are the so-called “electric quadrupole transitions.” These must satisfy the selection rules, {A.47},

$$l_H = l_L \text{ or } l_L \pm 2 \quad m_H = m_L \text{ or } m_L \pm 1 \text{ or } m_L \pm 2 \quad m_{s,H} = m_{s,L} \quad (6.28)$$

In addition $l_H = l_L = 0$ is not possible for such transitions.

Key Points

- The quantum numbers of two atom states must satisfy certain selection rules for transitions between them to be possible.
- The rules depend on the type of transition.
- Some example rules were given.

6.3.6 Angular momentum conservation

According to the selection rules of the previous subsection, the 2s, or ψ_{200} , hydrogen atom state cannot decay to the 1s, or ψ_{100} , ground state through an electric dipole transition, because $l_H = l_L = 0$. That transition is therefore forbidden. That does not mean it does not take place; just forbid your kids something. But this particular excited state cannot return to the ground state through a magnetic dipole transition either, because without orbital angular momentum, only the spin responds to the magnetic field. It cannot return to the ground state through an electric quadrupole transition, because these do not allow $l_H = l_L = 0$ transitions.

In fact, the 2s state cannot decay to the ground state through any electric or magnetic multipole transition. There is a simple argument to see why not, much simpler than trying to work out the selection rules for all possible multipole transitions. Both the 2s and 1s states have zero angular momentum. Conservation of angular momentum then requires that the photon emitted in the transition must have zero angular momentum too, and a photon cannot, [10].

A photon is a boson that has spin one. Now 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, $j = 0$. In fact, it can be deduced from figure 10.6 that that happens for a linear combination of the three states that each have zero net angular momentum in the chosen z -direction:

$$|j=0\rangle = \sqrt{\frac{1}{3}}|m=1, m_s=-1\rangle - \sqrt{\frac{1}{3}}|m=0, m_s=0\rangle + \sqrt{\frac{1}{3}}|m=-1, m_s=1\rangle$$

However, a photon is not a normal particle; it is a relativistic particle with rest mass zero that can only move at the speed of light. It turns out that when the

z -axis is taken in the direction of propagation, the photon can only have $m_s = 1$ or $m_s = -1$; the state $m_s = 0$ does not exist, chapter 12.2.3. So the combination above with zero net angular momentum is not possible for a photon.

(There is a more intuitive way to understand this peculiar behavior. According to quantum mechanics, angular momentum is related to angular variation. An atom that has zero angular momentum is spherically symmetric. Now classical electrostatics says that a spherical charge distribution has the same electric field outside the atom regardless of how the radial charge distribution changes inside the atom. And if the electromagnetic field outside the atom does not change with time, it cannot radiate energy away.)

So, what does happen to the $2s$ state? It turns out that in an extremely high vacuum, in which the state is not messed up by collisions with other atoms or particles, the dominant decay is by the emission of two photons, rather than a single one. This takes forever on quantum scales; the $2s$ state survives about a tenth of a second rather than maybe a nanosecond for an electric dipole transition. You see why ψ_{210} was cited as the typical excited state in this section, rather than the more obvious ψ_{200} .

This example illustrates how powerful conservation law arguments can be. Most of the selection rules of the previous subsection, as well as many not listed there, can be understood using a simple rule for the angular momentum of the emitted photon, [10]: The net angular momentum of the emitted photon can be taken to be $j = 1$ in dipole transitions. It increases by one unit for each next higher level: quadrupole transitions have $j = 2$, octupole ones $j = 3$, etcetera.

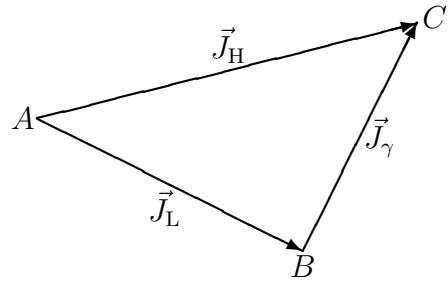


Figure 6.5: Triangle inequality.

The angular momentum of the photon restricts what the angular momentum of the final atomic state can be. In particular, conservation of angular momentum says that the angular momentum vector of the excited state ψ_H must be the sum of that of the lower state ψ_L plus that of the photon:

$$\hat{J}_H = \hat{J}_L + \hat{J}_\gamma$$

where γ indicates the photon.

Now if two vectors of lengths J_L and J_γ are added together in *classical physics*, as in figure 6.5, then the length of the resultant vector satisfies the so-called “triangle inequality”

$$|J_L - J_\gamma| \leq J_H \leq J_L + J_\gamma$$

To see why, note from the figure that J_H must be less than $J_L + J_\gamma$ or the combined lengths of the sides AB and BC in figure 6.5 would not be enough to reach point C from point A even if aligned with AC . Similarly J_H has to be long enough to cover the length difference between J_L and J_γ .

Classical physics also says that the change in the atomic z -momentum $J_{H,z} - J_{L,z}$ equals the z -component $J_{\gamma,z}$ of the photon angular momentum. The magnitude of that component cannot be more than the length of the complete vector J_γ , so

$$|J_{H,z} - J_{L,z}| \leq J_\gamma$$

In quantum mechanics however, things are somewhat different. Angular momentum is quantized. The length of an angular momentum vector is given in terms of an azimuthal quantum number j as $J = \sqrt{j(j+1)\hbar}$. In these terms, the triangle inequality becomes, chapter 10.1,

$$|j_L - j_\gamma| \leq j_H \leq j_L + j_\gamma \quad (6.29)$$

Also a z -component of angular momentum is given in terms of a magnetic quantum number m_j as $J_z = m_j\hbar$. It is still true that $J_{H,z} - J_{L,z} = J_{\gamma,z}$, but now it is the magnetic quantum number $m_{j,\gamma}$ whose magnitude is limited by the azimuthal quantum j_γ , so

$$|m_{j,H} - m_{j,L}| \leq j_\gamma \quad (6.30)$$

The quantum numbers j and m_j must be integer or half integer.

Ignoring spin-orbit interaction and transitions that simply flip over the electron spin, electron spin can be ignored. In that case the relevant atomic angular momentum can be taken to be just orbital, with azimuthal quantum number l instead of j . Then the requirements become

$$|l_L - j_\gamma| \leq l_H \leq l_L + j_\gamma \quad |m_H - m_L| \leq j_\gamma$$

Orbital angular momentum quantum numbers l and m must be integer, as must be j_γ .

To use these inequalities, the angular momentum quantum number j_γ of the photon is needed. It turns out that in dipole transitions the emitted photon has $j_\gamma = 1$. Then the above inequalities show that in such transitions the atomic angular momentum quantum numbers cannot change by more than one unit

each. Quadrupole transitions have $j_\gamma = 2$. They allow two-unit atomic angular momentum changes. However, quadrupole transitions are much slower than dipole ones. They can normally be ignored if dipole transitions are possible. Octupole transitions have $j_\gamma = 3$ and allow three-unit changes, but they are much slower still than dipole transitions. In general, 2^ℓ -pole transitions have $j_\gamma = \ell$ and can change the nuclear angular momentum quantum numbers by up to that amount. The larger ℓ , the slower the transition.

The quantum triangle inequality also implies that transitions from $j_L = 0$ to $j_H = 0$ are not possible at any multipole level ℓ .

Some selection rules are not explained by angular momentum conservation. For example, how come that l must change one unit in electric dipole transitions, while it must remain the same in magnetic dipole transitions? Angular momentum conservation does not explain that. These gaps are filled in by the next subsection.

Key Points

- □ Angular momentum conservation restricts what transitions are possible.
- □ A 2^ℓ -pole transition emits a photon with angular momentum $j_\gamma = \ell$.
Here ℓ is a positive integer.
- □ The larger ℓ , the slower the transition.
- □ Quantum rules must be used to find the restrictions put on the atomic angular momenta by the photon angular momentum. They are given by (6.29) and (6.30).

6.3.7 Parity

The previous subsection demonstrated the power of conservation laws in figuring out what physical interactions can occur. Physicists such as Wigner have therefore developed these arguments to a fine art. One reason is that they continue to work for systems that are not by far as well understood as an hydrogen atom. In particular they are of great value in understanding atomic nuclei.

Mathematicians discovered that conservation laws relate to fundamental symmetries of nature, section 6.2. For example, conservation of angular momentum turns out to be a mathematical rephrasing of the fact that the physics of a system is the same regardless of how the system is oriented compared to the surrounding empty space.

Physicists then identified another helpful symmetry of nature; nature looks perfectly fine when viewed in a mirror. Your left hand changes into a right hand when seen in a mirror, but there is no fundamental difference between the two;

lots of people write with their left hand. Physicists figured nature had to behave exactly the same way when seen in a mirror. And it does, as long as the weak nuclear force does not play a part. That force can safely be ignore here.

In three dimensions, a mathematically neat way of doing the mirroring is to replace every \vec{r} by $-\vec{r}$. This is equivalent to a mirroring followed by a 180° rotation around the line normal to the mirror. The operator that does that is called the “inversion operator.” Its eigenvalues must have magnitude one, because the integrated square magnitude of the wave function does not change when seen in the mirror. The eigenvalues must also be real, because the inversion operator is Hermitian; taking it to the other side in inner product amounts to just a change in the names of the integration variables. Therefore, the eigenvalues can only be plus or minus one. If the eigenvalue is one, parity is called even, if it is minus one, parity is odd.

Parity behaves somewhat different from angular momentum. While the angular momenta of the individual particles of a system must be *added* together, their parities must be *multiplied* together. Two particles each with parity -1 have a net parity of $(-1)^2 = 1$. (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 sign when $\vec{r}_1 \rightarrow -\vec{r}_1$ and ψ_2 changes sign when $\vec{r}_2 \rightarrow -\vec{r}_2$, then the total wave function Ψ stays the same. 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 in terms of a logarithm of an eigenvalue. For details, see section 6.2.)

Now apply that to transitions between an atomic high energy state ψ_H and a low energy state ψ_L plus a photon. If the parity of the lower energy state ψ_L is indicated by π_L and that of the photon by π_γ , then their combined parity is $\pi_L\pi_\gamma$. Since parity is conserved, that must equal the parity of the high energy state; $\pi_H = \pi_L\pi_\gamma$. It follows that if the photon has even parity, $\pi_\gamma = 1$, then the atomic parity must stay the same during the transition. If the photon has odd parity, $\pi_\gamma = -1$, then the atomic parity must flip over.

It turns out that the photon emitted in allowed, or electric dipole, transitions can be taken to have odd parity, [10]. So during allowed transitions, the atomic parity must flip over. The question is now, what does that mean in terms of the atomic quantum numbers. As shown in note {A.15}, an electron in an orbit with an even orbital azimuthal quantum number l has even parity. An electron in an orbit with an odd azimuthal quantum number l has odd parity. Therefore, in allowed transitions the azimuthal quantum number l must change from odd to even or vice versa. Combine that with the angular momentum constraint that the orbital angular momentum can change at most one unit, and you have that l must change by exactly one unit. That explains one of the selection rules given in 6.3.5.

The photon emitted in magnetic dipole transitions has even parity. That

means that the atomic parity cannot change. Angular momentum conservation says that l can change by up to one unit, but since such change would flip over the parity, l must stay unchanged. That explains the corresponding selection rule given in 6.3.5.

More generally, in electric 2^ℓ -pole transitions the atomic parity changes when ℓ is odd and stays unchanged when ℓ is even. It is just the opposite way around for magnetic 2^ℓ -pole transitions. For them, the atomic parity stays the same when ℓ is odd and changes when ℓ is even. In that sense, electric and magnetic multipole transitions complement each other. If the electric transition is forbidden by parity conservation at a level allowed by angular momentum conservation, then the magnetic one is possible and vice versa. However, since the electron in the hydrogen atom moves fairly slow compared to the speed of light, the magnetic transitions tend to be quite slow compared to the corresponding electric ones.

Key Points

- □ The inversion operator replaces any position vector \vec{r} by $-\vec{r}$. It has the effect of mirroring nature.
- □ The parity of a system is even if it stays the same under inversion.
The parity is odd if it changes sign.
- □ As long as the weak nuclear force is not a factor, parity is conserved.
- □ That puts restrictions on what atomic transitions are possible beyond those given by angular momentum conservation.

6.3.8 Absorption of a single weak wave

This subsection discusses what happens to an atom in the ground state if you perturb it with a single coherent wave of electromagnetic radiation over a time interval t_c .

Like in previous subsections, it will be assumed that only two atomic energy states are involved, a low energy state ψ_L and an excited state ψ_H . The amplitudes of these states are \bar{a} and \bar{b} respectively. They satisfy the evolution equations

$$i\hbar\dot{\bar{a}} = \overline{H}_{LH}\bar{b} \quad i\hbar\dot{\bar{b}} = \overline{H}_{HL}\bar{a}$$

where the Hamiltonian perturbation coefficient was found to be

$$\overline{H}_{LH} = \overline{H}_{HL}^* = eE_0 \cos(\omega t - \phi) e^{-i\omega_p t} \langle \psi_L | z | \psi_H \rangle$$

Here E_0 is the amplitude of the perturbing electromagnetic wave and ω its frequency. Also ω_p is the frequency of the photon released in a transition from the excited state to the ground state.

To reduce the writing, it is convenient to introduce an abbreviation for the time-independent part of \bar{H}_{LH} :

$$\omega_1 \equiv \frac{1}{\hbar} e E_0 \langle \psi_L | z | \psi_H \rangle \quad (6.31)$$

As the symbol suggests, ω_1 has units of frequency like ω and ω_p . However, for the time being it is just a concise way of writing the effective strength level of the perturbation.

The evolution equations become in terms of ω_1 :

$$\dot{\bar{a}} = -i\omega_1 \cos(\omega t - \phi) e^{-i\omega_p t} \bar{b} \quad \dot{\bar{b}} = -i\omega_1 \cos(\omega t - \phi) e^{i\omega_p t} \bar{a} \quad (6.32)$$

Unfortunately, this system does not have a simple solution.

Simplification is needed. According to the Euler formula (1.5), the cosine comes apart into two exponentials:

$$\cos(\omega t - \phi) = \frac{e^{i(\omega t - \phi)} + e^{-i(\omega t - \phi)}}{2}$$

If this is substituted into (6.32), one exponential gives rise to a factor $e^{\pm i(\omega + \omega_p)t}$. Now for light, ω_p is extremely large, in the order of $10^{15}/\text{s}$, and adding ω makes it even larger, so this exponential fluctuates extremely rapidly in value and averages out to zero over any reasonable time interval. For that reason, this exponential is usually ignored. (If the relevant time interval t_c becomes so short that $\omega_p t_c$ is no longer large, this is no longer justified.)

Keeping only the exponential with the frequency opposite to ω_p , by approximation the equations (6.32) become:

$$\dot{\bar{a}} = -i\omega_1 e^{-i\phi} \frac{e^{i(\omega - \omega_p)t}}{2} \bar{b} \quad \dot{\bar{b}} = -i\omega_1 e^{i\phi} \frac{e^{-i(\omega - \omega_p)t}}{2} \bar{a} \quad (6.33)$$

These equations can be solved analytically. In fact, the same equations are solved for nuclear magnetic resonance in chapter 10.8. But the solution is messy.

Therefore in this section an approximate analysis called “time-dependent perturbation theory” will be used. Time-dependent perturbation theory assumes that the level of perturbation, here given by ω_1 , is small.

It will also be assumed that the atom starts in the lower energy state, and more simply, from $a_0 = 1$, $b_0 = 0$. If \bar{a} starts at one and its changes, as given by (6.33), are small, then \bar{a} will stay about one. So in the equation for \bar{b} in (6.33), the factor \bar{a} can just be ignored. That allows it to be easily solved:

$$\bar{b}^L = \omega_1 e^{i\phi} \frac{e^{-i(\omega - \omega_p)t} - 1}{2(\omega - \omega_p)} \quad (6.34)$$

where the superscript L merely indicates that this solution starts from the lower energy state.

This solution is valid as long as \bar{a}^L remains close to one. That means that the effective strength ω_1 of the applied wave must be weak enough and/or the time interval t_c during which the wave is applied short enough. To be more precise, the formula requires that $\omega_1 t_c$ is small, in addition to the earlier requirement that $\omega_p t_c$ is large. (To check that, note that the formula should definitely not be allowed to fail in the critical case that $\omega = \omega_p$; then \bar{b} is only small if $\omega_1 t$ is.)

Of course, a real-life atom in a gas would suffer other perturbations than just the applied electromagnetic wave. Periodically, collisions with neighboring atoms will occur, and the preexisting ambient electromagnetic field will disturb the atoms too. However, if the electromagnetic wave is applied for a very short time interval, there will not be enough time for any of these effects to act, and the above expression correctly describes the evolution. The range of times in which interactions with the environment are unlikely to occur is called the “collisionless regime.” This is typically a very short time interval. To get a decent collisionless response from the atom during such a short time interval, the effective field strength ω_1 can be jacked way up by using concentrated laser light.

At the time t_c that the wave is turned off, the atom will have a “transition probability” that it can be found in the excited state equal to the square magnitude of \bar{b}^L . To find it, take a factor $e^{i\phi - \frac{1}{2}i(\omega - \omega_p)t_c}$ out of (6.34), use Euler on the remainder, and take the square absolute value. That gives the transition probability $P_{L \rightarrow H}$ from low to high as:

$$P_{L \rightarrow H} \equiv |\bar{b}^L|^2 = \frac{1}{4}|\omega_1|^2 t_c^2 \left(\frac{\sin \frac{1}{2}(\omega - \omega_p)t_c}{\frac{1}{2}(\omega - \omega_p)t_c} \right)^2 \quad \frac{1}{\omega_p} \ll t_c \ll \frac{1}{\omega_1} \quad (6.35)$$

The made assumptions are explicitly listed while the definition of ω_1 can be found in (6.31).

Take a large number of atoms in the ground state and apply the wave. Then just sit back and give the surroundings time to “measure” the atoms. The larger the transition probability, the more of these atoms will be found to be in the elevated energy state and to transition back to the ground state while emitting a photon of energy $\hbar\omega_p$. (The excited atoms have some average “lifetime” dependent on the ambient electromagnetic field. Compare the remaining subsections.) So the higher the transition probability above, the more photons you will get.

There is only a decent transition probability for perturbation frequencies ω close to the photon frequency ω_p . In particular, the fraction within parentheses in (6.35) has a maximum of one at $\omega = \omega_p$, (using l’Hôpital), and is negligibly small unless $(\omega - \omega_p)t_c$ is finite. Even in that range, the transition probability

cannot be more than $\frac{1}{4}|\omega_1|^2 t_c^2$, so $\omega_1 t_c$ must be finite too for a decent transition probability. True, $\omega_1 t_c$ must formally be small because of the small perturbation assumption, but you want to stretch it as far as you can, even if you would have to use the exact analysis instead. If both $(\omega - \omega_p)t_c$ and $\omega_1 t_c$ must be finite, then the range of frequencies ω around ω_p for which there is a decent response must be comparable to $|\omega_1|$. The physical meaning of $|\omega_1|$ is therefore as a frequency *range* rather than as a frequency by itself. It is the typical range of frequencies around the photon frequency ω_p for which there is a decent response to perturbations at this strength level.

Since a small range of frequencies can be absorbed, the observed line in the absorption spectrum is not going to be a mathematically thin line, but will have a small width. Such an effect is known as “spectral line broadening” {A.48}.

Key Points

- If an atom is perturbed by a coherent electromagnetic wave over a time interval t_c , then the final transition probability $P_{L \rightarrow H}$ to a higher energy state is given by (6.31) and (6.35).
- The time interval t_c must stay within the collisionless regime, in which the atom can be assumed to be undisturbed by its surroundings.
- The given expression for the transition probability requires that it remains small at time t_c . That is a consequence of a small perturbation assumption.
- Spectral lines are not truly mathematically thin lines. There is some broadening.

6.3.9 Absorption of incoherent radiation

Under normal conditions, an atom is not subjected to just a single electromagnetic wave, but to “broadband” incoherent radiation of all frequencies moving in all directions. In that case, you have to integrate the effects of all waves together.

In addition, usually interactions with the environment, including neighboring atoms, occur frequently instead of rarely. This is called the “collision-dominated regime.” Collisions between atoms may be modeled as elastic, since in the absence of interaction with the electromagnetic field, the Schrödinger equation conserves the probabilities of the global energy states of the particles. However, they will definitely randomize the coefficients a and b of the low and high energy states ψ_L and ψ_H of the individual atoms.

For a typical time interval t_c between collisions, the approximations made for a single wave in the previous subsection continue to hold. In particular, the

frequency ω_p of a photon released in a transition from a high to a low energy state may be of the order of $10^{15}/\text{s}$. Relevant times, such as atomic collisions times and spontaneous decay times, are normally much larger than 10^{-15} s . But they are still small enough that the transition probability that develops between collisions remains small, as long as the radiation level is low enough.

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 will be denoted as P_L . Similarly, the probability that an atom can be found in the higher energy state is indicated by P_H . For a single atom, these probabilities are given by the square coefficients $|a|^2$ and $|b|^2$. Therefore, P_L and P_H will be defined as the averages of $|a|^2$ respectively $|b|^2$ over all atoms. More simply, P_L can be identified with the fraction of atoms in the high energy state and P_H with the fraction in the high energy state.

In these terms it turns out, {A.49}, that the atom fractions P_L and P_H evolve in time according to the evolution equations

$$\boxed{\frac{dP_L}{dt} = -B_{L \rightarrow H}\rho(\omega_p) P_L + B_{H \rightarrow L}\rho(\omega_p) P_H + A_{H \rightarrow L} P_H + \dots} \quad (6.36)$$

$$\boxed{\frac{dP_H}{dt} = +B_{L \rightarrow H}\rho(\omega_p) P_L - B_{H \rightarrow L}\rho(\omega_p) P_H - A_{H \rightarrow L} P_H + \dots} \quad (6.37)$$

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_L of low energy atoms that is available to be excited. It is also proportional to the energy $\rho(\omega)$, per unit volume and per unit frequency range, of the electromagnetic radiation that does the excitation, evaluated at the photon frequency. The constant $B_{L \rightarrow H}$ is called the “transition rate” from L to H. Similarly, the second term in the right hand 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 $A_{H \rightarrow L}$ is called the “spontaneous emission rate.” The second equation can be understood similarly.

If there are transitions between more than two states involved, all their effects should be summed together; that is indicated by the dots in (6.36) and (6.37).

The constants 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. Also, Einstein in those pre-Schrödinger equation days treated the atoms as being in the low and high energy states for certain. So, so do the various textbooks. And these same textbooks typically define “measurement” as the sort of thing a physicist does in a lab. The implied notion that physicists would be carefully measuring the energy of each of countless atoms in the lab and in space (to get the atoms in energy eigenstates), on a continuing basis, is rather, ahem, interesting?

Anyway, as shown in note {A.49},

$$B_{L \rightarrow H} = B_{H \rightarrow L} = \frac{\pi |\langle \psi_L | e\vec{r} | \psi_H \rangle|^2}{3\hbar^2 \epsilon_0} \quad (6.38)$$

where $\epsilon_0 = 8.854\,19\,10^{-12} \text{ C}^2/\text{J m}$ is the permittivity of space. The spontaneous emission rate $A_{H \rightarrow L}$ is derived in the next subsection.

Key Points

- In the presence of an incoherent electromagnetic field with energy density $\rho(\omega)$, the fractions P_L of low energy atoms and P_H of high energy atoms satisfy the evolution equations (6.36) and (6.37).
- The coefficients in the equations are called the Einstein A and B coefficients.
- The B coefficients give the relative response of transitions to incoherent radiation. They are given by (6.38).

6.3.10 Spontaneous emission of radiation

Einstein derived the spontaneous emission rate of radiation based on a relatively simple argument. 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

$$\frac{dP_H}{dt} = B_{L \rightarrow H} \rho(\omega_p) P_L - B_{H \rightarrow L} \rho(\omega_p) P_H - A_{H \rightarrow L} P_H$$

where $\rho(\omega)$ is the ambient electromagnetic field energy density, ω_p 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 (6.37) derived in the previous section.

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 $\rho(\omega)$. The equilibrium version of the formula above gives the energy density as

$$\rho(\omega_p) = \frac{A_{H \rightarrow L}/B_{H \rightarrow L}}{(B_{L \rightarrow H}P_L/B_{H \rightarrow L}P_H) - 1}$$

Equating this to Planck's blackbody spectrum as derived in chapter 9.14.5 gives

$$\frac{A_{H \rightarrow L}/B_{H \rightarrow L}}{(B_{L \rightarrow H}P_L/B_{H \rightarrow L}P_H) - 1} = \frac{\hbar}{\pi^2 c^3} \frac{\omega_p^3}{e^{\hbar\omega_p/k_B T} - 1}$$

The atoms can be modeled as distinguishable particles. Therefore the ratio P_H/P_L can be found from the Maxwell-Boltzmann formula of chapter 5.14; that gives the ratio as $e^{-(E_H-E_L)/k_B T}$, or $e^{-\hbar\omega_p/k_B T}$ in terms of the photon frequency. It then follows that for the two expressions for $\rho(\omega_p)$ to be equal,

$B_{L \rightarrow H} = B_{H \rightarrow L} \quad \frac{A_{H \rightarrow L}}{B_{H \rightarrow L}} = \frac{\hbar\omega_p^3}{\pi^2 c^3}$

(6.39)

where $B_{H \rightarrow L}$ was given in (6.38).

That $B_{L \rightarrow H}$ must equal $B_{H \rightarrow L}$ was already mentioned in section 6.3.2. But it was not self-evident when Einstein wrote the paper; Einstein really invented stimulated emission here. The valuable new result is the formula for the spontaneous emission rate $A_{H \rightarrow L}$.

If you like to think of spontaneous emission as being due to perturbations by the ground state electromagnetic field, the ratio $A_{H \rightarrow L}/B_{H \rightarrow L}$ above would be the ground state energy density ρ_{gs} at frequency ω_p . In terms of the ideas of chapter 9.14.5, it corresponds to one photon in each radiation mode. That is not quite right: in the ground state of the electromagnetic field there is half a photon in each mode. It is just like a harmonic oscillator, which has half an energy quantum $\hbar\omega$ left in its ground state. The photon that the excited atom interacts with is its own, through a twilight effect, chapter 12.2.4.

Still, as (6.39) indicates, the “vacuum energy” of empty space becomes infinite like ω^3 at infinite frequencies. In the early days of relativistic quantum mechanics, this was seen as an ununderstood error in the theory that had to be fixed up some way. However, over time physicists came to accept that such energy blow up is really what happens; it gives rise to confirmed theoretical predictions. Presumably the blow-up describes extremely short scale processes that cannot be observed. Their nature may remain unknown to us, because what we observe is the same regardless of the precise details of these short-scale processes.

Due to the spontaneous emission rate, atoms pumped up to an excited energy state ψ_H will decay to lower energy states over time when left alone, say isolated in a closed box at absolute zero temperature. The governing equation is

$$\frac{dP_H}{dt} = -[A_{H \rightarrow L_1} + A_{H \rightarrow L_2} + A_{H \rightarrow L_3} + \dots] P_H$$

where the sum is over all the lower energy states that exist. The resulting expression for the number of atoms that can be found in the elevated energy state at a given time is

$$P_H(t) = P_H(0)e^{-t/\tau} \quad \tau = \frac{1}{A_{H \rightarrow L_1} + A_{H \rightarrow L_2} + A_{H \rightarrow L_3} + \dots}$$

(6.40)

The constant τ is called the “lifetime” of the excited state. No it is not really a lifetime, except in some average sense, but $1/\tau$ is the average fraction of excited atoms that disappears per unit time. The “half-life”

$$\tau_{1/2} = \tau \ln 2$$

is the time it takes for the number of atoms that can be found in the excited state to decrease by about half.

Isolated atoms in a box that is at room temperature are bathed in thermal blackbody radiation as well as vacuum energy. So stimulated emission will add to spontaneous emission. However, at room temperature, blackbody radiation has negligible energy in the visible light range, and transitions in this range will not really be affected.

Key Points

- Spontaneous decay of a fraction P_H of excited atoms follows the evolution equation (6.40).
- The constants in it are called the Einstein A coefficients. They are given by (6.39) in terms of the B coefficients derived in the previous subsection.
- The half-life is the time that it takes for half of a large number of excited atoms to decay.
- The half-life of the excited atom state is defined in terms of the A coefficients by (6.3.10).

6.4 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

$$\hat{\vec{r}} = (\hat{x}, \hat{y}, \hat{z})$$

and the linear momentum operator

$$\hat{\vec{p}} = (\hat{p}_x, \hat{p}_y, \hat{p}_z) = \frac{\hbar}{i} \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

are the same for all systems. So, you only need to find their eigenfunctions once.

6.4.1 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:

$$\begin{aligned}\hat{x}R_{\underline{x}\underline{y}\underline{z}}(x, y, z) &= \underline{x}R_{\underline{x}\underline{y}\underline{z}}(x, y, z) \\ \hat{y}R_{\underline{x}\underline{y}\underline{z}}(x, y, z) &= \underline{y}R_{\underline{x}\underline{y}\underline{z}}(x, y, z) \\ \hat{z}R_{\underline{x}\underline{y}\underline{z}}(x, y, z) &= \underline{z}R_{\underline{x}\underline{y}\underline{z}}(x, y, z)\end{aligned}$$

(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 6.6.

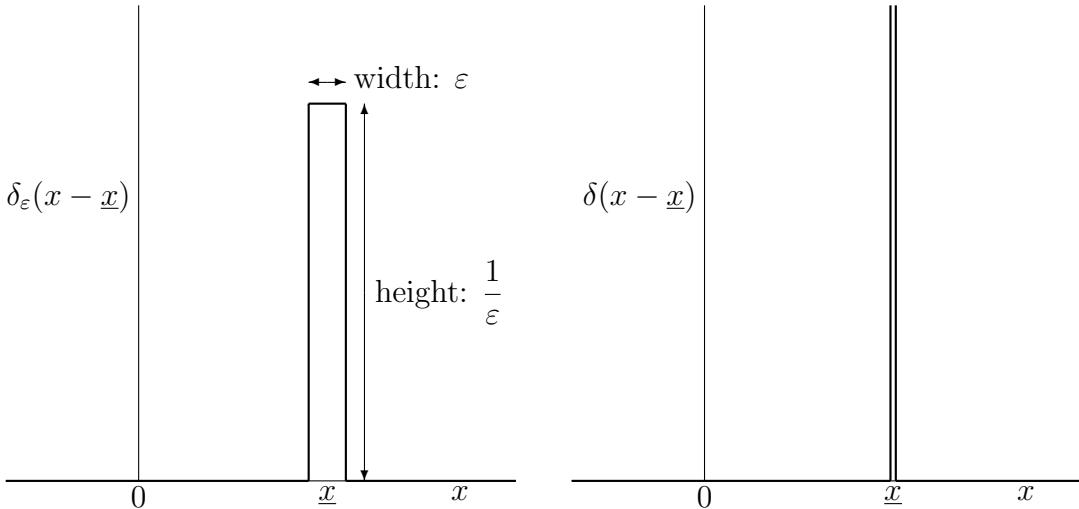


Figure 6.6: Approximate Dirac delta function $\delta_\varepsilon(x - \underline{x})$ is shown left. The true delta function $\delta(x - \underline{x})$ is the limit when ε becomes zero, and is an infinitely high, infinitely thin spike, shown right. It is the eigenfunction corresponding to a position \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, $f(\underline{x})$:

$$\langle \delta(x - \underline{x}) | f(x) \rangle = \int_{x=-\infty}^{\infty} \delta(x - \underline{x}) f(x) dx = \int_{x=-\infty}^{\infty} \delta(x - \underline{x}) f(\underline{x}) dx = f(\underline{x}) \quad (6.41)$$

(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:

$$R_{\underline{x}\underline{y}\underline{z}}(x, y, z) = \delta(x - \underline{x})\delta(y - \underline{y})\delta(z - \underline{z}) \equiv \delta^3(\vec{r} - \vec{\underline{r}}) \quad (6.42)$$

Here $\delta^3(\vec{r} - \vec{\underline{r}})$ is the three-dimensional delta function, a spike at position $\vec{\underline{r}}$ 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

$$c_{\underline{x}\underline{y}\underline{z}}(t) = \Psi(\underline{x}, \underline{y}, \underline{z}; t) \quad (6.43)$$

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 accuracy of an *infinite* number of digits would be zero. So, the properly normalized eigenfunctions would have been useless anyway.

Instead, according to Born's statistical interpretation of chapter 2.1, the expression

$$|\Psi(x, y, z; t)|^2 dx dy dz$$

gives the probability of finding the particle in an infinitesimal volume $dx dy dz$ 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.

6.4.2 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:

$$\frac{\hbar}{i} \frac{\partial X(x)}{\partial x} = p_x X(x)$$

The solution is a complex exponential:

$$X(x) = A e^{ip_x x / \hbar}$$

where A is a constant.

Just like the position eigenfunction earlier, the linear momentum eigenfunction has a normalization problem too. 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 non-quantum ones, are likely to make a different choice.)

The problems for the y and z -linear momentum have similar solutions, so the full eigenfunction for linear momentum takes the form:

$$P_{p_x p_y p_z}(x, y, z) = \frac{1}{\sqrt{2\pi\hbar^3}} e^{i(p_x x + p_y y + p_z z) / \hbar}$$

(6.44)

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:

$$\boxed{\Phi(p_x, p_y, p_z; t) = \frac{1}{\sqrt{2\pi\hbar^3}} \langle e^{i(p_x x + p_y y + p_z z)/\hbar} | \Psi \rangle} \quad (6.45)$$

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

$$|\Phi(p_x, p_y, p_z; t)|^2 dp_x dp_y dp_z$$

gives the probability of finding the linear momentum within a small momentum range $dp_x dp_y dp_z$ 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:

$$\boxed{\Psi(x, y, z; t) = \frac{1}{\sqrt{2\pi\hbar^3}} \langle e^{-i(p_x x + p_y y + p_z z)/\hbar} | \Phi \rangle_{\vec{p}}} \quad (6.46)$$

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:

$$\boxed{\Psi(x, y, z; t) = \frac{1}{\sqrt{2\pi\hbar^3}} \int_{\text{all } \vec{p}} \int \int \Phi(p_x, p_y, p_z; t) e^{i(p_x x + p_y y + p_z z)/\hbar} dp_x dp_y dp_z} \quad (6.47)$$

Mathematicians prove this formula under the name “Fourier Inversion Theorem”, {A.50}. 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 (6.44). The only real difference is that the sum has become an integral since \vec{p} has continuous values, not discrete ones.

- The linear momentum eigenfunctions are complex exponentials of the form:

$$\frac{1}{\sqrt{2\pi\hbar}^3} e^{i(p_x x + p_y y + p_z z)/\hbar}$$

- 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.
-

6.5 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:

$$v = \frac{p}{m} = \text{constant} \quad x = vt + x_0 \quad \text{for Newtonian motion in free space}$$

(Note that p is used to indicate p_x in this and the following sections.)

6.5.1 Solution of the Schrödinger equation.

As discussed in section 6.1, 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 $e^{-iEt/\hbar}$. 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:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} = E\psi \quad (6.48)$$

Solutions to this equation take the form of exponentials

$$\psi_E = A e^{\pm i\sqrt{2mE}x/\hbar}$$

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 \hat{p}_x derived in the previous section:

$$\psi_E = \frac{1}{\sqrt{2\pi\hbar}} e^{ipx/\hbar} \quad \text{with } p = \pm\sqrt{2mE} \quad (6.49)$$

The reason that the momentum eigenfunctions are also energy eigenfunctions is that the energy is all kinetic energy, and the kinetic operator equals $\hat{T} = \hat{p}^2/2m$. So eigenfunctions with precise momentum p have precise energy $p^2/2m$.

As shown by (6.47) 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:

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \Phi(p, t) e^{ipx/\hbar} dp$$

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 $e^{-iEt/\hbar}$. 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:

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \Phi_0(p) e^{ip(x - \frac{p}{2m}t)/\hbar} dp \quad (6.50)$$

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 (6.50).
- The function Φ_0 is still to be chosen to produce whatever physical situation is to be described.

6.5.2 Component wave solutions

Before trying to interpret the complete obtained solution (6.50) for the wave function of a particle in free space, it is instructive first to have a look at the component solutions, defined by

$$\psi_w \equiv e^{ip(x - \frac{p}{2m}t)/\hbar} \quad (6.51)$$

These solutions will be called component waves; both their real and imaginary parts are sinusoidal, as can be seen from the Euler formula (1.5).

$$\psi_w = \cos\left(p\left(x - \frac{p}{2m}t\right)/\hbar\right) + i \sin\left(p\left(x - \frac{p}{2m}t\right)/\hbar\right)$$

In figure 6.7, 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 the

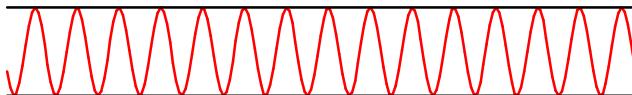


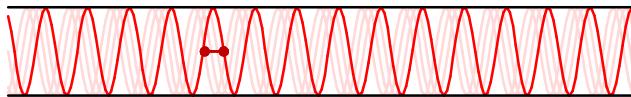
Figure 6.7: The real part (red) and envelope (black) of an example wave.

“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 6.8. This is unexpected, because



The html version of this document has an animation of the motion.

Figure 6.8: The wave moves with the phase speed.

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 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.

6.5.3 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 (6.50) 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 6.9. 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.

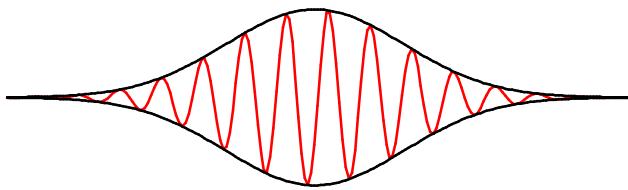


Figure 6.9: The real part (red) and magnitude or envelope (black) of a wave packet. (Schematic).

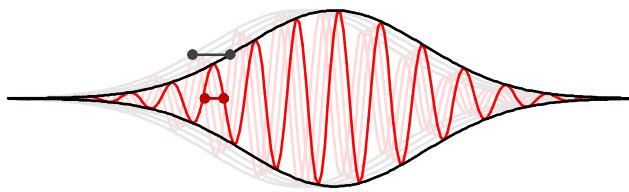
The vertical separation between the black lines is again a measure of the probability of finding the particle near that location. It is seen that the possible locations of the particle are now restricted to a finite region, the region in which the vertical distance between the black lines is nonzero.

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 6.1.8, for such a particle the uncertainty in position is negligible. The wave packet in which the particle can be found, as sketched in figure 6.9, 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 6.10 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.

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 6.1.8 with a speed $\langle p \rangle/m$, where $\langle p \rangle$ is the expectation value of momentum,



The html version of this document has an animation of the motion.

Figure 6.10: The velocities of wave and envelope are not equal.

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.
 - □ This region propagates with the classical particle velocity.
-

6.5.4 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:

component wave: $\psi_w = e^{i(kx - \omega t)}$

(6.52)

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.50}.)

For a particle in free space according to the previous subsection:

$$k = \frac{p}{\hbar} \quad \omega = \frac{p^2}{2m\hbar}$$

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

dispersion relation: $\omega = \omega(k)$

(6.53)

It really defines the physics of the wave propagation.

Since the waves are of the form $e^{ik(x - \frac{\omega}{k}t)}$, the wave is constant if $x = (\omega/k)t$ plus any constant. Such points move with the

phase velocity: $v_p \equiv \frac{\omega}{k}$

(6.54)

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

group velocity: $v_g \equiv \frac{d\omega}{dk}$

(6.55)

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

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{\Phi}_0(k) e^{i(kx - \omega t)} dk$$

(6.56)

Here $\bar{\Phi}_0$ is the complex amplitude of the waves. The combination $\bar{\Phi}_0 e^{-i\omega t}$ 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 $\bar{\Phi}_0(k)$ 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 note {A.51}. 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

$$v_{g1}t < x < v_{g2}t$$

(In case $v_{g1} > v_{g2}$, 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:

$$\Psi(x, t) \sim \frac{e^{\mp i\pi/4}}{\sqrt{|v'_{g0}|t}} \bar{\Phi}_0(k_0) e^{i(k_0 x - \omega_0 t)} \quad v_{g0} = \frac{x}{t}$$

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'_{g0} is the derivative of the group speed at that point, and \mp stands for the sign of $-v'_{g0}$.

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

$$d(k_0x - \omega_0t) = k_0dx - \omega_0dt + xdk_0 - t d\omega_0$$

and observe that the final two terms cancel because $d\omega_0/dk_0$ 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

$$\Psi(x, t) \sim e^{-i\pi/4} \sqrt{\frac{m}{t}} \Phi_0\left(\frac{mx}{t}\right) e^{imx^2/2\hbar t}$$

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 $e^{i(kx-\omega t)}$.
- 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 $d\omega/dk$. It describes how fast wave packets move.
- Relatively simple expressions exist for the wave function of wave packets at large times.

6.5.5 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

$$-\frac{\hbar^2}{2m_e} \frac{\partial^2 \psi}{\partial x^2} + V(x)\psi = E\psi \quad (6.57)$$

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 2.5.8. 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 (6.57) take the form of “Bloch waves.”

$$\psi_k^P(x) = \psi_{p,k}^P(x)e^{ikx} \quad (6.58)$$

where $\psi_{p,k}^P$ 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 3.4.1. 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^{ikd} for *some* real value k . And that means that if you write the eigenfunction in the Bloch form (6.58), then the exponential will produce the eigenvalue during a shift. So the part $\psi_{p,k}^P$ 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 $\psi_{p,k}^P$ 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 5.22, 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 $\psi_{p,k}^P$ 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:

$$\boxed{p_{\text{cm}} = \hbar k} \quad (6.59)$$

(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, {A.52}:

$$\boxed{v = \frac{dE^P}{dp_{\text{cm}}} \quad p_{\text{cm}} = \hbar k} \quad (6.60)$$

(In three dimensions, replace the p -derivative by $1/\hbar$ times the gradient with respect to \vec{k} .) In free space, $E^P = \hbar\omega$ and $p_{\text{cm}} = \hbar k$, 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, {A.52}:

$$\boxed{\frac{dp_{\text{cm}}}{dt} = F_{\text{ext}} \quad p_{\text{cm}} = \hbar k} \quad (6.61)$$

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,

$$\frac{dv}{dt} = \frac{d^2 E^p}{dp_{\text{cm}}^2} \frac{dp_{\text{cm}}}{dt} = \frac{d^2 E^p}{dp_{\text{cm}}^2} F_{\text{ext}}$$

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:

$$m_{\text{eff}} \equiv 1 / \frac{d^2 E^{\text{p}}}{dp_{\text{cm}}^2} \quad p_{\text{cm}} = \hbar k \quad (6.62)$$

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

$$E^{\text{p}} = E_{\text{min}}^{\text{p}} + \frac{1}{2} \frac{\partial^2 E^{\text{p}}}{\partial k_x^2} k_x^2 + \frac{1}{2} \frac{\partial^2 E^{\text{p}}}{\partial k_y^2} k_y^2 + \frac{1}{2} \frac{\partial^2 E^{\text{p}}}{\partial k_z^2} k_z^2 + \dots$$

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

$$E^{\text{p}} = E_{\text{min}}^{\text{p}} + \frac{1}{2} \frac{1}{m_{\text{eff},x}} p_{\text{cm},x}^2 + \frac{1}{2} \frac{1}{m_{\text{eff},y}} p_{\text{cm},y}^2 + \frac{1}{2} \frac{1}{m_{\text{eff},z}} p_{\text{cm},z}^2 + \dots \quad (6.63)$$

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.
-

6.6 Almost Classical Motion [Descriptive]

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 6.1.8, 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 note {A.53}, if you want to understand them better, or create some yourself.

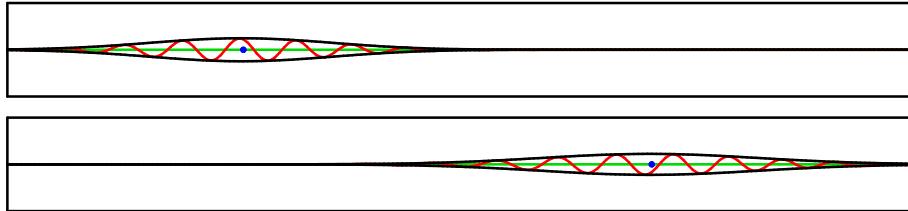
6.6.1 Motion through free space

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 6.11 shows it at two different times. 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.

6.6.2 Accelerated motion

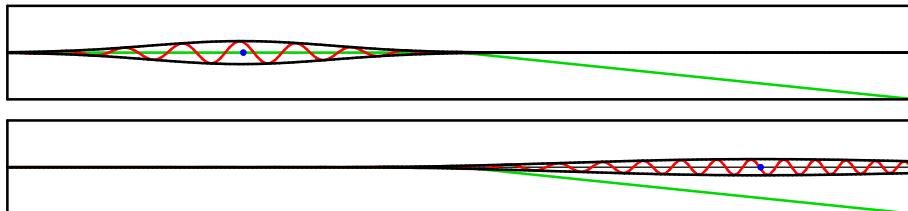
Figure 6.12 shows the motion when the potential energy (shown in green) ramps down starting from the middle of the plotted range. Physically this corresponds



The html version of this document has an animation of the motion to show that it is indeed at constant speed.

Figure 6.11: A particle in free space.

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.



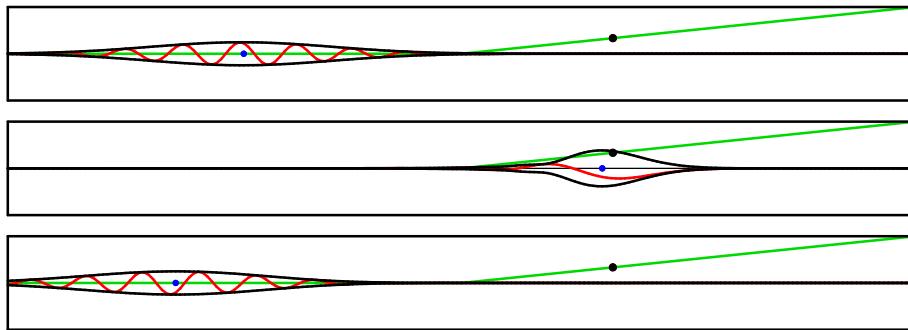
The html version of this document has an animation of the motion.

Figure 6.12: An accelerating particle.

6.6.3 Decelerated motion

Figure 6.13 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.



The html version of this document has an animation of the motion.

Figure 6.13: An decelerating particle.

6.6.4 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 2.6, but only now, near the end of part I, can the classical picture of a particle oscillating back and forward actually be created.

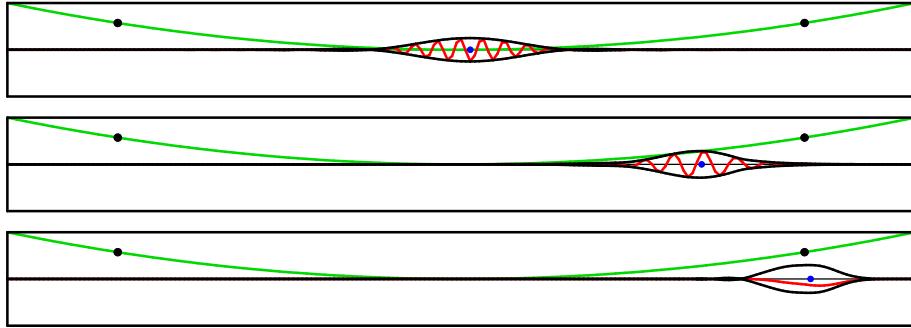
There are some mathematical differences from the previous cases, because the energy levels of the harmonic oscillator are discrete, unlike those of the particles that are able to escape. But if the energy levels are far enough above the ground state, localized wave packets similar to the ones in free space may be formed, {A.53}. The animation in figure 6.14 gives the motion of a wave packet whose nominal energy is hundred times the ground state energy.

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.



The html version of this document has an animation of the motion.

Figure 6.14: Unsteady solution for the harmonic oscillator. The third picture shows the maximum distance from the nominal position that the wave packet reaches.

6.7 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 the previous section. 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.

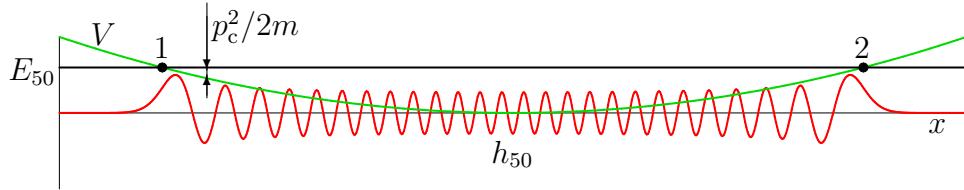


Figure 6.15: Harmonic oscillator potential energy V , eigenfunction h_{50} , and its energy E_{50} .

The WKB approximation is based on the rapid spatial variation of energy eigenfunctions with almost macroscopic energies. As an example, figure 6.15 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

$$p_c \equiv \sqrt{2m(E - V)} \quad (6.64)$$

In quantum mechanics, the large momentum implies the rapid oscillation of the wave function since quantum mechanics associates the linear momentum with the operator $\hbar d/dx$ 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

$$-\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V\psi = E\psi$$

take the $V\psi$ term to the other side and then rewrite $E - V$ in terms of the classical linear momentum. That produces

$$\frac{d^2\psi}{dx^2} = -\frac{p_c^2}{\hbar^2}\psi \quad (6.65)$$

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,

$$\psi \sim c_f e^{ip_c x/\hbar} + c_b e^{-ip_c x/\hbar}$$

where the constants c_f and c_b are arbitrary. (The subscripts denote whether the wave speed of the corresponding term is forward or backward.) It turns out, {A.54}, that to make the above expression work over more than one period, it is necessary to replace $p_c x$ by the anti-derivative $\int p_c dx$; furthermore, the “constants” c_f and c_b must be allowed to vary from period to period proportional to $1/\sqrt{p_c}$.

In short, the WKB approximation of the wave function is:

$$\text{classical WKB: } \psi \approx \frac{1}{\sqrt{p_c}} [C_f e^{i\theta} + C_b e^{-i\theta}] \quad \theta \equiv \frac{1}{\hbar} \int p_c dx \quad (6.66)$$

where C_f and C_b are now true constants.

If you ever glanced at notes such as {A.12}, {A.15}, and {A.17}, 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 (1.5). It produces

$$\boxed{\text{rephrased WKB: } \psi \approx \frac{1}{\sqrt{p_c}} [C_c \cos \theta + C_s \sin \theta] \quad \theta \equiv \frac{1}{\hbar} \int p_c dx} \quad (6.67)$$

The constants C_c and C_s are related to the original constants C_f and C_b as

$$\boxed{C_c = C_f + C_b \quad C_s = iC_f - iC_b \quad C_f = \frac{1}{2}(C_c - iC_s) \quad C_b = \frac{1}{2}(C_c + iC_s)} \quad (6.68)$$

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 2.5. 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_c must be zero for ψ to be zero at x_1 . 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

$$\boxed{\text{particle between impenetrable walls: } \frac{1}{\hbar} \int_{\underline{x}=x_1}^{x_2} p_c(\underline{x}) d\underline{x} = n\pi} \quad (6.69)$$

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 6.15. 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.55}. 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:

$$\boxed{\text{particle trapped between turning points: } \frac{1}{\hbar} \int_{\underline{x}=x_1}^{x_2} p_c(x) dx = (n - \frac{1}{2})\pi} \quad (6.70)$$

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 6.15 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:

$$\boxed{\text{tunneling WKB: } \psi \approx \frac{1}{\sqrt{|p_c|}} [C_p e^\gamma + C_n e^{-\gamma}] \quad \gamma \equiv \frac{1}{\hbar} \int |p_c| dx} \quad (6.71)$$

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 $\theta = \int p_c dx / \hbar$.
- The wave function can be written as (6.66) or (6.67), whatever is more convenient.
- For a particle stuck between impenetrable walls, the energy eigenvalues can be found from (6.69).

- For a particle stuck between a gradually increasing potential at both sides, the energy eigenvalues can be found from (6.70).
 - The “tunneling” wave function in regions that classically the particle is forbidden to enter can be approximated as (6.71). It is in terms of the antiderivative $\gamma = \int |p_c| dx / \hbar$.
-

6.7 Review Questions

1 Use the equation

$$\frac{1}{\hbar} \int_{\underline{x}=x_1}^{x_2} p_c(\underline{x}) d\underline{x} = n\pi$$

to find the WKB approximation for the energy levels of a particle stuck in a pipe of chapter 2.5.5. The potential V is zero inside the pipe, given by $0 \leq x \leq \ell_x$

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.

2 Use the equation

$$\frac{1}{\hbar} \int_{\underline{x}=x_1}^{x_2} p_c(\underline{x}) d\underline{x} = (n - \frac{1}{2})\pi$$

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 $E = \frac{1}{2}m\omega x_2^2$.

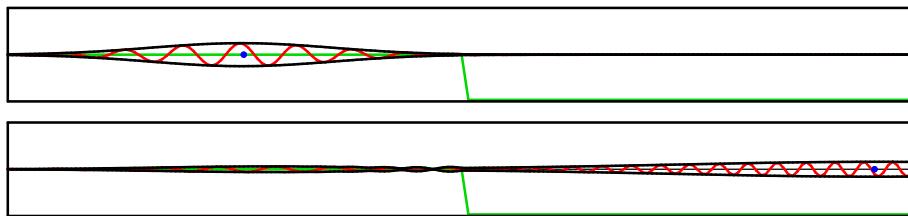
In this case too, the WKB approximation produces the exact energy eigenvalues. That, however, is just a coincidence; the classical WKB wave functions are certainly not exact; they become infinite at the turning points. As the example h_{50} above shows, the true wave functions most definitely do not.

6.8 Scattering

The motion of the wave packets in section 6.6 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.

6.8.1 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 6.16; 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.



The html version of this document has an animation of the motion.

Figure 6.16: 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 2.1 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.

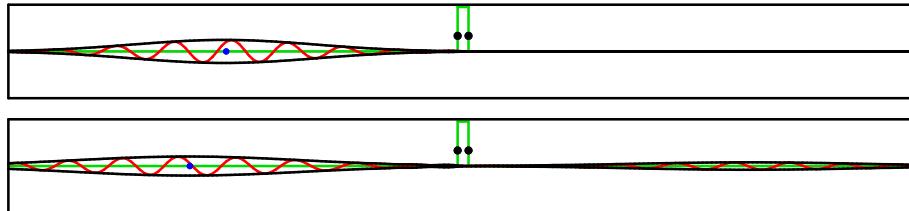
Partial reflections are the norm for potentials that vary nontrivially on quantum scales, but this example adds a second twist. Classically, a *decelerating* force is needed to turn a particle back, but here the force is everywhere accelerating only! As an actual physical example of this weird behavior, neutrons trying to enter nuclei experience attractive forces that come on so quickly that they may be repelled by them.

6.8.2 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 6.17 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.

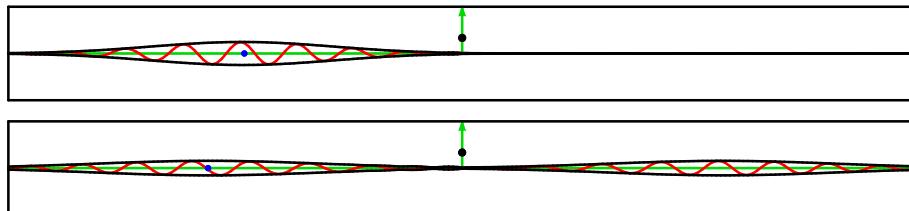


The html version of this document has an animation of the motion.

Figure 6.17: 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 6.18; 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 6.18, 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.



The html version of this document has an animation of the motion.

Figure 6.18: 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

- o— If the potential varies nontrivially on quantum scales, wave packets do not move like classical particles.
- o— A wave packet may split into separate parts that move in different ways.
- o— A wave packet may be reflected by an accelerating force.
- o— A wave packet may tunnel through regions that a classical particle could not enter.

6.9 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.

Consider an arbitrary scattering potential like the one in figure 6.19. 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

$$\psi_E = C_f e^{ip_c x/\hbar} + C_b e^{-ip_c x/\hbar}$$

where $p_c = \sqrt{2m(E - V)}$ is the classical momentum and C_f and C_b are constants. When eigenfunctions of slightly different energies are combined together,

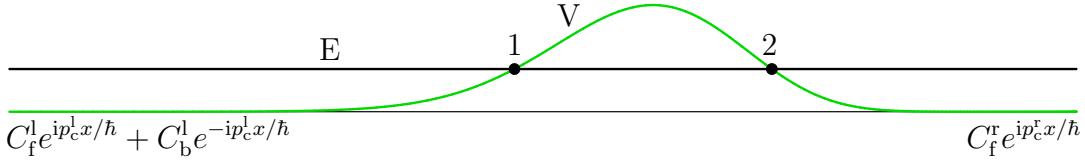


Figure 6.19: Schematic of a scattering potential and the asymptotic behavior of an example energy eigenfunction for a wave packet coming in from the far left.

the terms $C_f e^{ip_c x/\hbar}$ produce wave packets that move forwards in x , graphically from left to right, and the terms $C_b e^{-ip_c x/\hbar}$ 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_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_b is normally not zero. In fact, the term $C_b^l e^{-ip_c x/\hbar}$ 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.57},

$$R = \frac{|C_b^l|^2}{|C_f^l|^2} \quad (6.72)$$

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

$$T = \frac{p_c^r |C_f^r|^2}{p_c^l |C_f^l|^2} \quad p_c^l = \sqrt{2m(E - V_l)} \quad p_c^r = \sqrt{2m(E - V_r)} \quad (6.73)$$

where p_c^l respectively p_c^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.55}. It is

$$\boxed{T \approx e^{-2\gamma_{12}} \quad \gamma_{12} = \frac{1}{\hbar} \int_{x_1}^{x_2} |p_c| dx \quad |p_c| = \sqrt{2m(V - E)}} \quad (6.74)$$

where x_1 and x_2 are the “turning points” in figure 6.19, 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 note {A.56}.

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.

Part II

Gateway Topics

Chapter 7

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.

7.1 The Variational Method

Solving the equations of quantum mechanics is typically difficult, so approximations must usually be made. One very effective way to find an approximate ground state is the variational principle. This section gives some of the basic ideas, including ways to apply it best, and how to find eigenstates of higher energy in similar ways.

7.1.1 Basic variational statement

The variational method is based on the observation that the ground state is the state among all allowable wave functions that has the lowest expectation value of energy:

$$\langle E \rangle \text{ is minimal for the ground state wave function.} \quad (7.1)$$

The variational method has already been used to find the ground states for the hydrogen molecular ion, chapter 3.5, and the hydrogen molecule, chapter 4.2. The general procedure is to guess an approximate form of the wave function, invariably involving some parameters whose best values you are unsure about. Then search for the parameters that give you the lowest expectation value of the total energy; those parameters will give your best possible approximation to the true ground state {A.22}. In particular, you can be confident that the true ground state energy is no higher than what you compute, {A.24}.

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.

If you want to get more accurate values, you will need to increase the number of parameters. The molecular example solutions were based on the atom 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/or coefficients of the functions, or the coefficients of the polynomials.

7.1.2 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 could search forever. Maybe you could try some clever genetic algorithm.

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 partial 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. 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. Instead a typical physicist would say that the change δf , (or df), in f due to a small change δx in x must be zero. It is the same thing, since $\delta f = f' \delta x$, so that if f' is zero, then so is δf . But mathematicians do not like the word small, since it has no rigorous meaning. On the other hand, in physics you may not like to talk about derivatives, 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, and there is often more than one possible choice for x , with none preferred under all circumstances. (And in

practice, the word “small” does have an unambiguous meaning: it means that you must ignore everything that is of square magnitude or more in terms of the “small” quantities.)

In physics terms, the fact that the expectation energy must be minimal in the ground state means that you must have:

$$\delta\langle E \rangle = 0 \text{ for all acceptable small changes in wave function} \quad (7.2)$$

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.

7.1.3 Example application using Lagrangian multipliers

As an example of how you can apply the variational formulation of the previous subsection 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 you 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 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 penalty factor is carefully tuned, the second term can cancel any erroneous gain or decrease in expectation energy due to going out of bounds, {A.58}.

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 *unchanged* 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

$$(\langle \psi + \delta\psi | H | \psi + \delta\psi \rangle - \langle \psi | H | \psi \rangle) - \epsilon (\langle \psi + \delta\psi | \psi + \delta\psi \rangle - \langle \psi | \psi \rangle) = 0$$

Multiplying out, canceling equal terms and ignoring terms that are quadratically small in $\delta\psi$, you get

$$\langle \delta\psi | H | \psi \rangle + \langle \psi | H | \delta\psi \rangle - \epsilon (\langle \delta\psi | \psi \rangle + \langle \psi | \delta\psi \rangle) = 0$$

That is not yet good enough to say something specific about. But remember that you can exchange the sides of an inner product if you add a complex conjugate, so

$$\langle \delta\psi | H | \psi \rangle + \langle \delta\psi | H | \psi \rangle^* - \epsilon (\langle \delta\psi | \psi \rangle - \langle \delta\psi | \psi \rangle^*) = 0$$

Also remember that you can allow any change $\delta\psi$ you want, including the $\delta\psi$ you are now looking at times i. That means that you also have:

$$\langle i\delta\psi | H | \psi \rangle + \langle i\delta\psi | H | \psi \rangle^* - \epsilon (\langle i\delta\psi | \psi \rangle + \langle i\delta\psi | \psi \rangle^*) = 0$$

or using the fact that numbers come out of the left side of an inner product as complex conjugates

$$-i\langle \delta\psi | H | \psi \rangle + i\langle \delta\psi | H | \psi \rangle^* - \epsilon (-i\langle \delta\psi | \psi \rangle + i\langle \psi | \delta\psi \rangle^*) = 0$$

If you divide out a $-i$ and then average with the original equation, you get rid of the complex conjugates:

$$\langle \delta\psi | H | \psi \rangle - \epsilon \langle \delta\psi | \psi \rangle = 0$$

You can now combine them 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, {A.59}.

7.2 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.

7.2.1 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^n . Roughly speaking, the mass m_j^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_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:

$$\hat{T}^E = - \sum_{i=1}^I \frac{\hbar^2}{2m_e} \nabla_i^2 = - \sum_{i=1}^I \frac{\hbar^2}{2m_e} \left(\frac{\partial^2}{\partial r_{1i}^2} + \frac{\partial^2}{\partial r_{2i}^2} + \frac{\partial^2}{\partial r_{3i}^2} \right). \quad (7.3)$$

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,

$$\hat{T}^N = - \sum_{j=1}^J \frac{\hbar^2}{2m_j^n} \nabla_j^{n2} = - \sum_{j=1}^J \frac{\hbar^2}{2m_j^n} \left(\frac{\partial^2}{\partial r_{1j}^n} + \frac{\partial^2}{\partial r_{2j}^n} + \frac{\partial^2}{\partial r_{3j}^n} \right). \quad (7.4)$$

where $\vec{r}_j^n = (r_{1j}^n, r_{2j}^n, r_{3j}^n)$ 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:

$$V^{\text{NE}} = - \sum_{i=1}^I \sum_{j=1}^J \frac{Z_j e^2}{4\pi\epsilon_0} \frac{1}{r_{ij}} \quad (7.5)$$

where $r_{ij} \equiv |\vec{r}_i - \vec{r}_j^{\text{n}}|$ is the distance between electron number i and nucleus number j , and $\epsilon_0 = 8.85 \cdot 10^{-12} \text{ C}^2/\text{J m}$ is the permittivity of space.

Next there is the potential energy due to the electron-electron repulsions:

$$V^{\text{EE}} = \frac{1}{2} \sum_{i=1}^I \sum_{\substack{j=1 \\ j \neq i}}^I \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}} \quad (7.6)$$

where $r_{ii} \equiv |\vec{r}_i - \vec{r}_{\underline{i}}|$ 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,

$$V^{\text{NN}} = \frac{1}{2} \sum_{j=1}^J \sum_{\substack{j'=1 \\ j' \neq j}}^J \frac{Z_j Z_{\underline{j}} e^2}{4\pi\epsilon_0} \frac{1}{r_{jj'}} \quad (7.7)$$

where $r_{jj'} \equiv |\vec{r}_j^{\text{n}} - \vec{r}_{\underline{j}}^{\text{n}}|$ 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

$$[\hat{T}^{\text{E}} + \hat{T}^{\text{N}} + V^{\text{NE}} + V^{\text{EE}} + V^{\text{NN}}] \psi = E\psi \quad (7.8)$$

Here ψ is a function of the position and spin coordinates of all the electrons and all the nuclei, in other words:

$$\psi = \psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_I, S_{zI}, \vec{r}_1^{\text{n}}, S_{z1}^{\text{n}}, \vec{r}_2^{\text{n}}, S_{z2}^{\text{n}}, \dots, \vec{r}_J^{\text{n}}, S_{zJ}^{\text{n}}) \quad (7.9)$$

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.

7.2.2 The 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,

$$\hat{T}^E = - \sum_{i=1}^I \frac{\hbar^2}{2m_e} \left(\frac{\partial^2}{\partial r_{1i}^2} + \frac{\partial^2}{\partial r_{2i}^2} + \frac{\partial^2}{\partial r_{3i}^2} \right)$$

$$\hat{T}^N = - \sum_{j=1}^J \frac{\hbar^2}{2m_j^n} \left(\frac{\partial^2}{\partial r_{1j}^n} + \frac{\partial^2}{\partial r_{2j}^n} + \frac{\partial^2}{\partial r_{3j}^n} \right)$$

then what would seem more reasonable than to ignore the kinetic energy \hat{T}^N of the nuclei? It has those heavy masses in the bottom.

An alternative, and better, way of phrasing the assumption that \hat{T}^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 4.2. 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 ψ^E to distinguish them from the true solutions ψ that do not use the Born-Oppenheimer approximation. Mathematically ψ^E will still be a function of the electron and nuclear positions:

$$\psi^E = \psi^E(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_I, S_{zI}; \vec{r}_1^n, S_{z1}^n, \vec{r}_2^n, S_{z2}^n, \dots, \vec{r}_J^n, S_{zJ}^n). \quad (7.10)$$

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 ψ^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

$$[\hat{T}^E + V^{NE} + V^{EE} + V^{NN}] \psi^E = (E^E + V^{NN}) \psi^E,$$

which just means that the eigenvalue is called $E^E + V^{NN}$ instead of simply E^E . The reason is that you can then get rid of V^{NN} , and obtain the electron wave function eigenvalue problem in the more concise form

$$\boxed{[\hat{T}^E + V^{NE} + V^{EE}] \psi^E = E^E \psi^E} \quad (7.11)$$

After all, for given nuclear coordinates, V^{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_{gs}^E , the lowest eigenvalue of the electronic problem (7.11) 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:

$$\boxed{\text{nominal ground state condition: } E_{gs}^E + V^{NN} \text{ is minimal}} \quad (7.12)$$

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 $E_{gs}^E + V^{NN}$ 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 (7.12) above does have big limitations, so the next subsection discusses a more advanced approach.

7.2.3 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 certain 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 cooperative 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

$$\left[\hat{T}^N + V^{NN} + E_1^E \right] \psi_1^N = E \psi_1^N$$

where ψ_1^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^E .

So, what is this E_1^E ? Easy, it is the electronic ground state energy E_{gs}^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 compute E_1^E for a sufficient number of relevant nuclear positions, tabulate the results somehow, and interpolate them as needed. E_1^E is then a known function function of the nuclear positions and so is V^{NN} . Proceed to solve for the wave function for the nuclei ψ_1^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 $V^{\text{NN}} + E_1^E$ 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^E , (for simplicity, it will be assumed that the electron energy levels are numbered with a single index n ,) just solve

$$\boxed{[\hat{T}^N + V^{\text{NN}} + E_n^E] \psi_n^N = E \psi_n^N} \quad (7.13)$$

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_n^E and $E_{\bar{n}}^E$, 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 {A.60}. There is no need to worry, though; these reversed-spin solutions go their own way and do not affect the validity of (7.13). It is spatial, rather than spin nonuniqueness that is a concern.

There is a derivation of the nuclear eigenvalue problem (7.13) in note {A.61}, showing what the ignored terms are and why they can usually be ignored.

7.3 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 7.2 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.

7.3.1 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:

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \dots, \vec{r}_i, S_{zi}, \dots, \vec{r}_I, S_{zI})$$

where \vec{r}_i is the position of electron number i , and S_{zi} its spin in a chosen z -direction, with measurable values $\frac{1}{2}\hbar$ and $-\frac{1}{2}\hbar$. Of course, what answer you get for the wave function will also depend on where the nuclei are, but in this section, the nuclei are supposed to be at given positions, so to reduce the clutter, the dependence of the electron wave function on the nuclear positions will not be explicitly shown.

Hartree-Fock approximates the wave function in terms of a set of *single-electron* functions, each a product of a spatial function and a spin state:

$$\psi_1^s(\vec{r})\uparrow_1(S_z), \psi_2^s(\vec{r})\uparrow_2(S_z), \psi_3^s(\vec{r})\uparrow_3(S_z), \dots$$

where \uparrow stands for either spin-up, \uparrow , or spin-down, \downarrow . (By definition, function $\uparrow(S_z)$ equals one if the spin S_z is $\frac{1}{2}\hbar$, and zero if it is $-\frac{1}{2}\hbar$, while function $\downarrow(S_z)$ equals zero if S_z is $\frac{1}{2}\hbar$ and one if it is $-\frac{1}{2}\hbar$.) These single-electron functions are called “orbitals” or “spin orbitals.” The reason is that you tend to think of them as describing a single electron being in orbit around the nuclei with a particular spin. Wrong, of course: the electrons do not have reasonably defined positions on these scales. But you do tend to think of them that way anyway.

The spin orbitals are taken to be an orthonormal set. Note that any two spin orbitals are automatically orthogonal if they have opposite spins: spin states are orthonormal so $\langle \uparrow | \downarrow \rangle = 0$. If they have the same spin, their spatial orbitals will need to be orthogonal.

Single-electron functions can be combined into multi-electron functions by forming products of them of the form

$$a_{n_1, n_2, \dots, n_I} \psi_{n_1}^s(\vec{r}_1)\uparrow_{n_1}(S_{z1}) \psi_{n_2}^s(\vec{r}_2)\uparrow_{n_2}(S_{z2}) \dots \psi_{n_I}^s(\vec{r}_I)\uparrow_{n_I}(S_{zI})$$

where n_1 is the number of the single-electron function used for electron 1, n_2 the number of the single-electron function used for electron 2, and so on, and a_{n_1, n_2, \dots, n_I} is a suitable numerical constant. Such a product wave function is called a “Hartree product.”

Now if you use enough single-electron functions, with all their Hartree products, you can approximate any multi-electron wave function to arbitrarily high accuracy. Unfortunately, using many of them produces a problem much too big to be solved on even the most powerful computer. So you really want to use as little of them as possible. But you cannot use too few either; as chapter 4.7 explained, nature imposes an “antisymmetrization” requirement: the complete wave function that you write must change sign whenever any two electrons are exchanged, in other words when you replace \vec{r}_i, S_{zi} by $\vec{r}_{\underline{i}}, S_{\underline{zi}}$ and vice-versa for

any pair of electrons numbered i and \underline{i} . That is only possible if you use at least I different single-electron functions for your I electrons. This is known as the Pauli exclusion principle: any group of $I - 1$ electrons occupying the minimum of $I - 1$ single-electron functions “exclude” an additional I -th electron from simply entering the same functions. The I -th electron will have to find its own single-electron function to add to the mix.

The basic Hartree-Fock approximation uses the absolute minimum that is possible, just I different single-electron functions for the I electrons. In that case, the wave function Ψ can be written as a single “Slater determinant:”

$$\frac{a}{\sqrt{I!}} \begin{vmatrix} \psi_1^s(\vec{r}_1) \downarrow_1(S_{z1}) & \psi_2^s(\vec{r}_1) \downarrow_2(S_{z1}) & \dots & \psi_n^s(\vec{r}_1) \downarrow_n(S_{z1}) & \dots & \psi_I^s(\vec{r}_1) \downarrow_I(S_{z1}) \\ \psi_1^s(\vec{r}_2) \downarrow_1(S_{z2}) & \psi_2^s(\vec{r}_2) \downarrow_2(S_{z2}) & \dots & \psi_n^s(\vec{r}_2) \downarrow_n(S_{z2}) & \dots & \psi_I^s(\vec{r}_2) \downarrow_I(S_{z2}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \psi_1^s(\vec{r}_i) \downarrow_1(S_{zi}) & \psi_2^s(\vec{r}_i) \downarrow_2(S_{zi}) & \dots & \psi_n^s(\vec{r}_i) \downarrow_n(S_{zi}) & \dots & \psi_I^s(\vec{r}_i) \downarrow_I(S_{zi}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \psi_1^s(\vec{r}_I) \downarrow_1(S_{zI}) & \psi_2^s(\vec{r}_I) \downarrow_2(S_{zI}) & \dots & \psi_n^s(\vec{r}_I) \downarrow_n(S_{zI}) & \dots & \psi_I^s(\vec{r}_I) \downarrow_I(S_{zI}) \end{vmatrix} \quad (7.14)$$

where a is a constant of magnitude one. As chapter 4.7 explained, a Slater determinant is really equivalent to a sum of $I!$ Hartree products, each with the single-electron functions in a different order. It is the one wave function obtainable from the I single-electron functions that is antisymmetric with respect to exchanging any two of the I electrons.

Displaying the Slater determinant fully as above may look impressive, but it is a lot to read. Therefore, from now on it will be abbreviated as

$$\Psi = \frac{a}{\sqrt{I!}} \left| \det(\psi_1^s \downarrow_1, \psi_2^s \downarrow_2, \dots, \psi_n^s \downarrow_n, \dots, \psi_I^s \downarrow_I) \right\rangle. \quad (7.15)$$

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 {A.62}. 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 $\psi_1^s \downarrow_1, \psi_2^s \downarrow_2, \dots$ at random. The cleverness in the Hartree-Fock approach will be in writing down equations for these single-electron functions that produce the best approximation possible with a single Slater determinant.

This section will reserve the term “orbitals” specifically for the single-electron functions that provide the best single-determinant approximation. In those terms, if the Hartree-Fock orbitals provide the best single-determinant approximation, their results will certainly be better than the solutions that were written

down for the atoms in chapter 4.9, because those were really single Slater determinants. In fact, you could find much more accurate ways to average out the effects of the neighboring electrons than just putting them in the nucleus like the section on atoms essentially did. You could smear them out over some optimal area, say. But the solution you will get doing so will be no better than you could get using Hartree-Fock.

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. Spin then really only directly affects the antisymmetrization requirements.

Things are straightforward if you try to solve, say, a helium atom. In the exact ground state, the two electrons are in the spatial wave function that has the absolutely lowest energy, regardless of any antisymmetrization concerns. This spatial wave function is symmetric under electron exchange since the two electrons are identical. The antisymmetrization requirement is met since the electrons assume the singlet configuration,

$$\frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}},$$

for their combined spins.

The approximate Hartree-Fock wave function for helium you would correspondingly take to be

$$\frac{1}{\sqrt{2}}|\det(\psi_1^s\uparrow, \psi_2^s\downarrow)\rangle$$

and then you would make things easier for yourself by postulating a priori that the spatial orbitals are the same, $\psi_1^s = \psi_2^s$. Lo and behold, when you multiply out the Slater determinant,

$$\frac{1}{\sqrt{2}} \begin{vmatrix} \psi_1^s(\vec{r}_1)\uparrow(S_{z1}) & \psi_1^s(\vec{r}_1)\downarrow(S_{z1}) \\ \psi_1^s(\vec{r}_2)\uparrow(S_{z2}) & \psi_1^s(\vec{r}_2)\downarrow(S_{z2}) \end{vmatrix} = \psi_1^s(\vec{r}_1)\psi_1^s(\vec{r}_2) \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}},$$

it automagically reproduces the correct singlet spin state. And you only need to find one spatial orbital instead of two.

As discussed in chapter 4.9, 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 $|\det(\psi_1^s\uparrow, \psi_1^s\downarrow, \psi_3^s\uparrow, \psi_3^s\downarrow)\rangle/\sqrt{4!}$, 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 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 $\psi_3^s\uparrow$ or $\psi_3^s\downarrow$. Lets assume you take $\psi_3^s\uparrow$, so the wave function is

$$\frac{1}{\sqrt{3!}} |\det(\psi_1^s\uparrow, \psi_2^s\downarrow, \psi_3^s\uparrow)\rangle$$

You have introduced a bias in the determinant: there is now a real difference between $\psi_1^s\uparrow$ and $\psi_2^s\downarrow$: $\psi_1^s\uparrow$ has the same spin as the third spin orbital, and $\psi_2^s\downarrow$ opposite.

If you find the best approximation among *all* possible orbitals $\psi_1^s\uparrow$, $\psi_2^s\downarrow$, and $\psi_3^s\uparrow$, you will end up with spatial orbitals ψ_1^s and ψ_2^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. In the case of lithium, you would then demand that ψ_2^s equals ψ_1^s . 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 $\psi_3^s\uparrow$ 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 $2p_x$, $2p_y$, and $2p_z$ 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.

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 $\psi_1^s \uparrow$ and $\psi_2^s \downarrow$ with unequal spatial orbitals multiplies out to

$$\frac{1}{\sqrt{2}} |\det(\psi_1^s \uparrow, \psi_2^s \downarrow)\rangle = \frac{\psi_1^s(\vec{r}_1)\psi_2^s(\vec{r}_2) \uparrow(S_{z1})\downarrow(S_{z2}) - \psi_2^s(\vec{r}_1)\psi_1^s(\vec{r}_2) \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

or, writing the spin combinations in terms of singlets and triplets,

$$\begin{aligned} & \frac{\psi_1^s(\vec{r}_1)\psi_2^s(\vec{r}_2) + \psi_2^s(\vec{r}_1)\psi_1^s(\vec{r}_2)}{2} \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}} + \\ & \frac{\psi_1^s(\vec{r}_1)\psi_2^s(\vec{r}_2) - \psi_2^s(\vec{r}_1)\psi_1^s(\vec{r}_2)}{2} \frac{\uparrow(S_{z1})\downarrow(S_{z2}) + \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}} \end{aligned}$$

So, the spin will be some combination of zero spin (the singlet) and spin one (the triplet), and the combination will be different at different locations of the electrons to boot. However, it may be noted that unrestricted wave functions are commonly used as first approximations of doublet and triplet states anyway [32, p. 105].

To show that all this can make a real difference, take the example of the hydrogen molecule, chapter 4.2, when the two nuclei are far apart. The correct electronic ground state is

$$\frac{\psi_l(\vec{r}_1)\psi_r(\vec{r}_2) + \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)}{\sqrt{2}} \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

where $\psi_l(\vec{r}_1)\psi_r(\vec{r}_2)$ is the state in which electron 1 is around the left proton and electron 2 around the right one, and $\psi_r(\vec{r}_1)\psi_l(\vec{r}_2)$ is the same state but with the electrons reversed. Note that the spin state is a singlet one with zero net spin.

Now try to approximate it with a restricted closed shell Hartree-Fock wave function of the form $|\det(\psi_1^s \uparrow, \psi_1^s \downarrow)\rangle / \sqrt{2}$. The determinant multiplies out to

$$\psi_1^s(\vec{r}_1)\psi_1^s(\vec{r}_2) \frac{\uparrow(S_{z1})\downarrow(S_{z2}) - \downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}$$

Now ψ_1^s will be something like $(\psi_l + \psi_r)/\sqrt{2}$; the energy of the electrons is lowest when they are near the nuclei. But if $\psi_1^s(\vec{r}_1)$ is appreciable when electron 1 is near say the left nucleus, then $\psi_1^s(\vec{r}_2)$ is also appreciable when electron 2 is near the *same* nucleus, since it is the exact same function. So there is a big chance of finding both electrons together near the same nucleus. That is all wrong, since the electrons repel each other: if one electron is around the left nucleus,

the other should be around the right one. The computed energy, which should be that of two neutral hydrogen atoms far apart, will be much too high. Note however that you do get the correct spin state. Also, at the nuclear separation distance corresponding to the ground state of the complete molecule, the errors are much less, [32, p. 166]. Only when you are “breaking the bond” (dissociating the molecule, i.e. taking the nuclei apart) do you get into major trouble.

If instead you would use unrestricted Hartree-Fock, $|\det(\psi_1^s\uparrow, \psi_2^s\downarrow)\rangle/\sqrt{2}$, you should find $\psi_1^s = \psi_l$ and $\psi_2^s = \psi_r$ (or vice versa), which would produce a wave function

$$\frac{\psi_l(\vec{r}_1)\psi_r(\vec{r}_2)\uparrow(S_{z1})\downarrow(S_{z2}) - \psi_r(\vec{r}_1)\psi_l(\vec{r}_2)\downarrow(S_{z1})\uparrow(S_{z2})}{\sqrt{2}}.$$

This would produce the correct energy, though the spin would now be all wrong. Little in life is ideal, is it?

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 $\psi_1^s, \psi_2^s, \dots$ 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.

7.3.2 The Hamiltonian

The non-relativistic 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:

$$\hat{T}^E = -\sum_{i=1}^I \frac{\hbar^2}{2m_e} \nabla_i^2 = -\sum_{i=1}^I \frac{\hbar^2}{2m_e} \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \right). \quad (7.16)$$

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:

$$V^{NE} = -\sum_{i=1}^I \sum_{j=1}^J \frac{Z_j e^2}{4\pi\epsilon_0 r_{ij}} \quad (7.17)$$

where $r_{ij} \equiv |\vec{r}_i - \vec{r}_j^n|$ is the distance between electron number i and nucleus number j , and $\epsilon_0 = 8.85 \cdot 10^{-12} \text{ C}^2/\text{J m}$ 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

$$V^{\text{EE}} = \frac{1}{2} \sum_{i=1}^I \sum_{\substack{i=1 \\ i \neq i}}^I \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}} \quad (7.18)$$

where $r_{ii} \equiv |\vec{r}_i - \vec{r}_{\underline{i}}|$ is the distance between electron number i and electron number \underline{i} . Half of this repulsion energy will be blamed on electron i and half on electron \underline{i} , accounting for the factor $\frac{1}{2}$.

Without this interaction 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 4.9 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,

$$H = \sum_{i=1}^I h_i^e + \frac{1}{2} \sum_{i=1}^I \sum_{\substack{i=1 \\ i \neq i}}^I v_{ii}^{\text{ee}} \quad (7.19)$$

where h_i^e is the single-electron Hamiltonian of electron i ,

$$h_i^e = -\frac{\hbar^2}{2m_e} \nabla_i^2 + \sum_{j=1}^J \frac{Z_j e^2}{4\pi\epsilon_0} \frac{1}{r_{ij}} \quad (7.20)$$

and v_{ii}^{ee} is the electron i to electron \underline{i} repulsion potential energy

$$v_{ii}^{\text{ee}} = \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}}. \quad (7.21)$$

Note that $h_1^e, h_2^e, \dots, h_I^e$ 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_{12}^{ee} , $v_{13}^{\text{ee}}, \dots, v_{I(I-2)}^{\text{ee}}, v_{I(I-1)}^{\text{ee}}$ is just in which pair of electrons you talk about.

7.3.3 The expectation value of energy

As was discussed in more detail in section 7.1, 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 is defined as $\langle E \rangle = \langle \Psi | H \Psi \rangle$. There is a problem with using this expression as stands, though. Look once again at the arsenic atom example. There are 33 electrons in it, so you could try to choose 33 promising single-electron functions to describe it. You could then try to multiply out the Slater determinant for Ψ , integrate the inner products of the individual terms on a computer, and add it all together. However, the inner product of each pair of terms involves an integration over 99 scalar coordinates. Taking 10 locations along each axis to perform the integration, you would need to compute 10^{99} values for each pair of terms. And there are $33!$ terms in the Slater determinant, or $(33!)^2 = 7.5 \cdot 10^{73}$ pairs of terms... A computer that could do that is unimaginable.

Fortunately, it turns out 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 produce only single-electron energy expectation values of the general form

$$E_n^e \equiv \langle \psi_n^s(\vec{r}) | h^e | \psi_n^s(\vec{r}) \rangle \quad (7.22)$$

You might think there should be an index i on \vec{r}_i and h_i^e to indicate which electron it is. But remember that an inner product is really an integral; this one is

$$\int_{\text{all } \vec{r}_i} \psi_n^s(\vec{r}_i)^* h_i^e \psi_n^s(\vec{r}_i) d^3 \vec{r}_i,$$

and that the name of the integration variable \vec{r}_i does not make any difference: you get the exact same value for electron 1 as for electron I or any other. So the value of i does not make a difference, and it will just be left away.

If there was just one electron and it was in single-electron state $\psi_n^s \uparrow_n$, E_n^e would be its expectation value of energy. Actually, of course, there are I electrons, each partly present in state $\psi_n^s \uparrow_n$ because of the way the Slater determinant writes out, and each electron turns out to contribute an equal share E_n^e/I to the total energy E_n^e associated with single-electron state $\psi_n^s \uparrow_n$.

The pair-repulsion Hamiltonians produce six-dimensional inner products that come in two types. The inner products of the first type will be indicated by J_{nn} , and they are

$$J_{nn} \equiv \langle \psi_n^s(\vec{r}) \psi_{\underline{n}}^s(\vec{r}) | v^{ee} | \psi_n^s(\vec{r}) \psi_{\underline{n}}^s(\vec{r}) \rangle \quad (7.23)$$

Again, what electrons \vec{r} and $\underline{\vec{r}}$ refer to is of no consequence. But written out as an integral for a specific set of electrons, using the expression for the pair energy v_{ii}^{ee} of the previous section, you get

$$\int_{\text{all } \vec{r}_i} \int_{\text{all } \vec{r}_{\underline{i}}} |\psi_n^s(\vec{r}_i)|^2 |\psi_{\underline{n}}^s(\vec{r}_{\underline{i}})|^2 \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}} d^3\vec{r}_i d^3\vec{r}_{\underline{i}}$$

If all you had was one electron i in single-electron state $\psi_n^s \uparrow_n$ and a second electron \underline{i} in single-electron state $\psi_{\underline{n}}^s \uparrow_{\underline{n}}$, this would be the expectation potential energy of their interaction. It would be the probability of electron i being near \vec{r}_i and electron \underline{i} being near $\vec{r}_{\underline{i}}$ times the Coulomb potential energy at those positions. For that reason these integrals are called “Coulomb integrals.”

The second type of integrals will be indicated by K_{nn} , and they are

$$K_{nn} \equiv \langle \psi_n^s(\vec{r}) \psi_{\underline{n}}^s(\vec{r}) | v^{\text{ee}} | \psi_{\underline{n}}^s(\vec{r}) \psi_n^s(\vec{r}) \rangle \quad (7.24)$$

These integrals are called the “exchange integrals.” Now don’t start thinking that they are there 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 them instead as Coulomb integrals with the electrons in the right hand side of the inner product exchanged.

The exchange integrals are a reflection of nature doing business in terms of an unobservable wave function, rather than the observable probabilities that appear in the Coulomb integrals. They are the equivalent of the twilight terms that have appeared before in two-state systems. Written out as integrals, you get

$$\int_{\text{all } \vec{r}_i} \int_{\text{all } \vec{r}_{\underline{i}}} \psi_n^s(\vec{r}_i)^* \psi_{\underline{n}}^s(\vec{r}_{\underline{i}})^* \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}} \psi_{\underline{n}}^s(\vec{r}_i) \psi_n^s(\vec{r}_{\underline{i}}) d^3\vec{r}_i d^3\vec{r}_{\underline{i}}.$$

Going back to the original question of the expectation energy of the complete system of I electrons, it turns out that it can be written in terms of the various inner products above as

$$\langle E \rangle = \sum_{n=1}^I E_n^e + \frac{1}{2} \sum_{n=1}^I \sum_{\underline{n}=1}^I J_{nn} - \frac{1}{2} \sum_{n=1}^I \sum_{\underline{n}=1}^I \langle \uparrow_n | \uparrow_{\underline{n}} \rangle^2 K_{nn} \quad (7.25)$$

The spin inner product $\langle \uparrow_n | \uparrow_{\underline{n}} \rangle$ is one if the orbitals have the same spin, and zero if they have opposite spin, so the square is somewhat superfluous. Consider it a reminder that if you want, you can shove the spins into K_{nn} to make it a spin, rather than spatial, orbital inner product.

If you want to see where all this comes from, the derivations are in note {A.63}. There are also some a priori things you can say about the Coulomb and exchange integrals, {A.64}; they are real, and additionally

$$J_{nn} = K_{nn} \quad J_{nn} \geq K_{nn} \geq 0 \quad J_{nn} = J_{\underline{n}n} \quad K_{nn} = K_{\underline{n}n} \quad (7.26)$$

So in terms of linear algebra, they are real symmetric matrices with nonnegative coefficients and the same main diagonal.

The analysis can easily be extended to generalized orbitals that take the form

$$\psi_n^P(\vec{r}, S_z) = \psi_{n+}^s(\vec{r})\uparrow(S_z) + \psi_{n-}^s(\vec{r})\downarrow(S_z).$$

However, the normal unrestricted spin-up or spin-down orbitals, in which either ψ_{n+}^s or ψ_{n-}^s is zero, already satisfy the variational requirement $\delta\langle E \rangle = 0$ even if generalized variations in the orbitals are allowed, {A.65}.

In any case, the expectation value of energy has been found.

7.3.4 The canonical Hartree-Fock equations

The previous section 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 7.1, {A.66}. In unrestricted Hartree-Fock, for every spatial orbital $\psi_n^s(\vec{r})$ there is an equation of the form:

$$h^e \psi_n^s(\vec{r}) + \sum_{\underline{n}=1}^I \left\langle \psi_{\underline{n}}^s(\vec{r}) \middle| v^{ee} \right| \psi_n^s(\vec{r}) \rangle \psi_n^s(\vec{r}) - \sum_{\underline{n}=1}^I \langle \downarrow_{\underline{n}} | \uparrow_n \rangle^2 \left\langle \psi_{\underline{n}}^s(\vec{r}) \middle| v^{ee} \right| \psi_n^s(\vec{r}) \rangle \psi_{\underline{n}}^s(\vec{r}) = \epsilon_n \psi_n^s(\vec{r}) \quad (7.27)$$

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 {A.66}.

Recall that h^e is the single-electron Hamiltonian consisting of its kinetic energy and its potential energy due to nuclear attractions, and that v^{ee} is the potential energy of repulsion between two electrons at given locations:

$$h^e = -\frac{\hbar^2}{2m_e} \nabla^2 - \sum_{j=1}^J \frac{Z_j e^2}{4\pi\epsilon_0 r_j} \quad r_j \equiv |\vec{r} - \vec{r}_j^n| \quad v^{ee} = \frac{e^2}{4\pi\epsilon_0} \frac{1}{r} \quad r \equiv |\vec{r} - \vec{r}|$$

So, if there were no electron-electron repulsions, i.e. $v^{ee} = 0$, the canonical equations above would turn into single-electron Hamiltonian eigenvalue problems of

the form $h^e \psi_n^s = \epsilon_n \psi_n^s$ 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 4.9: 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,

$$\mathcal{F} \psi_n^s(\vec{r}) \uparrow_n(S_z) = \epsilon_n \psi_n^s(\vec{r}) \uparrow_n(S_z)$$

where \mathcal{F} is called the “Fock operator,” and is written out further as:

$$\mathcal{F} = h^e + v^{\text{HF}}.$$

The first term in the Fock operator is the single-electron Hamiltonian. The mischief is in the innocuous-looking second term v^{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 (7.27) not described by the single-electron Hamiltonian h^e :

$$\begin{aligned} v^{\text{HF}} \psi^s(\vec{r}) \uparrow(S_z) &= \sum_{\underline{n}=1}^I \langle \psi_{\underline{n}}^s(\vec{r}) | v^{\text{ee}} | \psi_{\underline{n}}^s(\vec{r}) \rangle \psi^s(\vec{r}) \uparrow(S_z) \\ &\quad - \sum_{\underline{n}=1}^I \langle \psi_{\underline{n}}^s(\vec{r}) \uparrow_{\underline{n}}(S_{z1}) | v^{\text{ee}} | \psi^s(\vec{r}) \uparrow(S_{z1}) \rangle \psi_{\underline{n}}^s(\vec{r}) \uparrow_{\underline{n}}(S_z) \end{aligned}$$

To recover the canonical equations (7.27) from the Fock form, take an inner product with the spin $\uparrow(S_z)$. The definition of the Fock operator is unavoidably in terms of spin rather than spatial single-electron functions: 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^{HF} by itself, but only for v^{HF} applied to an arbitrary single-electron function $\psi^s \uparrow$. The reason is that v^{HF} is not a normal potential at all: the second term, the one due to the exchange integrals, does not multiply $\psi^s \uparrow$ by a potential function, it shoves it into an inner product! The Hartree-Fock “potential” v^{HF} is an *operator*, not a normal potential energy. Given a single-electron function, it produces another function.

Actually, even that is not quite true. The Hartree-Fock “potential” is only an operator *after* you have found the orbitals $\psi_1^s \uparrow_1, \psi_2^s \uparrow_2, \dots, \psi_{\underline{n}}^s \uparrow_{\underline{n}}, \dots, \psi_I^s \uparrow_I$ 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, {A.67}.

So how do you solve the canonical Hartree-Fock equations for the orbitals ψ_n^s ? If the Hartree-Fock potential v^{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^{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?

7.3.5 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 [32] 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.

Meaning of the orbital energies

In the single electron case, the “orbital energy” ϵ_n in the canonical Hartree-Fock equation

$$\begin{aligned} h^e \psi_n^s(\vec{r}) + \sum_{\underline{n}=1}^I \left\langle \psi_{\underline{n}}^s(\vec{r}) \middle| v^{ee} \right| \psi_n^s(\vec{r}) \rangle \psi_n^s(\vec{r}) \\ - \sum_{\underline{n}=1}^I \langle \downarrow_n | \uparrow_n \rangle^2 \left\langle \psi_{\underline{n}}^s(\vec{r}) \middle| v^{ee} \right| \psi_n^s(\vec{r}) \rangle \psi_n^s(\vec{r}) = \epsilon_n \psi_n^s(\vec{r}) \end{aligned}$$

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 $\psi_n^s(\vec{r})$, i.e. by putting $\psi_n^s(\vec{r})^*$ to the left of both sides and integrating over \vec{r} . That produces

$$\epsilon_n = E_n^e + \sum_{\underline{n}=1}^I J_{n\underline{n}} - \sum_{\underline{n}=1}^I \langle \downarrow_n | \uparrow_{\underline{n}} \rangle^2 K_{n\underline{n}} \quad (7.28)$$

which consists of the single-electron energy E_n^e , Coulomb integrals $J_{n\underline{n}}$ and exchange integrals $K_{n\underline{n}}$ as defined in subsection 7.3.3. It can already be seen that if all the ϵ_n are summed together, it does not produce the total expectation energy (7.25), 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 $\psi_n^s \downarrow_n$ in any meaningful sense.

However, ϵ_n can still be viewed as an approximate ionization energy. Assume that the electron is removed from orbital $\psi_n^s \downarrow_n$, leaving the electron at infinite distance at rest. No, scratch that; all electrons share orbital $\psi_n^s \downarrow_n$, not just one. Assume that one electron is removed from the system and that the remaining $I - 1$ electrons stay out of the orbital $\psi_n^s \downarrow_n$. 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_n^e 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 $n = n$ twice, but that is zero anyway since $J_{nn} = K_{nn}$.)

So by the removal of the electron “from” (read: and) orbital $\psi_n^s \downarrow_n$, 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 $\psi_n^s \downarrow_n$ 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 [32, 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], {A.70}. 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 $\psi_{I+1}^s \downarrow_{I+1}$. 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 [32, 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.

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,

$$\begin{aligned} -\frac{\hbar^2}{2m_e} \nabla^2 \psi_n^s(\vec{r}) + v^{\text{Ne}} \psi_n^s(\vec{r}) + \sum_{\underline{n}=1}^I \langle \psi_{\underline{n}}^s | v^{\text{ee}} | \psi_{\underline{n}}^s \rangle \psi_n^s(\vec{r}) \\ - \sum_{\underline{n}=1}^I \langle \uparrow_{\underline{n}} | \uparrow_{\underline{n}} \rangle^2 \langle \psi_{\underline{n}}^s | v^{\text{ee}} | \psi_{\underline{n}}^s \rangle \psi_{\underline{n}}^s(\vec{r}) = \epsilon_n \psi_n^s(\vec{r}) \end{aligned}$$

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^{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

$$\text{slowest decaying orbitals: } \psi_n^s(\vec{r}) \sim \exp(-\sqrt{-2m_e\epsilon_n}r/\hbar + \dots)$$

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

$$\psi_n^s(\vec{r}) \sim \exp(-\sqrt{2m_e|\epsilon_m|_{\min, \text{same spin, no ss}}}r/\hbar + \dots) \quad (7.29)$$

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 $\psi_n^s(\vec{r})$ 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 10.7, but with multiple spherically distributed charges rather than one, and the net charge being zero.)

Hartree-Fock limit

The Hartree-Fock approximation greatly simplifies finding a many-dimensional wave function. But really, solving the “eigenvalue problems” (7.27) 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, \dots . 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.

Configuration interaction

According to the previous subsubsection, 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 bottom line is that the Hartree-Fock solution still has a finite error, because a wave function cannot in general be described accurately using only a single Slater determinant. So what is the point in computing the wrong numbers to ten digits accuracy?

You might think that the error in the Hartree-Fock approximation would be called something like "Hartree-Fock error," "single determinant error," or "representation error," since it is due to an incomplete representation of the true wave function. However, the error is called "correlation energy" 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, {A.68}.

Anyway, in view of the fact that even an exact solution to the Hartree-Fock problem has a finite error, trying to get it exactly right is futile. At some stage, you would be much better off spending your efforts trying to reduce the inherent error in the Hartree-Fock approximation itself by including more determinants. As noted in section 4.7, 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,) orbitals $\psi_1^s \uparrow_1, \psi_2^s \uparrow_2, \dots$ have been found, the Hartree-Fock operator becomes just a Hermitian operator, and can be used to compute further orthonormal orbitals $\psi_{I+1}^s \uparrow_{I+1}, \psi_{I+2}^s \uparrow_{I+2}, \dots$. You can add these to the stew, 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 compute just one more orbital $\psi_{I+1}^s \uparrow_{I+1}$, you can already form I more Slater determinants: you can replace any of the I orbitals in the original determinant by the new function $\psi_{I+1}^s \uparrow_{I+1}$. So you can now approximate the true wave function by the more general expression

$$\begin{aligned}\Psi = & a_0 \left(| \det(\psi_1^s \uparrow_1, \psi_2^s \uparrow_2, \psi_3^s \uparrow_3, \dots, \psi_I^s \uparrow_I) \rangle \right. \\ & + a_1 | \det(\psi_{I+1}^s \uparrow_{I+1}, \psi_2^s \uparrow_2, \psi_3^s \uparrow_3, \dots, \psi_I^s \uparrow_I) \rangle \\ & + a_2 | \det(\psi_1^s \uparrow_1, \psi_{I+1}^s \uparrow_{I+1}, \psi_3^s \uparrow_3, \dots, \psi_I^s \uparrow_I) \rangle \\ & + \dots \\ & \left. + a_I | \det(\psi_1^s \uparrow_1, \psi_2^s \uparrow_2, \psi_3^s \uparrow_3, \dots, \psi_{I+1}^s \uparrow_{I+1}) \rangle \right).\end{aligned}$$

where the coefficients a_1, a_2, \dots 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 $|\det(\psi_{I+1}^s \uparrow_{I+1}, \psi_2^s \downarrow_2, \psi_3^s \downarrow_3, \dots, \psi_I^s \downarrow_I)\rangle$ is like a state where you excited an electron out of the lowest state $\psi_1^s \downarrow_1$ into an elevated energy state $\psi_{I+1}^s \uparrow_{I+1}$. (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 $\psi_{I+1}^s \uparrow_{I+1}$ in the Fock operator instead of $\psi_1^s \downarrow_1$. 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

$$\Psi = a_0 |\det(\psi_1^s \downarrow_1 + a_1 \psi_{I+1}^s \uparrow_{I+1}, \psi_2^s \downarrow_2 + a_2 \psi_{I+1}^s \uparrow_{I+1}, \dots, \psi_I^s \downarrow_I + a_I \psi_{I+1}^s \uparrow_{I+1})\rangle$$

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 $\psi_{I+2}^s \uparrow_{I+2}$ instead of $\psi_{I+1}^s \uparrow_{I+1}$, but the fact is that the 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 $\psi_{I+1}^s \uparrow_{I+1}$ and $\psi_{I+2}^s \uparrow_{I+2}$.

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 4.7. 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 maybe “excited-determinant correction” or so. However, it is called “configuration interaction,” because every non-expert will wonder whether the physicist is talking about the configuration of the nuclei or the electrons, (actually, it refers to the *practitioner* “configuring” all those determinants, no kidding,) and what it is interacting with (with the person bringing

in the coffee, of course. OK.) If you said that you were performing a “configuration interaction” while actually doing, say, some finite difference or finite element computation, just because it requires you to specify a configuration of mesh points, some people might doubt your sanity. But in physics, the standards are not so high.

Chapter 8

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 5. 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.

8.1 Molecular Solids [Descriptive]

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.69}.

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

$$V_{\text{LJ}} = C \left(d^{-12} - d^{-6} \right) \quad (8.1)$$

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 4.10 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, [18]. 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 a 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 [[2]] notes: “Surprisingly dipole-dipole attractions are fairly minor compared with 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.
-

8.2 Ionic Solids [Descriptive]

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 chlorine atom takes the loosely bound lone 3s electron away from a sodium (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 sodium 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 8.1. The sodium ion is a bit less than two Å in diameter, the chlorine one a bit less than four.

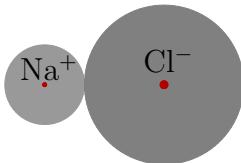


Figure 8.1: Billiard-ball model of the salt molecule.

The energetics of this process is rather interesting. Assume that you start out with a neutral sodium atom and a neutral chlorine atom that are far apart. To take the lone 2s electron out of the sodium atom, and leave it at rest at a position far from either the sodium or the chlorine atom, takes an amount of energy called the “ionization energy” of sodium. 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 chlorine. 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, {A.70}.)

Anyway, since it takes 5.14 eV to take the electron out of sodium, and you

get only 3.62 eV back by putting it into chlorine, 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* sodium; 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 sodium ion and negative chlorine ion stick together as in figure 8.1.

In other words, to create the widely separated positive sodium 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 $\epsilon_0 = 8.85 \cdot 10^{-12} \text{ C}^2/\text{J m}$ 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 4.10.

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 chlorine ions will also attract the sodium 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 8.2. 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 8.2 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.

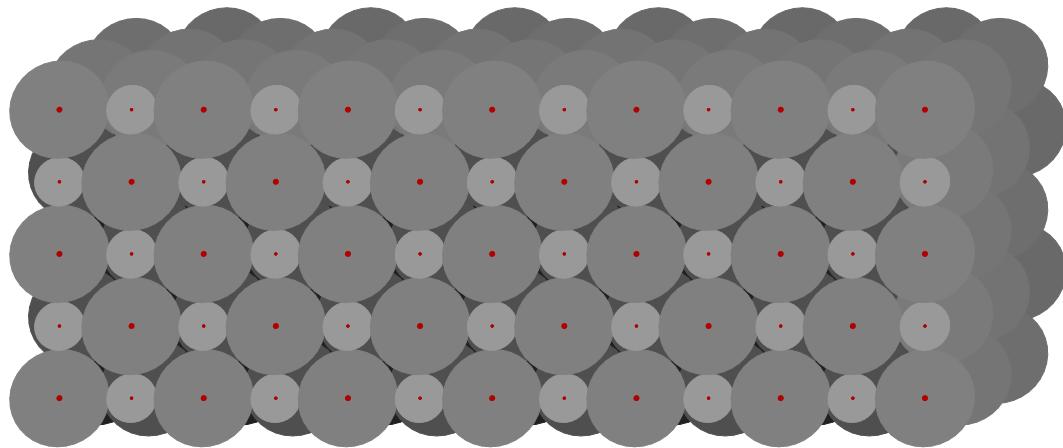


Figure 8.2: Billiard-ball model of a salt crystal.

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 8.2, think of the frontal plane as a chess board of black and white cubes, with a sodium nucleus in the center of each white cube and a chlorine nucleus in the center of each black one. The next plane of atoms can similarly be considered to consist 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,

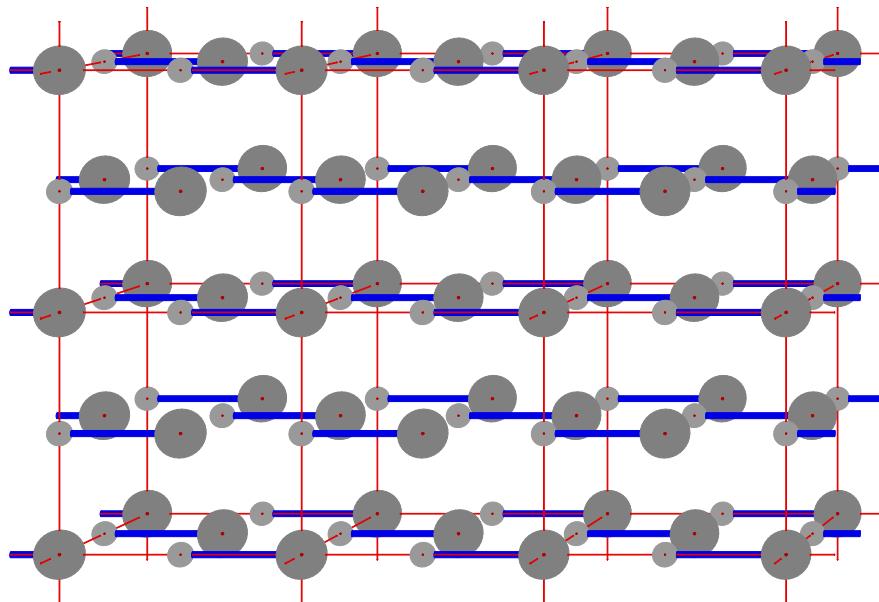


Figure 8.3: The salt crystal disassembled to show its structure.

called the “basis,” that are stacked together in a regular manner. One possible choice for the basis in salt is a single sodium ion plus a single chlorine ion to the right of it, like the molecule of figure 8.1. In figure 8.3 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 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 sodium 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 8.3. There are chlorine 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 *sodium* 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.
-

8.3 Metals [Descriptive]

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.

8.3.1 Lithium

The simplest metal is lithium. Before examining solid lithium, first consider once more the free lithium atom. Figure 8.4 gives a more realistic picture of the atom than the simplistic analysis of chapter 4.9 did. The atom is really made up of two tightly bound electrons in “|1s⟩” states very close to the nucleus, plus a loosely bound third “valence” electron in an expansive “|2s⟩” 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.

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

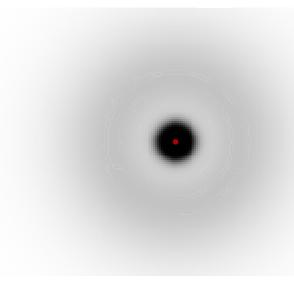


Figure 8.4: The lithium atom, scaled more correctly than in chapter 4.9

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.

Now consider solid lithium. The perfect lithium crystal would look as sketched in figure 8.5. 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.

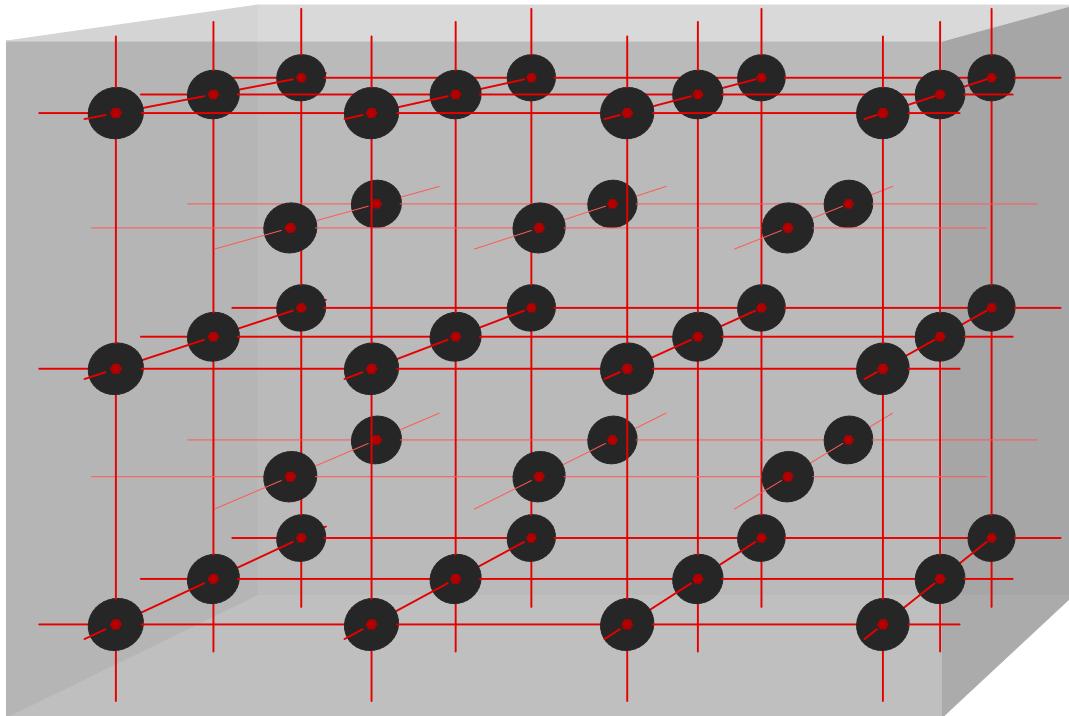


Figure 8.5: Body-centered-cubic (bcc) structure of lithium.

- 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.

8.3.2 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.

Fortunately, there is a dirty trick to get a “crystal” with only two atoms: assume that nature keeps repeating itself as indicated in figure 8.6. Mathe-

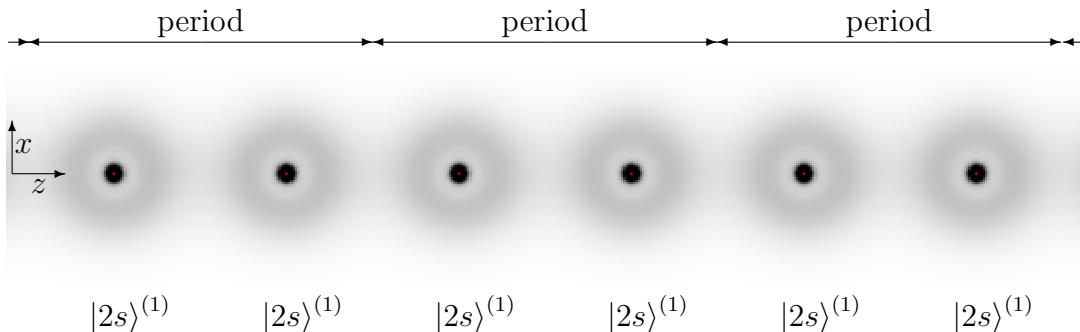


Figure 8.6: Fully periodic wave function of a two-atom lithium “crystal.”

matically, 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.

8.3.3 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.

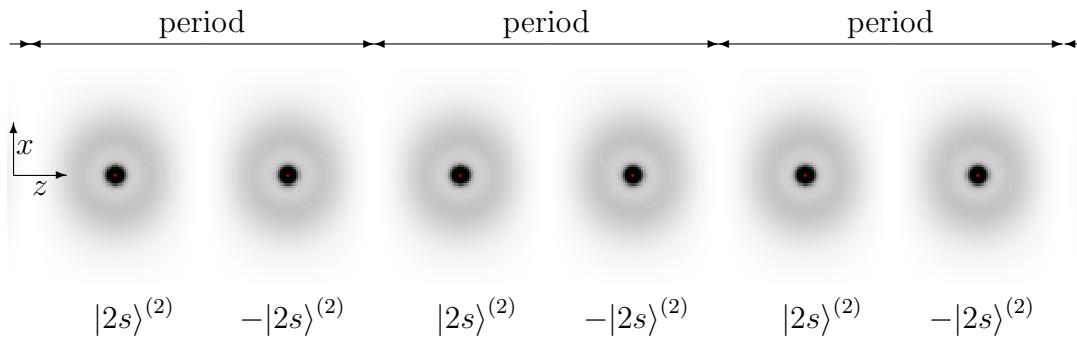


Figure 8.7: Flip-flop wave function of a two-atom lithium “crystal.”

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 8.6; 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 8.7. 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.

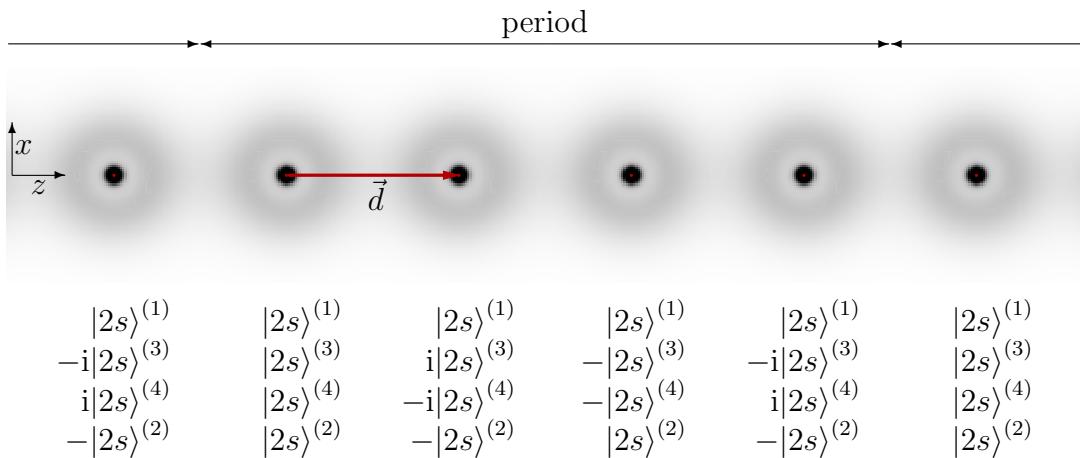


Figure 8.8: Wave functions of a four-atom lithium “crystal.” The actual picture is that of the fully periodic mode.

A one-dimensional crystal made up from four atoms is shown in figure 8.8. 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.

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.
-

8.3.4 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 non-trivial need.

The reason that the energy eigenfunctions take the form shown in figure 8.8 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 $e^{i2\pi\nu}$ 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 8.8. For the fully periodic eigenfunction $\nu = 0$, making the translation eigenvalue $e^{i2\pi\nu}$ 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 $e^{i\pi} = -1$ 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 (1.5), they change by factors $e^{\pm i\pi/2} = \pm i$ 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 $-\frac{1}{2} < \nu \leq \frac{1}{2}$. In particular for an even number of atoms J :

$$\nu = \frac{j}{J} \quad \text{for } j = -\frac{J}{2} + 1, -\frac{J}{2} + 2, -\frac{J}{2} + 3, \dots, \frac{J}{2} - 1, \frac{J}{2}$$

Note that for these values of ν , if you move over J atom spacings, $e^{i2\pi\nu J} = 1$ 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 $-J/2 < j \leq J/2$, hence $-\frac{1}{2} < \nu \leq \frac{1}{2}$: 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 $e^{i2\pi\nu}$ since $e^{i2\pi J/J} = e^{i2\pi} = 1$.

8.3.5 Floquet (Bloch) theory

Mathematically it is awkward to describe the energy eigenfunctions piecewise, as figure 8.8 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

$$z\hat{k} = u\vec{d} \quad (8.2)$$

where \vec{d} is the vector shown in figure 8.8 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 $e^{i2\pi\nu}$. A *continuous* function that does that is the exponential $e^{i2\pi\nu u}$. 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

$$\psi^p = e^{i2\pi\nu u} \psi_p^p$$

where ψ_p^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. {A.71}.

The periodic part ψ_p^p of the energy eigenfunctions is *not* the same as the $|2s\rangle^{(\cdot)}$ states of figure 8.8, because $e^{i2\pi\nu u}$ varies continuously with the crystal position $z = ud$, unlike the factors shown in figure 8.8. However, since the magnitude of $e^{i2\pi\nu u}$ is one, the magnitudes of ψ_p^p and the $|2s\rangle^{(\cdot)}$ 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

$$\boxed{\psi_k^p = e^{ikz} \psi_{p,k}^p \text{ with } \psi_{p,k}^p \text{ periodic on the atom scale } d} \quad (8.3)$$

The constant k in the exponential is called the wave number, and subscripts k have been added to ψ^p and ψ_p^p just to indicate that they will be different for different values of this wave number. Since the exponential must still equal $e^{i2\pi\nu u}$, clearly the wave number k is proportional to ν . Indeed, substituting $z = ud$ into e^{ikz} , k can be traced back to be

$$\boxed{k = \nu D \quad D = \frac{2\pi}{d} \quad -\frac{1}{2} < \nu \leq \frac{1}{2}} \quad (8.4)$$

8.3.6 Fourier analysis

As the previous subsection explained, the energy eigenfunctions in a crystal take the form of a Floquet exponential times a periodic function $\psi_{p,k}^p$. This periodic part is not normally an exponential. However, it is generally possible to write it as an infinite sum of exponentials:

$$\boxed{\psi_{p,k}^p = \sum_{m=-\infty}^{\infty} c_{km} e^{ik_m z} \quad k_m = mD \text{ for } m \text{ an integer}} \quad (8.5)$$

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 $\psi_{p,k}^P$ 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 $p_z = \hbar\partial/\text{i}\partial z$, 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 $\psi_{p,k}^P$, can be written as a sum of them. See also {A.5}.

8.3.7 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^{ikz} times a periodic function $\psi_{p,k}^P$. That periodic function can be written as a sum of Fourier exponentials $e^{ik_m z}$. It is a good idea to depict all those k -values graphically, to keep them apart. That is done in figure 8.9.

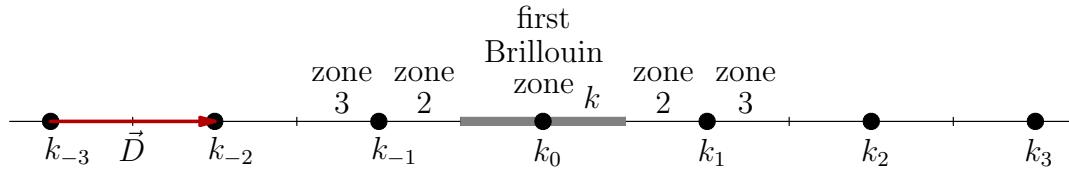


Figure 8.9: Reciprocal lattice of a one-dimensional crystal.

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 $-\frac{1}{2} < \nu \leq \frac{1}{2}$, is somewhere in the grey range in figure 8.9. This range is called the first “Brillouin zone.” It is an 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 $e^{ik_m z}$ from the Floquet exponential e^{ikz} . The $e^{ik_m z}$ can be absorbed in a redefinition of the Fourier series for the periodic part $\psi_{p,k}^P$ 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.

8.3.8 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.

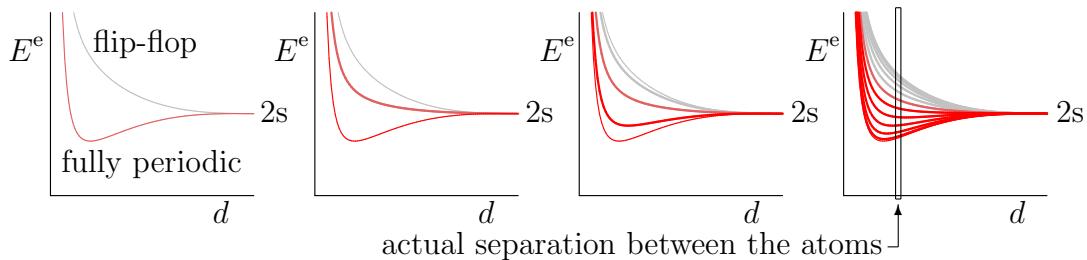


Figure 8.10: Schematic of energy bands.

For the two-atom crystal of figures 8.6 and 8.7, the answer is much like that for the hydrogen molecular ion of chapter 3.5 and hydrogen molecule of chapter 4.2. In particular, when the atom cores are far apart, the $|2s\rangle^{(1)}$ states are the same as the free lithium atom wave function $|2s\rangle$. 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 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 8.10.

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

8.10 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 8.10. 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 8.10. 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 8.10. 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 5.20, 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.

8.3.9 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 8.10, 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 8.11 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.

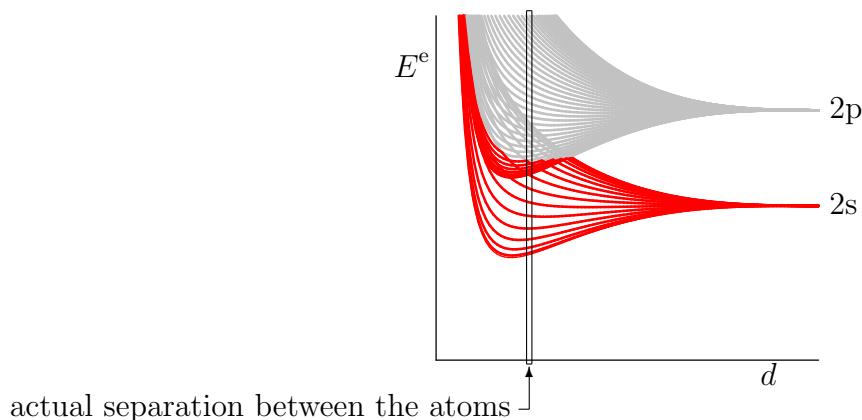


Figure 8.11: Schematic of merging bands.

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 8.10. 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 anti-symmetrically 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 8.10, 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.

8.3.10 Three-dimensional metals

The ideas of the previous subsections generalize towards three-dimensional crystals in a relatively straightforward way.

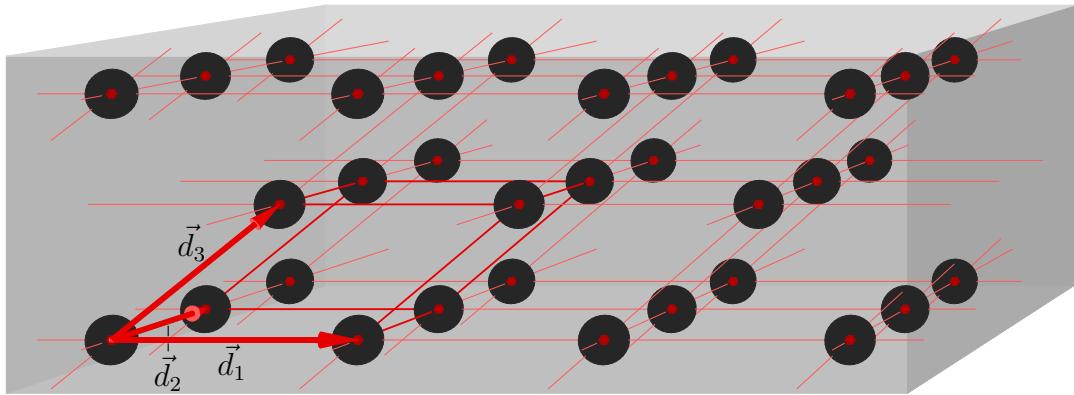


Figure 8.12: A primitive cell and primitive translation vectors of lithium.

As the lithium crystal of figure 8.12 illustrates, in a three-dimensional crystal there are three “primitive translation vectors.” The three dimensional Cartesian position \vec{r} can be written as

$$\vec{r} = u_1 \vec{d}_1 + u_2 \vec{d}_2 + u_3 \vec{d}_3 \quad (8.6)$$

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 8.5. 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 8.12, \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 8.12. 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 8.12, 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 8.5 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 built 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.

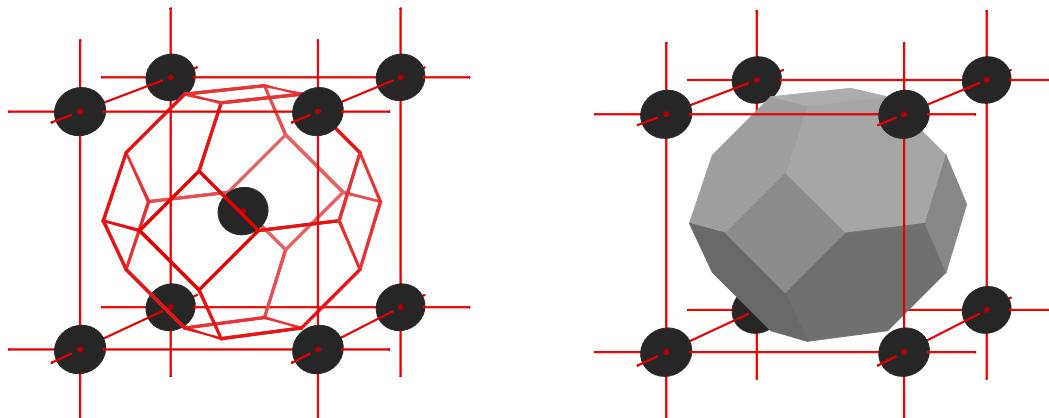


Figure 8.13: Wigner-Seitz cell of the bcc lattice.

Figure 8.13 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 8.12. 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 $e^{i2\pi\nu_1}$ when translated over \vec{d}_1 , by $e^{i2\pi\nu_2}$ when translated over \vec{d}_2 , and by $e^{i2\pi\nu_3}$ when translated over \vec{d}_3 . All that just means that they must take the Floquet (Bloch) function form

$$\psi^P = e^{i2\pi(\nu_1 u_1 + \nu_2 u_2 + \nu_3 u_3)} \psi_p^P,$$

where ψ_p^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:

$$u_1 = \frac{1}{2\pi} \vec{D}_1 \cdot \vec{r} \quad u_2 = \frac{1}{2\pi} \vec{D}_2 \cdot \vec{r} \quad u_3 = \frac{1}{2\pi} \vec{D}_3 \cdot \vec{r}$$

(8.7)

(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:

$$\begin{aligned} \vec{d}_1 \cdot \vec{D}_1 &= 2\pi, & \vec{d}_2 \cdot \vec{D}_1 &= 0, & \vec{d}_3 \cdot \vec{D}_1 &= 0, \\ \vec{d}_1 \cdot \vec{D}_2 &= 0, & \vec{d}_2 \cdot \vec{D}_2 &= 2\pi, & \vec{d}_3 \cdot \vec{D}_2 &= 0, \\ \vec{d}_1 \cdot \vec{D}_3 &= 0, & \vec{d}_2 \cdot \vec{D}_3 &= 0, & \vec{d}_3 \cdot \vec{D}_3 &= 2\pi. \end{aligned}$$

(8.8)

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

$$\psi_{\vec{k}}^p = e^{i\vec{k} \cdot \vec{r}} \psi_{p,\vec{k}}^p \quad \vec{k} = \nu_1 \vec{D}_1 + \nu_2 \vec{D}_2 + \nu_3 \vec{D}_3 \quad (8.9)$$

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 $\psi_{p,\vec{k}}^p$ too can be written in terms of exponentials. The appropriate Fourier exponentials are the eigenfunctions of the momentum operators in the three primitive directions. Converted to physical coordinates, it gives:

$$\begin{aligned} \psi_{p,\vec{k}}^p &= \sum_{m_1} \sum_{m_2} \sum_{m_3} c_{p,\vec{k}\vec{m}} e^{i\vec{k}_{\vec{m}} \cdot \vec{r}} \\ \vec{k}_{\vec{m}} &= m_1 \vec{D}_1 + m_2 \vec{D}_2 + m_3 \vec{D}_3 \text{ for } m_1, m_2, \text{ and } m_3 \text{ integers} \end{aligned} \quad (8.10)$$

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 (8.8) 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, [19, 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 8.5.1 and 8.6 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 8.5 and 8.6 will look a bit closer at them quantitatively.

8.4 Covalent Materials [Descriptive]

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.

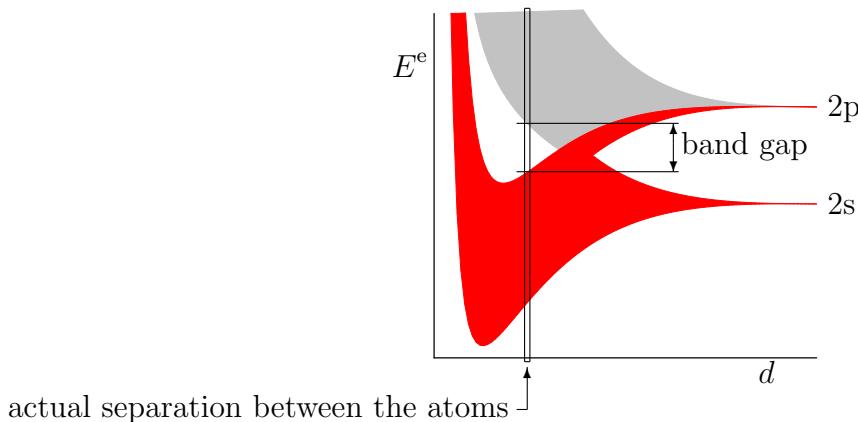


Figure 8.14: Schematic of crossing bands.

Figure 8.14 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 4.2.4), 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 5.4 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.

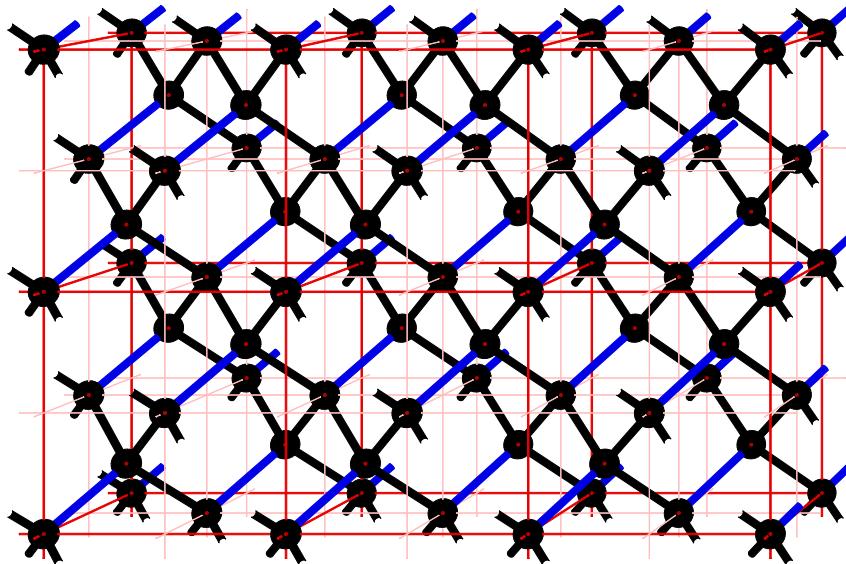


Figure 8.15: Ball and stick schematic of the diamond crystal.

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 8.15. For simplicity and clarity, the four hybrid bonds that attach each atom core to its four neighbors are shown as blue or black sticks rather than as a distribution of grey tones.

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 one 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 8.15 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 black ones.)

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 sodium 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.

8.5 Free-Electron Gas

Chapter 5 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

$$\psi_{\vec{k}}^{\text{p}}(\vec{r}) = \frac{1}{\sqrt{\mathcal{V}}} e^{i\vec{k}\cdot\vec{r}} = \frac{1}{\sqrt{\mathcal{V}}} e^{i(k_x x + k_y y + k_z z)} \quad (8.11)$$

Here the wave numbers are related to the box dimensions as

$$k_x = n_x \frac{2\pi}{\ell_x} \quad k_y = n_y \frac{2\pi}{\ell_y} \quad k_z = n_z \frac{2\pi}{\ell_z} \quad (8.12)$$

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 $\psi_{\vec{k}}^{\text{p}\downarrow}$ and a spin-down version $\psi_{\vec{k}}^{\text{p}\uparrow}$ 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.

8.5.1 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 5.17 of chapter 5.18. 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 8.3.10.) 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.

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 8.16.

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 5.17, would get very messy when crystal structure effects are added.

A lattice like the one shown in figure 8.16 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{i} , \hat{j} , 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.

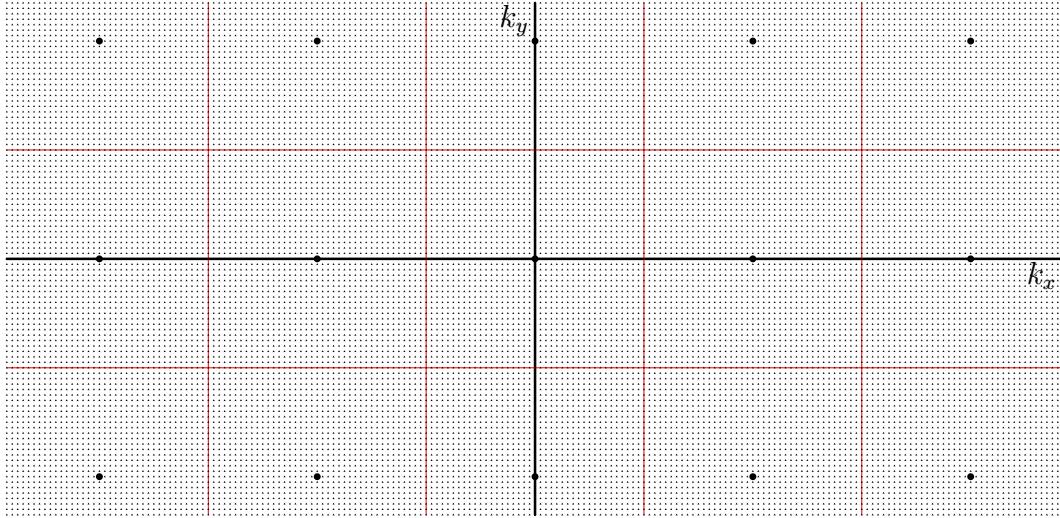


Figure 8.16: Assumed simple cubic reciprocal lattice, shown as black dots, in cross-section. The boundaries of the surrounding primitive cells are shown as thin red lines.

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 8.16 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 8.16 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 8.16.

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.

8.5.2 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 5.17, 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 8.17 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

$$\psi_{\vec{k}}^P = e^{i\vec{k} \cdot \vec{r}} \psi_{p,\vec{k}}^P$$

where the atom-scale periodic part $\psi_{p,\vec{k}}^P$ 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 $\psi_{p,\vec{k}}^P$ 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

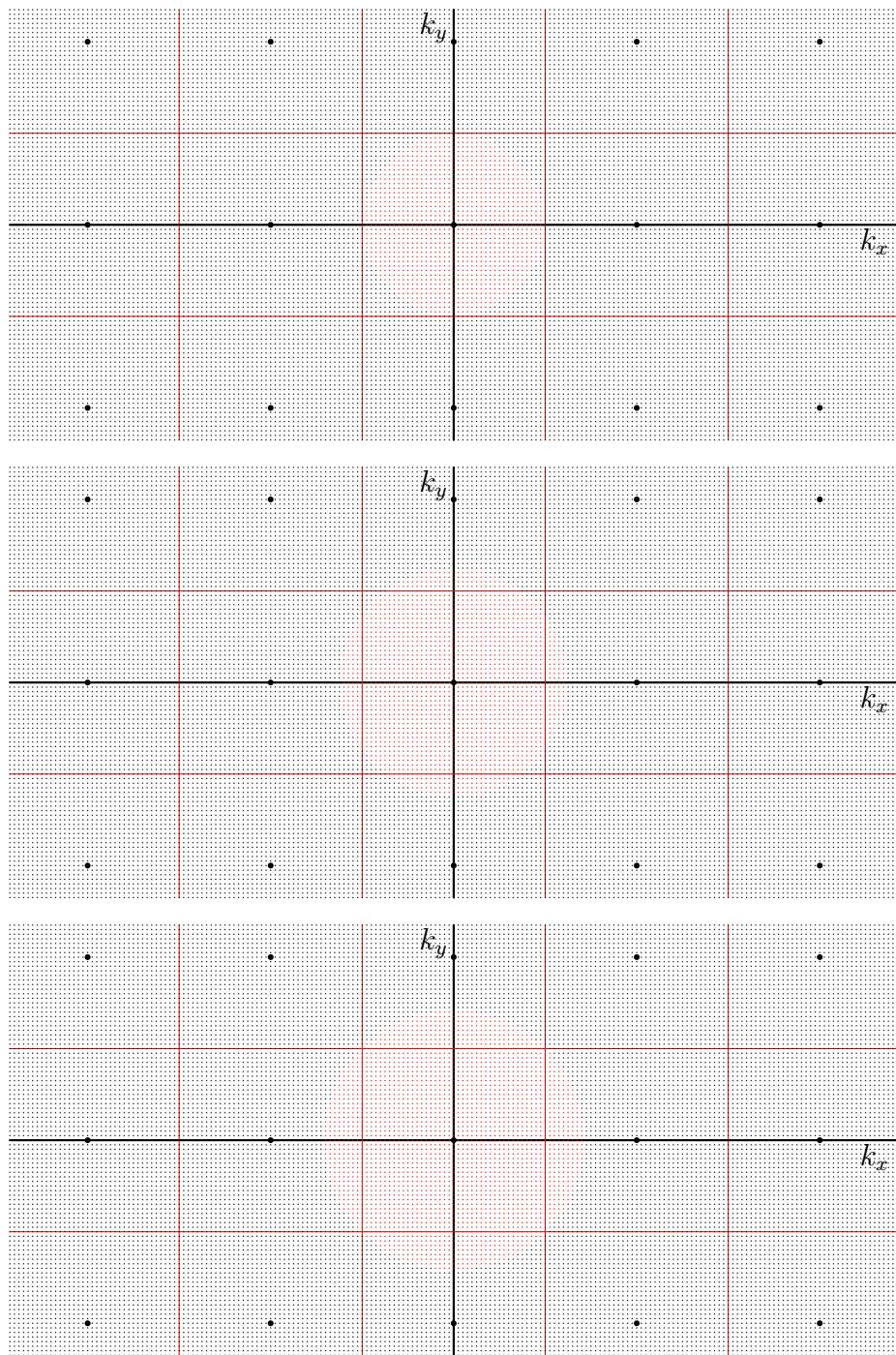


Figure 8.17: Occupied states for one, two, and three free electrons per physical lattice cell.

over the physical lattice cell; the rapid variation will now be absorbed into the lattice-cell part $\psi_{p,\vec{k}}^P$. 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, $\psi_{p,\vec{k}}^P$ 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 8.18a and 8.18b. 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.

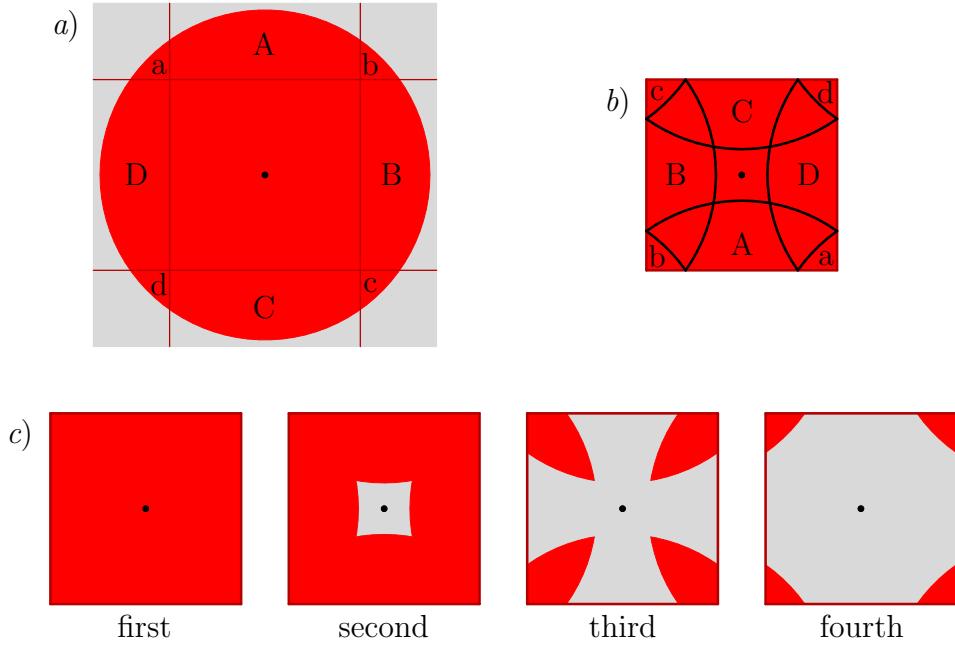


Figure 8.18: Redefinition of the occupied wave number vectors into Brillouin zones.

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 $\psi_{p,\vec{k}}^P$. 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 $\psi_{p,\vec{k}}^P$ 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 8.18b. 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 8.18c. 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.

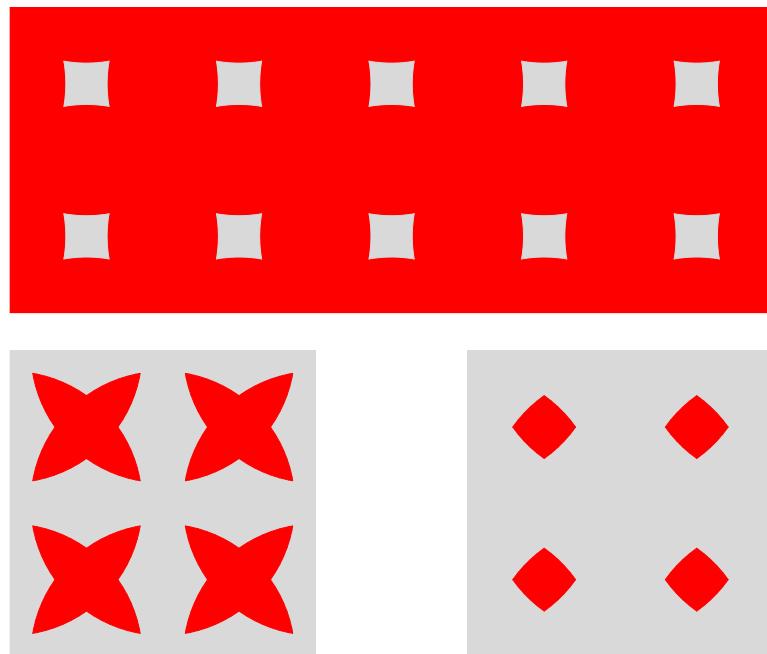


Figure 8.19: Second, third, and fourth Brillouin zones seen in the periodic zone scheme.

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 8.19. 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 8.19, increasing the esthetics.

8.6 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.

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 8.20. If there is no lattice potential, the corresponding energy eigenfunction is the free-electron one,

$$\psi_{k,0}^{\text{p}} = \frac{1}{\sqrt{8\ell_x\ell_y\ell_z}} e^{i(k_x x + k_y y + k_z z)}$$

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

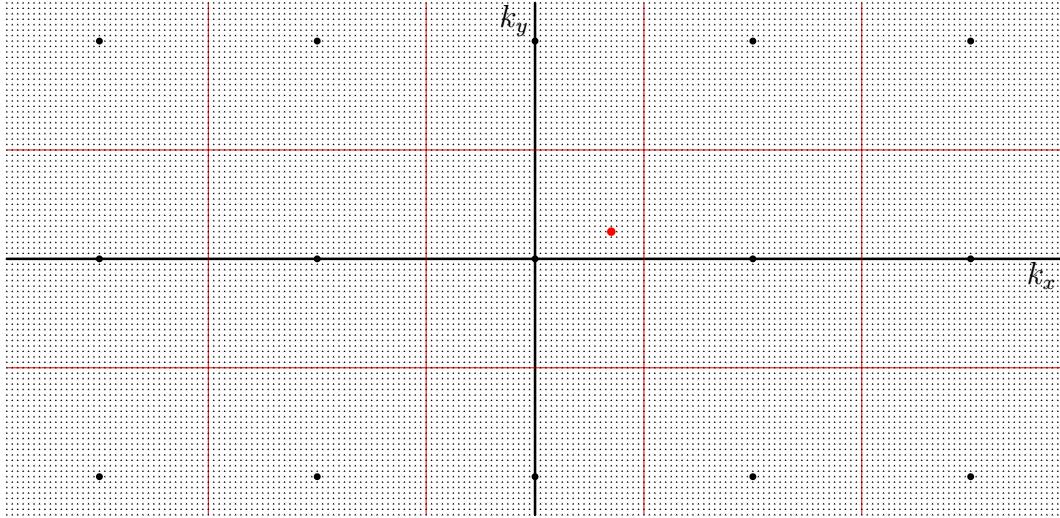


Figure 8.20: The red dot shows the wavenumber vector of a sample free electron wave function. It is to be corrected for the lattice potential.

the form

$$\psi_{\vec{k}}^p = \psi_{p,\vec{k}}^p e^{i(k_x x + k_y y + k_z z)}$$

where $\psi_{p,\vec{k}}^p$ is periodic on an atomic scale. If the lattice potential is weak, as assumed here,

$$\psi_{p,\vec{k}}^p \approx \frac{1}{\sqrt{8\ell_x\ell_y\ell_z}}$$

Also, the energy will be almost the free-electron one:

$$E_{\vec{k}}^e \approx E_{\vec{k},0}^e = \frac{\hbar^2}{2m_e} k^2$$

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.

8.6.1 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 chapter 12.1 and {A.114}. 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

$$H_{\vec{k}\vec{k}} \equiv \langle \psi_{\vec{k},0}^p | V \psi_{\vec{k},0}^p \rangle \quad (8.13)$$

where V is the lattice potential, and the $\psi_{\vec{k},0}^p$ are the free-electron energy eigenfunctions.

In those terms, the energy of the eigenfunction $\psi_{\vec{k}}$ with Floquet wave number \vec{k} is

$$E_{\vec{k}}^e \approx E_{\vec{k},0}^e + H_{\vec{k}\vec{k}} - \sum_{\vec{k} \neq \vec{k}} \frac{|H_{\vec{k}\vec{k}}|^2}{E_{\vec{k},0}^e - E_{\vec{k},0}^e} + \dots \quad (8.14)$$

Here $E_{\vec{k},0}^e$ 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.

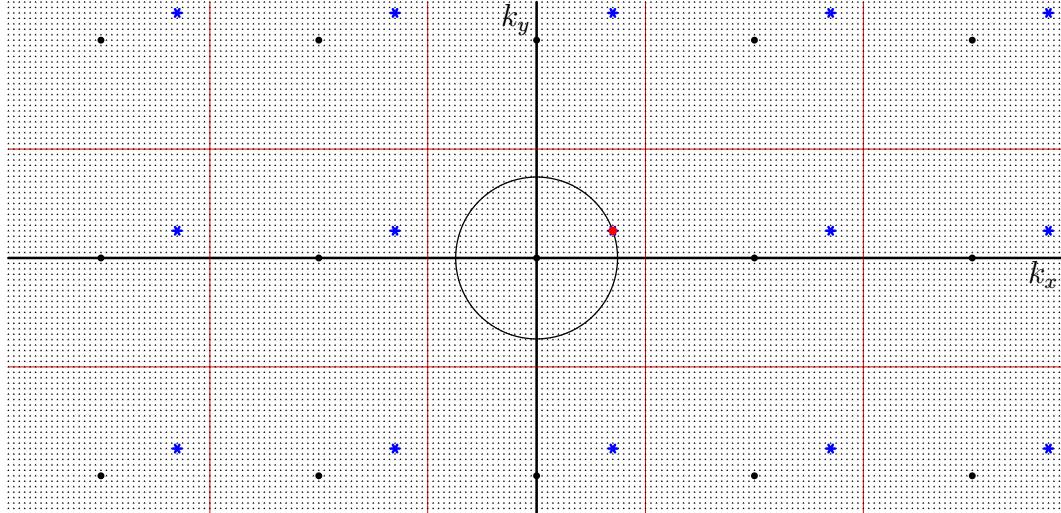


Figure 8.21: The grid of nonzero Hamiltonian perturbation coefficients and the problem sphere in wave number space.

That makes the sum in (8.14) 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 $\psi_{\vec{k},0}^p$. The only Hamiltonian perturbation coefficients that are nonzero occur for the \vec{k} values shown in figure 8.21 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 chapter 12.1.3 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 $\psi_{\vec{k},0}^P$ on the grid of blue stars that have the same free-electron energy $E_{\vec{k},0}^e$ as the eigenfunction $\psi_{\vec{k},0}^P$. You can infer the problem from (8.14); 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 $\psi_{\vec{k}}^P$. In particular, since the free-electron energy $E_{\vec{k},0}^e$ equals $\hbar^2 k^2 / 2m_e$, 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.

8.6.2 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 $E_{\vec{k},0}^e$ without lattice potential changes due to the lattice potential by an amount:

$$\Delta E_{\vec{k}}^e = - \sum_{\vec{k} \neq \vec{k}} \frac{|H_{\vec{k}\vec{k}}|^2}{E_{\vec{k},0}^e - E_{\vec{k},0}^e} \quad (8.15)$$

where the $H_{\vec{k}\vec{k}}$ 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 8.21, $\underline{\vec{k}}$ was an summation index over the blue grid points of that figure, and $E_{\vec{k},0}^e$ and $E_{\vec{k},0}^e$ were proportional to the square distances from the origin to points $\underline{\vec{k}}$, respectively \vec{k} . $E_{\vec{k},0}^e$ is also the energy level of the eigenfunction without lattice potential.

The expression above for the energy change is not valid when $E_{\vec{k},0}^e = E_{\vec{k},0}^e$, in which case it would incorrectly give infinite change in energy. However, it does apply when $E_{\vec{k},0}^e \approx E_{\vec{k},0}^e$, in which case it predicts unusually large changes in energy. The condition $E_{\vec{k},0}^e \approx E_{\vec{k},0}^e$ means that a blue star $\underline{\vec{k}}$ on the grid in figure 8.21 is almost the same distance from the origin as the red point \vec{k} itself.

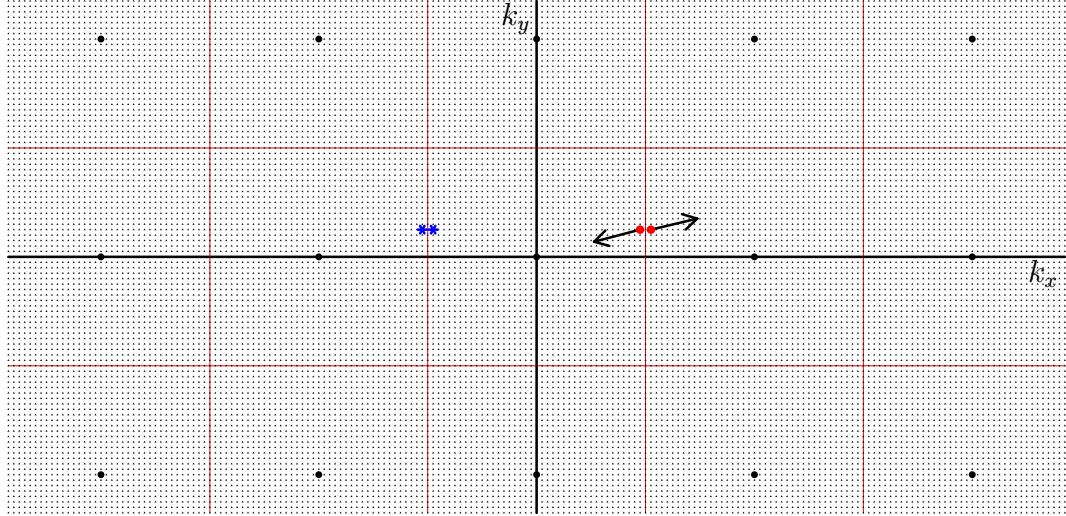


Figure 8.22: Tearing apart of the wave number space energies.

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 8.22 shows two neighboring states \vec{k} straddling the right-hand vertical plane of the cell, as well as their lattice \vec{k} -values that cause the unusually large energy changes.

For the left of the two states, $E_{\vec{k},0}^e$ is just a bit larger than $E_{\vec{k},0}^e$, so the energy change (8.15) 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; $E_{\vec{k},0}^e$ for that state is just a bit less than $E_{\vec{k},0}^e$, 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.

That is illustrated for an arbitrarily chosen example lattice potential in figure 8.23. 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.

But notice in figure 8.23 that the energy levels get torn apart along many

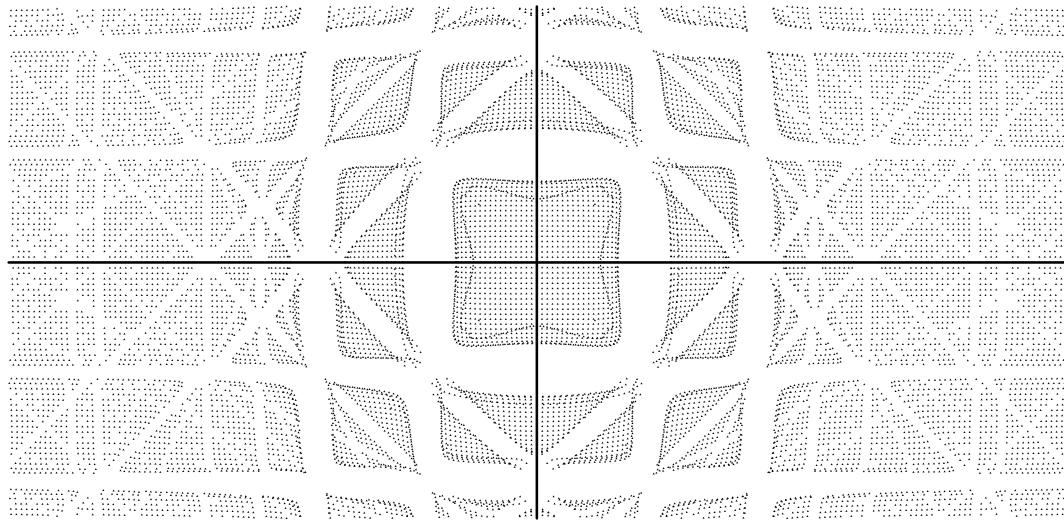


Figure 8.23: Effect of a lattice potential on the energy. The energy is represented by the square distance from the origin, and is relative to the energy at the origin.

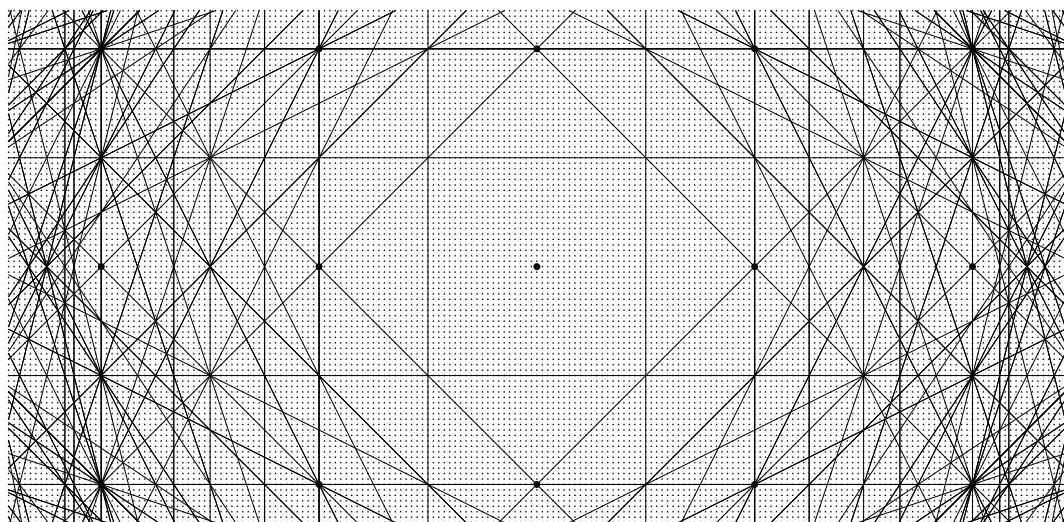


Figure 8.24: Bragg planes seen in wave number space cross section.

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 8.24 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 8.24). 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 8.24). They can be pushed into the first Brillouin zone using the fundamental translation vectors to combine into a Wigner-Seitz cell shape.

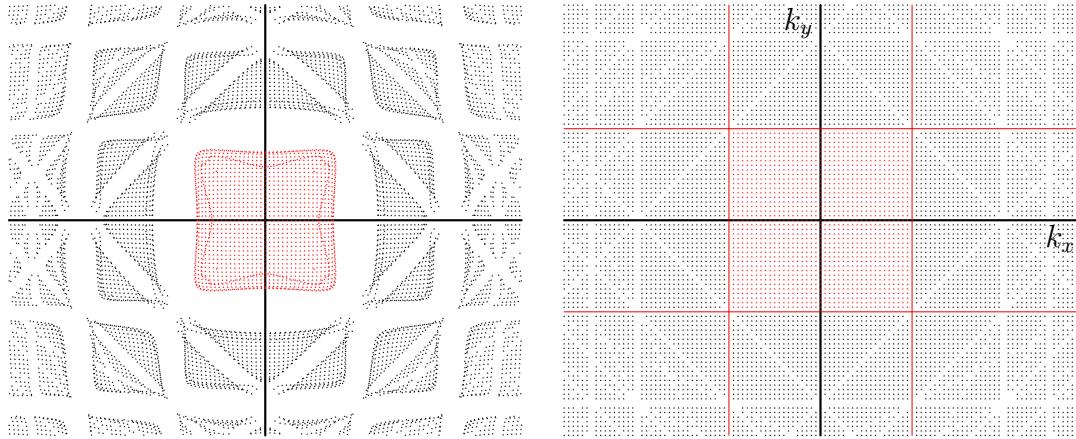


Figure 8.25: Occupied states for the energies of figure 8.23 if there are two valence electrons per lattice cell. Left: energy. Right: wave numbers.

For a sufficiently strong lattice potential like the one in figure 8.23, 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 8.25. 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 8.25 also illustrates that a lattice potential can change the Floquet wave number vectors that get occupied.

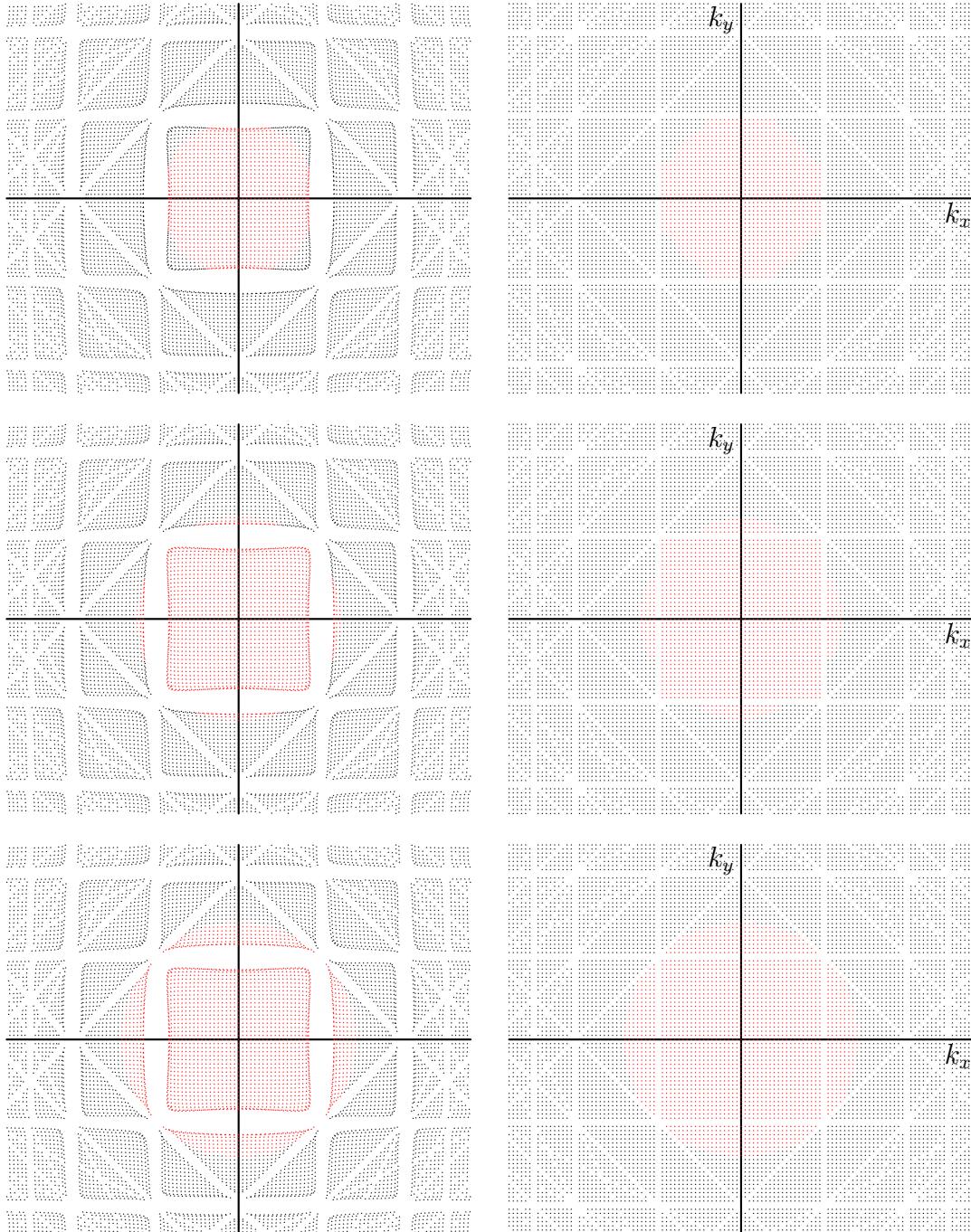


Figure 8.26: Smaller lattice potential. From top to bottom shows one, two and three valence electrons per lattice cell. Left: energy. Right: wave numbers.

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 8.17, 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 8.26. 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.

8.7 Additional Points [Descriptive]

This section mentions a couple of additional very basic issues in the quantum mechanics of solids.

8.7.1 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 4.9, 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.73}.

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.

Also, 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 $S, P, D, F, G, H, I, K, \dots$ 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. Also, combined angular momentum is not all that meaningful as a number by itself, so stating the correct amount will not give away too much of a secret.

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. {A.74}.

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.

8.7.2 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 8.11, 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 ener-

getic. 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.

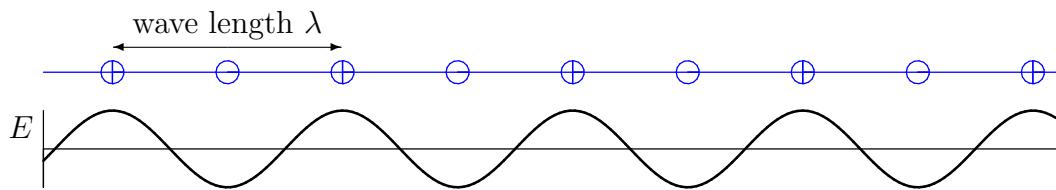


Figure 8.27: Depiction of an electromagnetic ray.

Figure 8.27 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 8.28. 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 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.

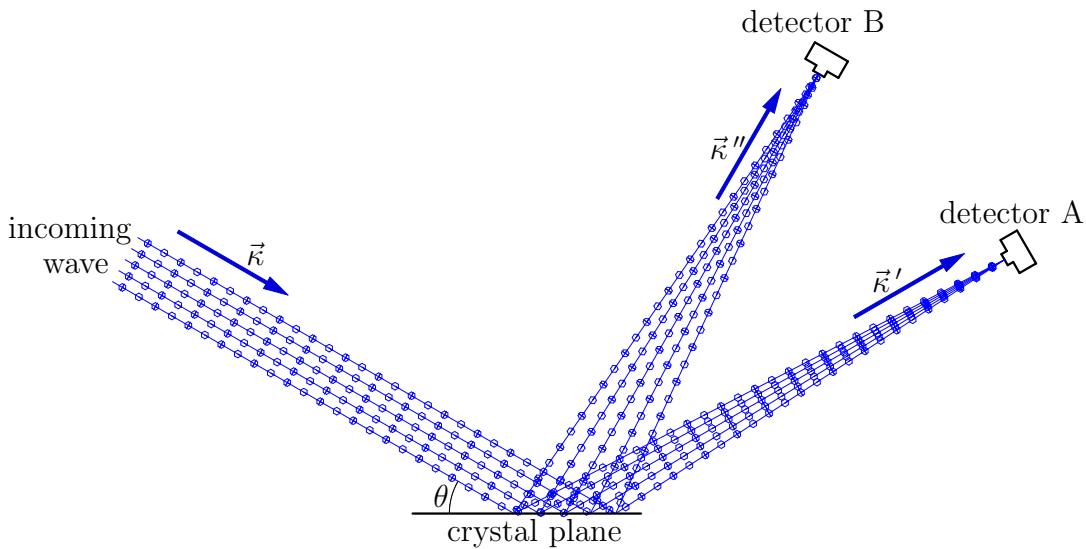


Figure 8.28: Law of reflection in elastic scattering from a plane.

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 8.29. 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. (Also note that whatever variations in phase there are in the signals arriving at the detector in figure 8.29 are artifacts: for graphical reasons the detector is much closer to the specimen than it should be. The spacing between planes should

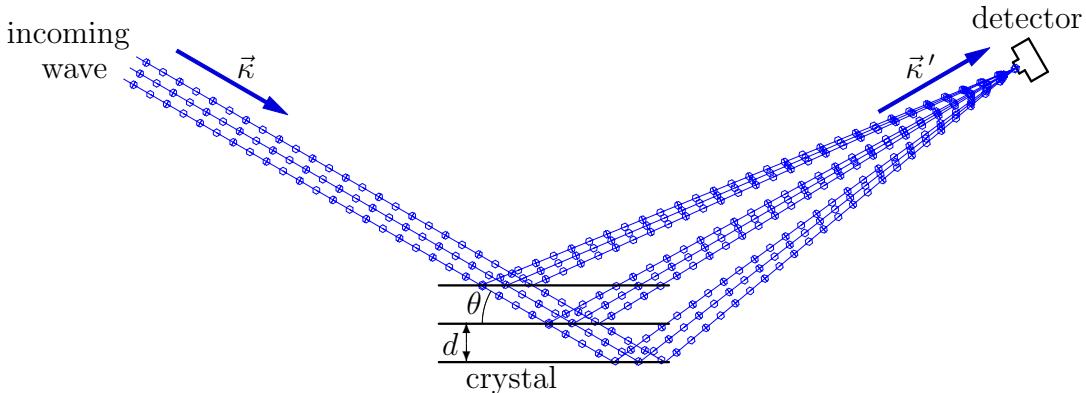


Figure 8.29: Scattering from multiple “planes of atoms”.

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:

$$2d \sin \theta = n\lambda \quad (8.16)$$

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

$$e^{i\kappa(s-ct)},$$

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 $e^{i\kappa\Delta s}$. 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 cancellation of signals at the detector.

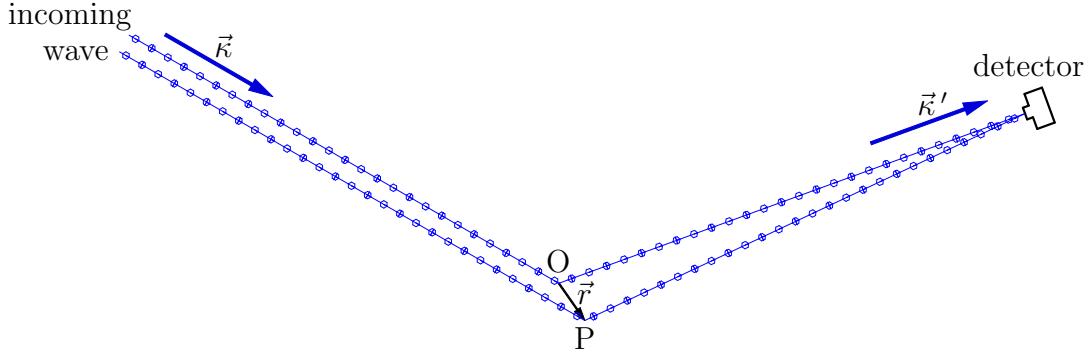


Figure 8.30: Difference in travel distance when scattered from P rather than O.

So, how much is the phase factor $e^{i\kappa\Delta s}$? Figure 8.30 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 $\vec{\kappa}$:

$$\Delta s_1 = \vec{r} \cdot \frac{\vec{\kappa}}{\kappa} \quad \text{so} \quad e^{i\kappa\Delta s_1} = e^{i\vec{\kappa} \cdot \vec{r}}.$$

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

$$\Delta s_2 = -\vec{r} \cdot \frac{\vec{\kappa}'}{\kappa} \quad \text{so} \quad e^{i\kappa\Delta s_2} = e^{-i\vec{\kappa}' \cdot \vec{r}}$$

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

$$e^{i(\vec{\kappa} - \vec{\kappa}') \cdot \vec{r}}.$$

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 8.28, 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 cancellation 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 8.29, 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 $e^{-id2\kappa \sin \theta}$. 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 $-i2\pi$. 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

$$D = C \int_{\text{all } \vec{r}} n(\vec{r}) e^{i(\vec{\kappa} - \vec{\kappa}') \cdot \vec{r}} d^3 \vec{r}$$

where C is some constant. Now the electron density is periodic on crystal lattice scale, so according to section 8.3.10 it can be written as a Fourier series, giving the signal as

$$D = C \sum_{\text{all } \vec{k}_n} \int_{\text{all } \vec{r}} n_{\vec{k}_n} e^{i(\vec{k}_n + \vec{\kappa} - \vec{\kappa}') \cdot \vec{r}} d^3 \vec{r}$$

where the \vec{k}_n wave number vectors form the reciprocal lattice and the numbers $n_{\vec{k}_n}$ are constants. Because the volume integration above extends over countless lattice cells, there will be massive cancellation of signal unless the exponential

is constant, which requires that the factor multiplying the position coordinate is zero:

$$\vec{k}_{\vec{n}} = \vec{\kappa}' - \vec{\kappa} \quad (8.17)$$

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

$$\vec{\kappa} \cdot \vec{k}_{\vec{n}} = -\frac{1}{2} \vec{k}_{\vec{n}} \cdot \vec{k}_{\vec{n}} \quad (8.18)$$

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 8.24. 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 $\hat{p}_z = \hbar\partial/\partial z$ are proportional to $e^{i\kappa z}$, 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 9

Basic and Quantum Thermodynamics

Chapter 5 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 4.7–4.9 and chapter 5.1–5.16.

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. Non-trivial 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.”

9.1 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 $\frac{3}{2}k_B T$ with $k_B = 1.380\,65\,10^{-23}$ 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 $\frac{3}{2}k_B T$. 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 $\frac{3}{2}k_B T$ 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 $k_B T$ 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 9.14.4 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 9.8 through 9.10.

9.2 Single-Particle and System Eigenfunctions

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 $\psi_1^P(\vec{r}, S_z)$, $\psi_2^P(\vec{r}, S_z)$, . . .

The basic case is that of noninteracting particles in a box, like discussed in chapter 5.2. For such particles the single-particle eigenfunctions take the spatial form

$$\psi_n^P = \sqrt{\frac{8}{\ell_x \ell_y \ell_z}} \sin(k_x x) \sin(k_y y) \sin(k_z z)$$

where k_x, k_y, k_z are constants, called the “wave number components.” Different

values for these constants correspond to different single-particle eigenfunctions, with single-particle energy

$$E_n^P = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2) = \frac{\hbar^2}{2m}k^2$$

The single-particle energy eigenfunctions will in this chapter be numbered as $n = 1, 2, 3, \dots, N$. Higher values of index n correspond to eigenfunctions of equal or higher energy E_n^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 ψ_q^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 ψ_1^P , and the system will be in its ground state

$$\psi_1^S = \psi_1^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2})\psi_1^P(\vec{r}_3, S_{z3})\psi_1^P(\vec{r}_4, S_{z4})\psi_1^P(\vec{r}_5, S_{z5}) \dots \psi_1^P(\vec{r}_I, S_{zI})$$

where I is the total number of particles in the system. This does assume that the single-particle ground state energy E_1^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:

$$\psi_q^S = \psi_{24}^P(\vec{r}_1, S_{z1})\psi_4^P(\vec{r}_2, S_{z2})\psi_7^P(\vec{r}_3, S_{z3})\psi_1^P(\vec{r}_4, S_{z4})\psi_6^P(\vec{r}_5, S_{z5}) \dots \psi_{54}^P(\vec{r}_{36}, S_{z36})$$

This system eigenfunction has an energy that is the sum of the 36 single-particle eigenstate energies involved:

$$E_q^S = E_{24}^P + E_4^P + E_7^P + E_1^P + E_6^P + \dots + E_{54}^P$$

E_8^P	ψ_{60}^P	ψ_{61}^P	ψ_{62}^P	ψ_{63}^P	ψ_{64}^P	ψ_{65}^P	ψ_{66}^P	ψ_{67}^P	ψ_{68}^P	ψ_{69}^P	ψ_{70}^P	ψ_{71}^P	ψ_{72}^P	ψ_{73}^P
E_7^P	ψ_{47}^P	ψ_{48}^P	ψ_{49}^P	ψ_{50}^P	ψ_{51}^P	ψ_{52}^P	ψ_{53}^P	ψ_{54}^P	ψ_{55}^P	ψ_{56}^P	ψ_{57}^P	ψ_{58}^P	ψ_{59}^P	
E_6^P	ψ_{35}^P	ψ_{36}^P	ψ_{37}^P	ψ_{38}^P	ψ_{39}^P	ψ_{40}^P	ψ_{41}^P	ψ_{42}^P	ψ_{43}^P	ψ_{44}^P	ψ_{45}^P	ψ_{46}^P		(17) (18)
E_5^P	ψ_{24}^P	ψ_{25}^P	ψ_{26}^P	ψ_{27}^P	ψ_{28}^P	ψ_{29}^P	ψ_{30}^P	ψ_{31}^P	ψ_{32}^P	ψ_{33}^P	ψ_{34}^P		(16)	
E_4^P	ψ_{15}^P	ψ_{16}^P	ψ_{17}^P	ψ_{18}^P	ψ_{19}^P	ψ_{20}^P	ψ_{21}^P	ψ_{22}^P	ψ_{23}^P					
E_3^P	ψ_8^P	ψ_9^P	ψ_{10}^P	ψ_{11}^P	ψ_{12}^P	ψ_{13}^P	ψ_{14}^P							
E_2^P	ψ_3^P	ψ_4^P	ψ_5^P	ψ_6^P	ψ_7^P									
E_1^P	ψ_1^P	ψ_2^P												

Figure 9.1: 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 9.1. 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 ψ_{24}^P box, particle 2 is inside the ψ_4^P 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^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 5.3. 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 5.12. Various formulae given here and in literature may need to be adjusted then.

Of course, in normal non-nano 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

$$\psi_q^S = \psi_{24}^P(\vec{r}_1, S_{z1})\psi_4^P(\vec{r}_2, S_{z2})\psi_7^P(\vec{r}_3, S_{z3})\psi_1^P(\vec{r}_4, S_{z4})\psi_6^P(\vec{r}_5, S_{z5}) \dots \psi_{54}^P(\vec{r}_{36}, S_{z36})$$

into

$$\psi_{\underline{q}}^S = \psi_{24}^P(\vec{r}_1, S_{z1})\psi_6^P(\vec{r}_2, S_{z2})\psi_7^P(\vec{r}_3, S_{z3})\psi_1^P(\vec{r}_4, S_{z4})\psi_4^P(\vec{r}_5, S_{z5}) \dots \psi_{54}^P(\vec{r}_{36}, S_{z36})$$

and that is simply a different wave function, because the states are different, independent functions. In terms of the pictorial representation figure 9.1, swapping the numbers “2” and “5” in the particles changes the picture.

As chapter 4.7 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 9.1, 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 9.2.

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 9.2 for bosons is totally unacceptable for fermions since it uses many of

E_8^P	ψ_{60}^P	ψ_{61}^P	ψ_{62}^P ○	ψ_{63}^P	ψ_{64}^P	ψ_{65}^P	ψ_{66}^P	ψ_{67}^P	ψ_{68}^P	ψ_{69}^P	ψ_{70}^P	ψ_{71}^P	ψ_{72}^P	ψ_{73}^P
E_7^P	ψ_{47}^P	ψ_{48}^P	ψ_{49}^P	ψ_{50}^P	ψ_{51}^P	ψ_{52}^P	ψ_{53}^P	ψ_{54}^P ○	ψ_{55}^P	ψ_{56}^P	ψ_{57}^P	ψ_{58}^P	ψ_{59}^P	
E_6^P	ψ_{35}^P	ψ_{36}^P	ψ_{37}^P	ψ_{38}^P	ψ_{39}^P	ψ_{40}^P	ψ_{41}^P	ψ_{42}^P	ψ_{43}^P	ψ_{44}^P	ψ_{45}^P	ψ_{46}^P ○ ○		
E_5^P	ψ_{24}^P ○	ψ_{25}^P	ψ_{26}^P	ψ_{27}^P	ψ_{28}^P	ψ_{29}^P	ψ_{30}^P	ψ_{31}^P ○	ψ_{32}^P	ψ_{33}^P	ψ_{34}^P ○			
E_4^P	ψ_{15}^P ○	ψ_{16}^P	ψ_{17}^P ○	ψ_{18}^P ○	ψ_{19}^P ○	ψ_{20}^P ○	ψ_{21}^P	ψ_{22}^P	ψ_{23}^P					
E_3^P	ψ_8^P ○	ψ_9^P ○	ψ_{10}^P	ψ_{11}^P	ψ_{12}^P ○ ○	ψ_{13}^P ○	ψ_{14}^P ○ ○							
E_2^P	ψ_3^P ○	ψ_4^P ○ ○ ○	ψ_5^P ○ ○ ○	ψ_6^P ○ ○ ○	ψ_7^P ○ ○ ○									
E_1^P	ψ_1^P ○ ○ ○ ○ ○ ○	ψ_2^P ○ ○ ○ ○ ○ ○												

Figure 9.2: Graphical depiction of an arbitrary system energy eigenfunction for 36 identical bosons.

E_8^P	ψ_{60}^P	ψ_{61}^P	ψ_{62}^P	ψ_{63}^P	ψ_{64}^P	ψ_{65}^P	ψ_{66}^P	ψ_{67}^P	ψ_{68}^P	ψ_{69}^P	ψ_{70}^P	ψ_{71}^P	ψ_{72}^P	ψ_{73}^P
E_7^P	ψ_{47}^P	ψ_{48}^P	ψ_{49}^P	ψ_{50}^P	ψ_{51}^P	ψ_{52}^P	ψ_{53}^P	ψ_{54}^P	ψ_{55}^P	ψ_{56}^P ●	ψ_{57}^P	ψ_{58}^P	ψ_{59}^P	
E_6^P	ψ_{35}^P	ψ_{36}^P ●	ψ_{37}^P	ψ_{38}^P ●	ψ_{39}^P	ψ_{40}^P	ψ_{41}^P	ψ_{42}^P	ψ_{43}^P ●	ψ_{44}^P	ψ_{45}^P ●	ψ_{46}^P		
E_5^P	ψ_{24}^P ●	ψ_{25}^P ●	ψ_{26}^P	ψ_{27}^P	ψ_{28}^P	ψ_{29}^P ●	ψ_{30}^P ●	ψ_{31}^P	ψ_{32}^P ●	ψ_{33}^P	ψ_{34}^P ●			
E_4^P	ψ_{15}^P ●	ψ_{16}^P ●	ψ_{17}^P	ψ_{18}^P ●	ψ_{19}^P ●	ψ_{20}^P ●	ψ_{21}^P ●	ψ_{22}^P ●	ψ_{23}^P ●					
E_3^P	ψ_8^P ●	ψ_9^P ●	ψ_{10}^P ●	ψ_{11}^P ●	ψ_{12}^P ●	ψ_{13}^P ●	ψ_{14}^P ●							
E_2^P	ψ_3^P ●	ψ_4^P ●	ψ_5^P ●	ψ_6^P ●	ψ_7^P ●									
E_1^P	ψ_1^P ●	ψ_2^P ●												

Figure 9.3: Graphical depiction of an arbitrary system energy eigenfunction for 33 identical fermions.

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 9.3.

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.

9.3 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 9.1–9.3 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 9.1–9.3. Doing so allows for more answerable questions such as: “How many system energy eigenfunctions ψ_q^S 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, \dots), then what is the number $Q_{\vec{I}}$ of system eigenfunctions ψ_q^S 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 note {A.75}. But here are the results. (Drums please.) The system eigenfunction counts for distinguishable particles, bosons, and fermions are:

$$Q_I^d = I! \prod_{\text{all } s} \frac{N_s^{I_s}}{(I_s)!} \quad (9.1)$$

$$Q_I^b = \prod_{\text{all } s} \frac{(I_s + N_s - 1)!}{(I_s)!(N_s - 1)!} \quad (9.2)$$

$$Q_I^f = \prod_{\text{all } s} \frac{(N_s)!}{(I_s)!(N_s - I_s)!} \quad (9.3)$$

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

$$Q_I^d = I! \frac{N_1^{I_1}}{(I_1)!} \frac{N_2^{I_2}}{(I_2)!} \frac{N_3^{I_3}}{(I_3)!} \frac{N_4^{I_4}}{(I_4)!} \dots$$

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

$$n! = \prod_{\underline{n}=1}^n \underline{n} = 1 \times 2 \times 3 \times \dots \times n$$

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.

The first example is illustrated in quantum-mechanical terms in figure 9.4. 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 $E_1^p = 1$

$E_3^P = 4$	ψ_5^P	ψ_6^P	ψ_7^P	ψ_8^P ④	ψ_9^P	ψ_{10}^P	ψ_{11}^P	ψ_{12}^P
$E_2^P = 2$	ψ_2^P	ψ_3^P ①	ψ_4^P ③					
$E_1^P = 1$	ψ_1^P ②							

Figure 9.4: Illustrative small model system having 4 distinguishable particles. The particular eigenfunction shown is arbitrary.

(arbitrary units), shelf 2 has $N_2 = 3$ single-particle states with energy $E_2^P = 2$, (note that $3 \approx 2\sqrt{2}$), and shelf 3 has $N_3 = 4\sqrt{4} = 8$ single-particle states with energy $E_3^P = 4$. 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 $E_3^P = 4$. 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.

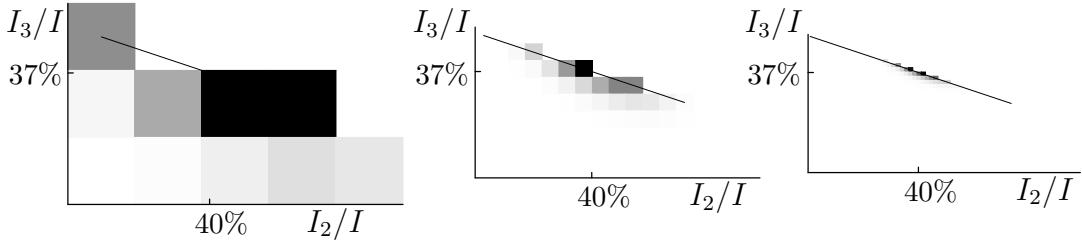


Figure 9.5: The number of system energy eigenfunctions for a simple model system with only three energy shelves. Positions of the squares indicate the numbers of particles on shelves 2 and 3; darkness of the squares indicates the relative number of eigenfunctions with those shelf numbers. Left: system 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 (9.1), is shown graphically in the left graph of figure 9.5. Darker squares indicate more eigenfunctions with those shelf occupation numbers. The oblique line in figure 9.5 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

$$\psi_1^S = \psi_1^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2})\psi_1^P(\vec{r}_3, S_{z3})\psi_1^P(\vec{r}_4, S_{z4}).$$

This is represented by the white square at the origin in the left graph of figure 9.5.

As another example, the darkest square in the left graph of figure 9.5 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,

$$\psi_3^P(\vec{r}_1, S_{z1})\psi_1^P(\vec{r}_2, S_{z2})\psi_4^P(\vec{r}_3, S_{z3})\psi_8^P(\vec{r}_4, S_{z4}),$$

is the one depicted in figure 9.4. It has particle 1 in single-particle state ψ_3^P , which is on shelf 2, particle 2 in ψ_1^P , which is on shelf 1, particle 3 in ψ_4^P which is on shelf 2, and particle 4 in ψ_8^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 ψ_5^P through ψ_{12}^P for the single-particle eigenfunction on shelf 3 that that particle can go into, times 3 possible choices ψ_2^P through ψ_4^P 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 (5.6). 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 9.5. Now the number of system energy eigenfunctions that have all particles on shelf 1 is not one, but $4^{16} = 4\,294\,967\,296$, 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 $1.4 \cdot 10^{24}$ 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 $1.2 \cdot 10^{77}$, a tremendous number, but totally humiliated by the $2.7 \cdot 10^{138}$ eigenfunctions that have occupation numbers $\vec{I} = (14, 27, 23)$. Taking the ratio of

these two numbers shows that there are $2.3 \cdot 10^{61}$ 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 9.5 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 9.5 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

$$E^S = I_1 E_1^P + I_2 E_2^P + I_3 E_3^P$$

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 9.5 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.

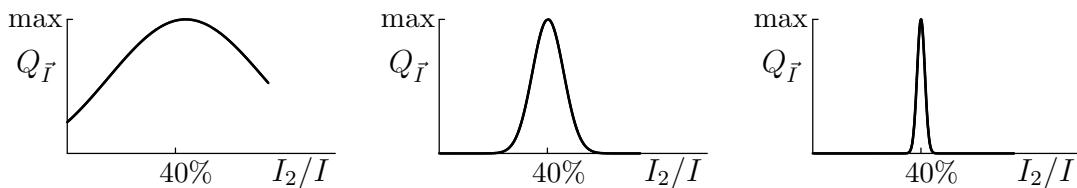


Figure 9.6: Number of energy eigenfunctions on the oblique energy line in 9.5. (The curves are mathematically interpolated to allow a continuously varying fraction of particles on shelf 2.) Left: 4 particles, middle: 64, right: 1,024.

Next assume that the system has exactly the energy of the oblique cut-off line in figure 9.5, with zero uncertainty. The number of energy eigenstates $Q_{\bar{I}}$ on that oblique line is plotted in figure 9.6 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 9.6 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^p , 40% on shelf 2 at energy E_2^p , and 37% on shelf 3 at energy E_3^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, {A.76},

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.

9.4 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 5 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, {A.77}, but they can be worked out, as in note {A.78}.

The final result is, of course, the particle energy distributions from chapter 5:

$$\iota^b = \frac{1}{e^{(E^p - \mu)/k_B T} - 1} \quad \iota^d = \frac{1}{e^{(E^p - \mu)/k_B T}} \quad \iota^f = \frac{1}{e^{(E^p - \mu)/k_B T} + 1}.$$

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 note {A.78} 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 9.14.4. 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 5.2, 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 note {A.84}. Note that for particles like photons that can be readily created or annihilated, there is no chemical potential; μ entered into the

derivation in note {A.78} 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 $e^{(E^P - \mu)/k_B T}$ 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 9.14.4

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.

9.5 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 assumed that the energy was certain. 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 certain 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 certain, 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 energy certainty is much bleaker. Surely a single box in the stack may randomly exchange a bit of energy with the other boxes. If the exchange of energy is completely random, a single box is likely to acquire an uncertainty in its energy equal to the typical exchanged amount times the square root of the number of boxes. That would be an unlimited amount of uncertainty if the number of boxes is made larger and larger. 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 note {A.79}. It turns out that when the temperature T , but not the energy of a system is certain, the system energy eigenfunctions ψ_q^S can be assigned probabilities of the form

$$P_q = \frac{1}{Z} e^{-E_q^S/k_B T} \quad (9.4)$$

where $k_B = 1.380\,65\,10^{-23}$ J/K is the Boltzmann constant. This equation for the probabilities is called the Gibbs “canonical probability distribution.” Feynman [13, 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

$$Z = \sum_{\text{all } q} e^{-E_q^S/k_B T} \quad (9.5)$$

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^S is a macroscopic energy, while the energy $k_B T$ 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.

So, what happens to the simple model system from section 9.3 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 9.7. 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 9.7, but it is no longer limiting energy; it is merely the energy at the most probable shelf occupation numbers.

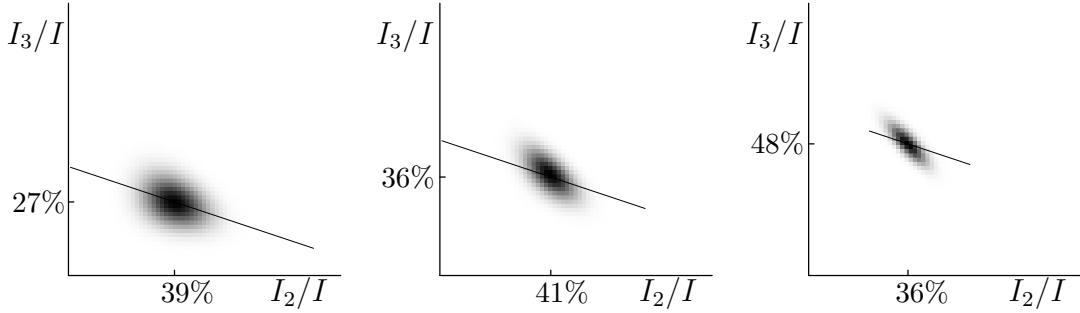


Figure 9.7: Probabilities of shelf-number sets for the simple 64 particle model system if there is uncertainty in energy. More probable shelf-number distributions are shown darker. Left: identical bosons, middle: distinguishable particles, right: identical fermions. The temperature is the same as in figure 9.5.

Equivalently, it is the “expectation energy” of the system, defined following the ideas of chapter 3.3.1 as

$$\langle E \rangle \equiv \sum_{\text{all } q} P_q E_q^S \equiv E$$

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 9.7 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, $k_B T = 1.85$ in the arbitrary energy units.

9.6 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 5. To do so, however, the first shelf must be taken to contain just a single, non degenerate ground state.

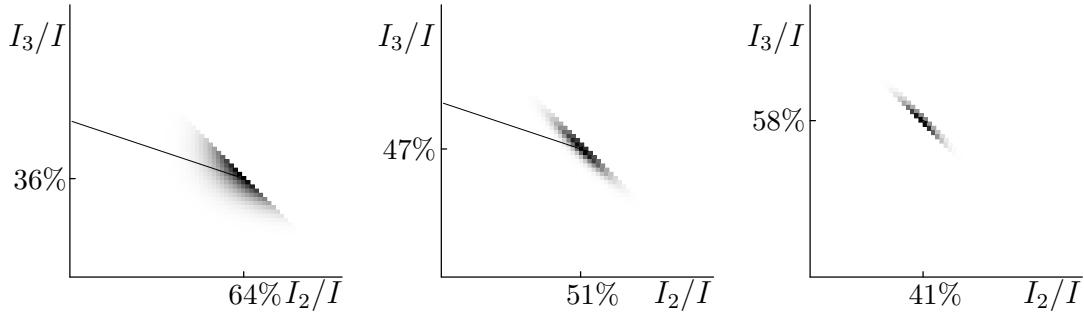


Figure 9.8: Probabilities of shelf-number sets for the simple 64 particle model system if shelf 1 is a non-degenerate ground state. Left: identical bosons, middle: distinguishable particles, right: identical fermions. The temperature is the same as in figure 9.7.

In that case, figure 9.7 of the previous section turns into figure 9.8. 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.

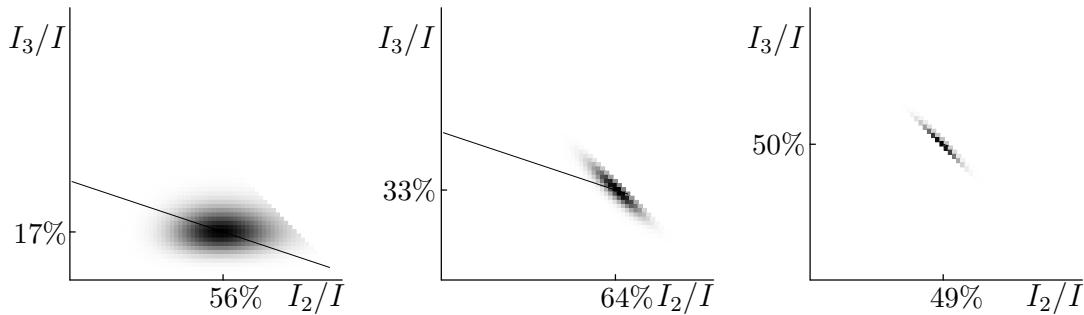


Figure 9.9: Like figure 9.8, but at a lower temperature.

If the temperature is lowered however, as in figure 9.9 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.

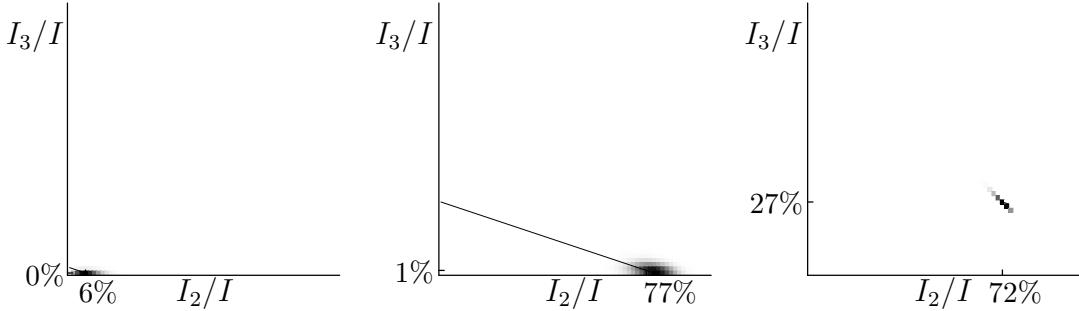


Figure 9.10: Like figure 9.8, but at a still lower temperature.

When the temperature is lowered still much lower, as shown in figure 9.10, 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 9.10 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 9.9.)

9.7 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

$$I_A = 6.0221 \cdot 10^{23}$$

That produces the “molar density” $\bar{\rho} \equiv I/I_A V$ and “molar specific volume” $\bar{v} \equiv VI_A/I$. 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 non-molar specific quantities is called the “molecular mass” M ; it is applied according to its dimensions of kg/kmol. The numerical value of the molecular 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,

$$E = \sum_{\text{all } q} P_q E_q^S \quad (9.6)$$

where P_q is the canonical probability of (9.4), (9.5). 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 (9.5) is differentiated with respect to temperature, you get

$$\left(\frac{\partial Z}{\partial T} \right)_{V \text{ constant}} = \frac{1}{k_B T^2} \sum_{\text{all } q} E_q^S e^{-E_q^S/k_B T}.$$

(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

$$E = k_B T^2 \left(\frac{\partial \ln Z}{\partial T} \right)_{V \text{ constant}} \quad (9.7)$$

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^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^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 dE_q^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

$$-P dV.$$

(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 $P = -dE_q^S/dV$. For a real system with uncertainty in energy and energy eigenfunction, the pressure is defined as the expectation value:

$$P = - \sum_{\text{all } q} P_q \frac{dE_q^S}{dV} \quad (9.8)$$

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:

$$P = k_B T \left(\frac{\partial \ln Z}{\partial V} \right)_{T \text{ constant}} \quad (9.9)$$

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 $-P dV$. 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

$$H = E + PV \quad (9.10)$$

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:

$$dE = \delta Q - P dV \quad (9.11)$$

In words, the internal energy of the system changes by the amount δQ of heat added plus the amount $-P dV$ 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 dE and dV 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,

$$de = \delta q - P dv,$$

it means that C_v is defined as $\delta q/dT$ when $dv = 0$. 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:

$$C_v \equiv \left(\frac{\partial e}{\partial T} \right)_v \quad (9.12)$$

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 $de + P dv$ and that is the change in enthalpy $h = e + Pv$. There is the first practical application of the enthalpy already! It follows that

$$C_p \equiv \left(\frac{\partial h}{\partial T} \right)_P \quad (9.13)$$

9.8 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 9.10, 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.

9.9 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.

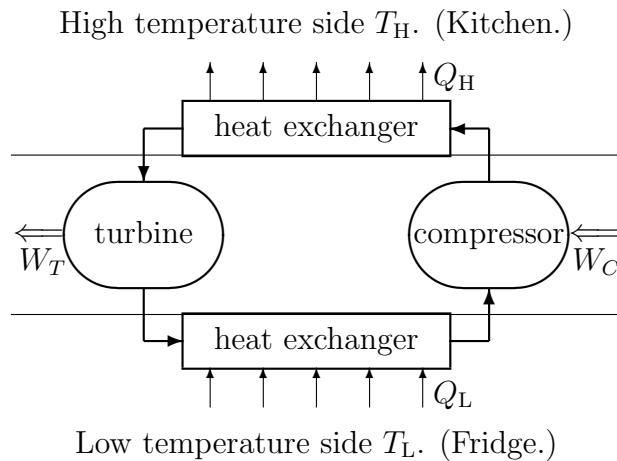


Figure 9.11: Schematic of the Carnot refrigeration cycle.

The theoretically simple refrigeration device is called a "Carnot cycle" refrigeration device, or Carnot heat pump. A schematic is shown in figure 9.11. 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 $-P dV$, 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, {A.80}, a very simple result is found. The ratio of the heat Q_H dumped by the device into the kitchen to the heat Q_L removed from the refrigerator is exactly the same as the ratio of the temperature of the kitchen T_H to that of the fridge T_L :

$$\boxed{\text{For an ideal cycle: } \frac{Q_H}{Q_L} = \frac{T_H}{T_L}} \quad (9.14)$$

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_H and Q_L . A “coefficient of performance” can be defined that is the ratio of the heat Q_L removed from the fridge to the required power input W :

$$\boxed{\text{For an ideal refrigeration cycle: } \beta \equiv \frac{Q_L}{W} = \frac{T_L}{T_H - T_L}} \quad (9.15)$$

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 9.11 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

$$\boxed{\text{For an ideal heat pump: } \beta' \equiv \frac{Q_H}{W} = \frac{T_H}{T_H - T_L}} \quad (9.16)$$

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!

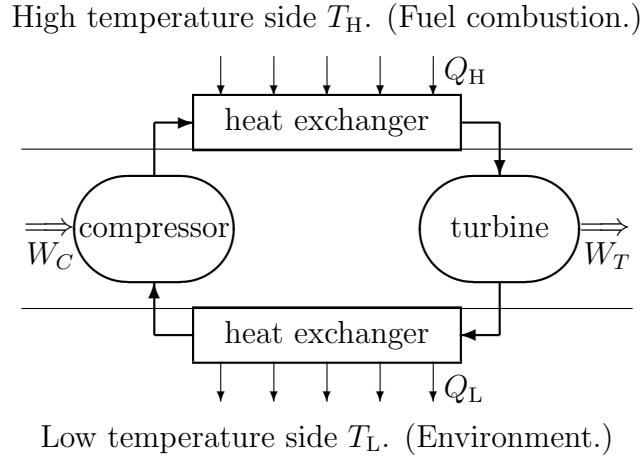


Figure 9.12: Schematic of the Carnot heat engine.

If you run the Carnot refrigeration cycle in reverse, as in figure 9.12, all arrows reverse and it turns into a “heat engine.” The device now takes in 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, it are the exhaust and radiator 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 non-reversed version. But the purpose is now to create work, so the “thermal efficiency” of a heat engine is defined as

$$\boxed{\text{For an ideal heat engine: } \eta_{\text{th}} \equiv \frac{W}{Q_H} = \frac{T_H - T_L}{T_H}} \quad (9.17)$$

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_H at the high temperature side, and slightly hotter than T_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_H and Q_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

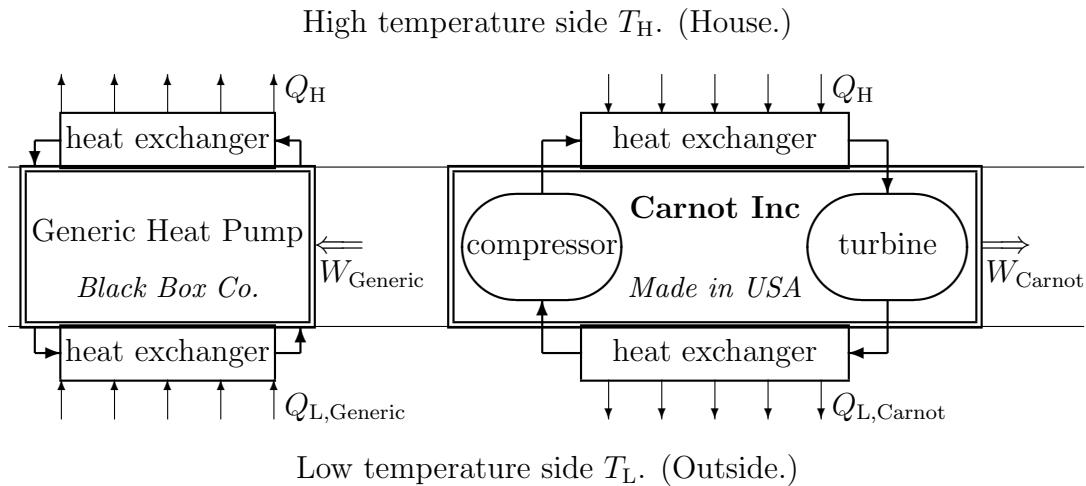


Figure 9.13: A generic heat pump next to a reversed Carnot one with the same heat delivery.

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 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 9.13. 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_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_H is a measure of the amount of fuel burned, so the non-reversible 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.

9.10 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 sys-

tems, 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 δQ_H of heat at a temperature T_H and give up an infinitesimal amount δQ_L at a temperature T_L . This is done so that $\delta Q_H/\delta Q_L = T_H/T_L$, or separating the two ends of the device, $\delta Q_H/T_H = \delta Q_L/T_L$. 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:

$$S \equiv S_{\text{ref}} + \int_{\text{reference state}}^{\text{desired state}} \frac{\delta Q}{T} \quad \text{along any reversible path} \quad (9.18)$$

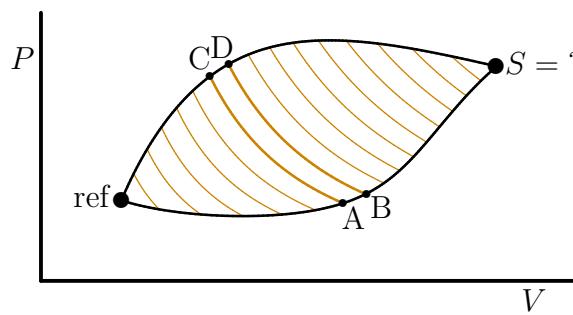


Figure 9.14: Comparison of two different integration paths for finding the entropy of a desired state. The two different integration paths are in black and the yellow lines are reversible adiabatic process lines.

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 9.14. 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 δQ_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:

$$\boxed{dS \geq \frac{\delta Q}{T}} \quad (9.19)$$

where the = applies if the change is reversible and > if it is not.

Note that the above formula is only valid if the system has an unambiguous 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 (9.18) 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 non equilibrium 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, {A.81}.

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 (9.18) 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 ψ_q^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 $\sum P_q E_q^S$, 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. Your 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 $k_B = 1.380\,65\,10^{-23}\text{ J/K}$; 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:

$$S = -k_B \sum P_q \ln(P_q)$$

(9.20)

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.82}.

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 note {A.83}, but here are the results. For systems in thermal equilibrium, is the entropy the same as the one given by the classical integration (9.18)? 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 [13, p. 8] gives an argument to show that the entropy of an isolated system always increases with time. Taking the time derivative of (9.20),

$$\frac{dS}{dt} = -k_B \sum_q [\ln(P_q) + 1] \frac{dP_q}{dt} = -k_B \sum_q \sum_r [\ln(P_q) + 1] R_{qr} [P_r - P_q],$$

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

$$k_B [\ln(P_r) + 1 - \ln(P_q) - 1] R_{qr} [P_r - P_q]$$

and that is always greater than 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 6.3

you might consider this more a validation of time dependent perturbation theory than of the expression for entropy.

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.

9.11 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 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.

9.12 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 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_{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”

$$F = E - TS \quad (9.21)$$

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 $P_{\text{surr}}(V_2 - V_1)$. 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”

$$G = H - TS \quad (9.22)$$

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 (9.11) with the differential second law (9.19) for reversible processes:

$$dE = T dS - P dV \quad (9.23)$$

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

$$\boxed{dH = T dS + V dP} \quad (9.24)$$

The differentials of the Helmholtz and Gibbs free energies are, after cleaning up with the two expressions immediately above:

$$\boxed{dF = -S dT - P dV} \quad (9.25)$$

and

$$\boxed{dG = -S dT + V dP} \quad (9.26)$$

Expression (9.25) 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 $P dV$ into $-V dP$ 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 (9.23) through (9.26) must be the partial derivatives of the quantity in the left hand side:

$$\left(\frac{\partial E}{\partial S} \right)_V = T \quad \left(\frac{\partial E}{\partial V} \right)_S = -P \quad \left(\frac{\partial T}{\partial V} \right)_S = - \left(\frac{\partial P}{\partial S} \right)_V \quad (9.27)$$

$$\left(\frac{\partial H}{\partial S} \right)_P = T \quad \left(\frac{\partial H}{\partial P} \right)_S = V \quad \left(\frac{\partial T}{\partial P} \right)_S = \left(\frac{\partial V}{\partial S} \right)_P \quad (9.28)$$

$$\left(\frac{\partial F}{\partial T} \right)_V = -S \quad \left(\frac{\partial F}{\partial V} \right)_T = -P \quad \left(\frac{\partial S}{\partial V} \right)_T = \left(\frac{\partial P}{\partial T} \right)_V \quad (9.29)$$

$$\left(\frac{\partial G}{\partial T} \right)_P = -S \quad \left(\frac{\partial G}{\partial P} \right)_T = V \quad \left(\frac{\partial S}{\partial P} \right)_T = - \left(\frac{\partial V}{\partial T} \right)_P \quad (9.30)$$

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 (9.23) and (9.24) by T and rearranging so that S becomes the quantity differenced. That produces

$$\left(\frac{\partial S}{\partial T}\right)_V = \frac{1}{T} \left(\frac{\partial E}{\partial T}\right)_V \quad \left(\frac{\partial S}{\partial V}\right)_T = \frac{1}{T} \left(\frac{\partial E}{\partial V}\right)_T + \frac{P}{T}$$

$$\left(\frac{\partial E}{\partial V}\right)_T = T^2 \left(\frac{\partial P/T}{\partial T}\right)_V \quad (9.31)$$

$$\left(\frac{\partial S}{\partial T}\right)_P = \frac{1}{T} \left(\frac{\partial H}{\partial T}\right)_P \quad \left(\frac{\partial S}{\partial P}\right)_T = \frac{1}{T} \left(\frac{\partial H}{\partial P}\right)_T - \frac{V}{T}$$

$$\left(\frac{\partial H}{\partial P}\right)_T = -T^2 \left(\frac{\partial V/T}{\partial T}\right)_P \quad (9.32)$$

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

$Pv = RT \quad \text{with} \quad R = \frac{k_B}{m} = \frac{R_u}{M} \quad R_u = 8.314\,472 \frac{\text{kJ}}{\text{kmol K}}$

(9.33)

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” $R_u = I_A k_B$ and the molecular mass $M = I_A m$. 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:

$\text{For ideal gases: } e, h, C_v, C_p = e, h, C_v, C_p(T) \quad C_P = C_v + R$

(9.34)

(The final relation is because $C_P = dh/dT = d(e + Pv)/dT$ with $de/dT = C_v$ 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 (9.30) 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 (9.27). Turned on its head, it says that

$$\boxed{\frac{1}{T} = \left(\frac{\partial S}{\partial E} \right)_V \text{ and other external parameters fixed}} \quad (9.35)$$

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 the 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

$$\boxed{\bar{\mu}_c \equiv \left(\frac{\partial G}{\partial n_c} \right)_{P,T}} \quad (9.36)$$

(If there is only one constituent, then $G = \bar{g}$ 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

$$G = \bar{\imath}_1 \bar{g}_1 + \bar{\imath}_2 \bar{g}_2 + \dots = \sum_c \bar{\imath}_c \bar{g}_c,$$

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

$$dG(P, T, \bar{\imath}_1, \bar{\imath}_2, \dots) = \frac{\partial G}{\partial T} dT + \frac{\partial G}{\partial P} dP + \frac{\partial G}{\partial \bar{\imath}_1} d\bar{\imath}_1 + \frac{\partial G}{\partial \bar{\imath}_2} d\bar{\imath}_2 + \dots$$

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,

$$dG = -S dT + V dP + \bar{\mu}_1 d\bar{\imath}_1 + \bar{\mu}_2 d\bar{\imath}_2 + \dots = -S dT + V dP + \sum_c \bar{\mu}_c d\bar{\imath}_c$$

(9.37)

This generalizes (9.26) 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 dG must be zero whenever $dT = dP = 0$, 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 $d\bar{\imath}_c$ in the amounts of the constituents. For example, in a liquid-vapor “phase equilibrium,” any additional amount of particles $d\bar{\imath}_f$ that condenses to liquid must equal the amount $-d\bar{\imath}_g$ 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 (9.37) it can be seen that the liquid and vapor phase must have the same chemical potential, $\bar{\mu}_f = \bar{\mu}_g$. 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, $d\mu_f = d\mu_g$, and substituting in the differential (9.26) for the Gibbs free energy, taking it on a molar basis since $\bar{\mu} = \bar{g}$,

$$-\bar{s}_f dT + \bar{v}_f dP = -\bar{s}_g dT + \bar{v}_g dP$$

and rearranging gives the Clausius-Clapeyron equation:

$$\frac{dP}{dT} = \frac{s_g - s_f}{v_g - v_f}$$

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 $s_{fg} = s_g - s_f$ and $v_{fg} = v_g - v_f$, making the formula above very convenient.

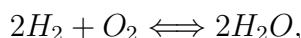
In case your tables do not have the entropies of the liquid and vapor phases, they often still have the “latent heat of vaporization,” also known as “enthalpy of vaporization” or similar, and in engineering thermodynamics books typically indicated by h_{fg} . That is the difference between the enthalpy of the saturated liquid and vapor phases, $h_{fg} = h_g - h_f$. If saturated liquid is turned into saturated vapor by adding heat under conditions of constant pressure and temperature, (9.24) shows that the change in enthalpy $h_g - h_f$ equals $T(s_g - s_f)$. So the Clausius-Clapeyron equation can be rewritten as

$$\frac{dP}{dT} = \frac{h_{fg}}{T(v_g - v_f)}$$

(9.38)

Because $T ds$ 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



the changes in the amounts of the constituents are related as

$$d\bar{n}_{H_2} = -2d\bar{r} \quad d\bar{n}_{O_2} = -1d\bar{r} \quad d\bar{n}_{H_2O} = 2d\bar{r}$$

where $d\bar{r}$ 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 $d\bar{r}$ 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 $dF = dG - PdV - VdP$, and substituting for dG from (9.37), that gives

$$dF = -SdT - PdV + \bar{\mu}_1 d\bar{v}_1 + \bar{\mu}_2 d\bar{v}_2 + \dots = -SdT - PdV + \sum_c \bar{\mu}_c d\bar{v}_c \quad (9.39)$$

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 $F = \bar{v}\bar{f}(T, \bar{v})$, but $(\partial \bar{v}\bar{f}(T, V/\bar{v}) / \partial \bar{v})_{TV}$ does not simply produce \bar{f} , even if $(\partial \bar{v}\bar{g}(T, P) / \partial \bar{v})_{TP}$ produces \bar{g} .

9.13 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 $P_q = e^{-E_q^S/kT} / Z$ can be used for the probabilities of the energy eigenfunctions.

Consider first the Helmholtz free energy:

$$F = E - TS = \sum_q P_q E_q^S + T k_B \sum_q P_q \ln(e^{-E_q^S/k_B T} / Z)$$

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

$$F = -k_B T \ln Z \quad (9.40)$$

That makes strike three for the partition function Z , since it already was able to produce the internal energy E , (9.7), and the pressure P , (9.9). 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 (9.9):

$$G = -k_B T \left[\ln Z - V \left(\frac{\partial \ln Z}{\partial V} \right)_T \right] \quad (9.41)$$

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 note {A.84} make three. Please note that whatever this book tells you thrice is absolutely true.

9.14 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 9.1 through 9.3 to be filled in for a concrete case.

For particles in a macroscopic box, the single-particle energy levels E^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 dE^p is approximately, following (5.6),

$$dN = V n_s \mathcal{D} dE^p = V \frac{n_s}{4\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \sqrt{E^p} dE^p \quad (9.42)$$

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^p is

$$\iota = \frac{1}{e^{(E^p - \mu)/k_B T} \pm 1}$$

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:

$$I = \int_{E^p=0}^{\infty} \iota V n_s \mathcal{D} dE^p = V \frac{n_s}{4\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \int_{E^p=0}^{\infty} \frac{\sqrt{E^p}}{e^{(E^p-\mu)/k_B T} \pm 1} dE^p$$

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:

$$E = \int_{E^p=0}^{\infty} E^p \iota n_s V \mathcal{D} dE^p = V \frac{n_s}{4\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \int_{E^p=0}^{\infty} \frac{E^p \sqrt{E^p}}{e^{(E^p-\mu)/k_B T} \pm 1} dE^p$$

The expression for the number of particles can be nondimensionalized by rearranging and taking a root to give

$$\frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3} = \left(\frac{n_s}{4\pi^2} \int_{u=0}^{\infty} \frac{\sqrt{u} du}{e^{u-u_0} \pm 1} \right)^{2/3} \quad u \equiv \frac{E^p}{k_B T} \quad u_0 \equiv \frac{\mu}{k_B T} \quad (9.43)$$

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 $k_B T$. 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” λ_{th} by writing the classical microscopic energy $k_B T$ in a quantum-like way:

$$k_B T \equiv 4\pi \frac{\hbar^2}{2m} \frac{1}{\lambda_{\text{th}}^2}$$

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

$$\frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3} = \frac{1}{4\pi} \left[\frac{\lambda_{\text{th}}}{(V/I)^{1/3}} \right]^2$$

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 (9.43) 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 $k_B T$ on the other side. While the two energy ratios are not the same, (9.43) 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

$$\boxed{\frac{E}{Ik_B T} = \int_{u=0}^{\infty} \frac{u\sqrt{u} du}{e^{u-u_0} \pm 1} \Big/ \int_{u=0}^{\infty} \frac{\sqrt{u} du}{e^{u-u_0} \pm 1} \quad u \equiv \frac{E^p}{k_B T} \quad u_0 \equiv \frac{\mu}{k_B T}} \quad (9.44)$$

The integral in the bottom arises when getting rid of the ratio of energies that forms using (9.43).

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 $k_B T$. It too is a unique function of u_0 , and as a consequence, also of the ratio of typical microscopic quantum and classical energies.

9.14.1 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 note {A.78}, the number of particles in the ground state is given by

$$I_1 = \frac{N_1 - 1}{e^{(E_1^p - \mu)/k_B T} - 1}. \quad (9.45)$$

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^p . To be precise, $E_1 - \mu$ must be small of order $k_B T/I$; 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^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 $\mu \approx E_1^p \approx 0$, the approximations being extremely close. If the ground state is unique, $N_1 = 1$, Bose-Einstein condensation simply occurs when $\mu = E_1^p \approx 0$.

You would therefore expect that you can simply put $u_0 = \mu/k_B T$ to zero in the integrals (9.43) and (9.44). However, if you do so (9.43) 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:

$$\frac{\frac{\hbar^2}{2m} \left(\frac{I - I_1}{V} \right)^{2/3}}{k_B T} = \left(\frac{n_s}{4\pi^2} \int_{u=0}^{\infty} \frac{\sqrt{u} du}{e^u - 1} \right)^{2/3} \quad \text{for BEC} \quad (9.46)$$

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 $\sqrt{u_1} = \sqrt{E_1^p/k_B T}$ is vanishingly small, not finite. Arguments given in note {A.78} 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, [28, 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

$$\frac{\frac{\hbar^2}{2m} \left(\frac{I - I_1}{V} \right)^{2/3}}{k_B T} = \frac{\frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3}}{k_B T_B} = \left(\frac{n_s}{8\pi^{3/2}} \zeta\left(\frac{3}{2}\right) \right)^{2/3} \quad T \leq T_B \quad (9.47)$$

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 $k_B T_B$. 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:

$$\frac{I_1}{I} = 1 - \left(\frac{T}{T_B} \right)^{3/2} \quad T \leq T_B \quad (9.48)$$

9.14.2 Fermions at low temperatures

Another application of the integrals (9.43) and (9.44) is to find the Fermi energy E_F^p and internal energy E of a system of weakly interacting fermions for vanishing temperature.

For low temperatures, the nondimensional energy ratio $u_0 = \mu/k_B T$ blows up, since $k_B T$ becomes zero and the chemical potential μ does not; μ becomes the Fermi energy E_F^p , chapter 5.10. To deal with the blow up, the integrals can be rephrased in terms of $u/u_0 = E^p/\mu$, which does not blow up.

In particular, the ratio (9.43) 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:

$$\frac{\frac{\hbar^2}{2m} \left(\frac{I}{V}\right)^{2/3}}{\mu} = \left(\frac{n_s}{4\pi^2} \int_{u/u_0=0}^{\infty} \frac{\sqrt{u/u_0} d(u/u_0)}{e^{u_0[(u/u_0)-1]} + 1} \right)^{2/3}$$

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

$$\frac{\frac{\hbar^2}{2m} \left(\frac{I}{V}\right)^{2/3}}{\mu} = \left(\frac{n_s}{4\pi^2} \int_{u/u_0=0}^1 \sqrt{u/u_0} d(u/u_0) \right)^{2/3} = \left(\frac{n_s}{6\pi^2} \right)^{2/3}$$

It follows that the Fermi energy is

$$E_F^p = \mu|_{T=0} = \left(\frac{6\pi^2}{n_s} \right)^{2/3} \frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3}$$

Physicists like to define a “Fermi temperature” as the temperature where the classical microscopic energy $k_B T$ becomes equal to the Fermi energy. It is

$$T_F = \frac{1}{k_B} \left(\frac{6\pi^2}{n_s} \right)^{2/3} \frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3} \quad (9.49)$$

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 (9.44) can be integrated in the same way and then shows that $E = \frac{3}{5} I \mu = \frac{3}{5} I E_F^p$. In short, at absolute zero, the average energy per particle is $\frac{3}{5}$ times E_F^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 5.10. However, the analysis in this subsection can be used to find the corrected expressions when the temperature is fairly small but not zero, {A.85}, 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:

$$C_v = \frac{\pi^2 k_B T}{2 E_F^p} \frac{k_B}{m} (1 + \dots) \quad (9.50)$$

where k_B/m is the gas constant R . All low-temperature expansions proceed in powers of $(k_B T/E_F^p)^2$, 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.

9.14.3 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 have no anti-derivatives among the normal functions?

Well, there is at least one. By definition, (9.8), the pressure is the expectation value of $-dE_q^S/dV$ where the E_q^S are the system energy eigenvalues. For weakly interacting particles in a box, chapter 5.2 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: $E_q^S = \text{const } V^{-2/3}$. Differentiating produces $dE_q^S/dV = -\frac{2}{3}E_q^S/V$ and taking the expectation value

$$PV = \frac{2}{3}E \quad (9.51)$$

This expression is valid for weakly interacting bosons and fermions even if the (anti)symmetrization requirements cannot be ignored.

9.14.4 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

$$\iota = \frac{1}{e^{(E^p - \mu)/k_B T} \pm 1}$$

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 (9.44) can now easily be done; the e^{u_0} factor divides away and an integration by parts in the numerator produces $E = \frac{3}{2}Ik_B T$. Plug it into the generalized ideal gas law (9.51) to get the normal “ideal gas law”

$$\boxed{PV = Ik_B T \quad \Longleftrightarrow \quad Pv = RT \quad R \equiv \frac{k_B}{m}} \quad (9.52)$$

Also, following (9.34),

$$e = \frac{3}{2} \frac{k_B}{m} T = C_v T \quad h = \frac{5}{2} \frac{k_B}{m} T = C_p T \quad C_v = \frac{3}{2} R \quad C_p = \frac{5}{2} R$$

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 9.15.

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 (9.43) 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 :

$$-\frac{\mu}{k_B T} = \frac{3}{2} \ln \left[k_B T \sqrt[3]{4\pi n_s^{-2/3} \frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3}} \right]$$

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

$$\boxed{s = C_v \ln \left[k_B T \sqrt[3]{4\pi n_s^{-2/3} \frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3}} \right] + C_p} \quad (9.53)$$

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.

9.14.5 Blackbody radiation

This section takes a closer look at blackbody radiation, discussed earlier in chapter 5.8. 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 {A.78} and {A.79}.

The second caveat is that the usual density of states (5.6) is nonrelativistic. It does not apply to photons, which move at the speed of light. For photons you must use the density of modes (5.7).

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, chapter 12.2.3. 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, (10.28); 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 (5.7).

The energy per unit box volume and unit frequency range found under the above caveats is Planck's blackbody spectrum already given in chapter 5.8:

$$\rho(\omega) \equiv \frac{d(E/V)}{d\omega} = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1} \quad (9.54)$$

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 5.8:

$$\boxed{\frac{E}{V} = \frac{\pi^2}{15\hbar^3 c^3} (k_B T)^4} \quad (9.55)$$

The number of particles may be found similar to the energy, by dropping the $\hbar\omega$ energy per particle from the integral. It is, [28, 36.24, with typo]:

$$\boxed{\frac{I}{V} = \frac{2\zeta(3)}{\pi^2 \hbar^3 c^3} (k_B T)^3} \quad \zeta(3) \approx 1.202 \quad (9.56)$$

Taking the ratio with (9.55), the average energy per photon may be found:

$$\boxed{\frac{E}{I} = \frac{\pi^4}{30\zeta(3)} k_B T \approx 2.7 k_B T} \quad (9.57)$$

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 $\int dE/T$ using (9.55), starting from absolute zero and keeping the volume constant:

$$\boxed{\frac{S}{V} = \frac{4\pi^2}{45\hbar^3 c^3} k_B (k_B T)^3} \quad (9.58)$$

Dividing by (9.56) shows the average entropy per photon to be

$$\frac{S}{I} = \frac{2\pi^4}{45\zeta(3)} k_B \quad (9.59)$$

independent of temperature.

The generalized ideal gas law (9.51) 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:

$$\boxed{PV = \frac{1}{3}E} \quad (9.60)$$

9.14.6 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 ω_{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 $\hbar\omega_{max}$

$$\boxed{k_B T_D = \hbar\omega_{max}} \quad (9.61)$$

The expression for the internal energy becomes, from (5.11) times 1.5:

$$\boxed{\frac{E}{V} = \int_0^{\omega_{max}} \frac{3\hbar}{2\pi^2 c_s^3} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1} d\omega} \quad (9.62)$$

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 (5.7), 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 (5.7) is used at high temperatures, the law of Dulong and Petit results, as described in section 9.15.

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.

9.15 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 9.14.4, shows that for an ideal gas, the molecules have $\frac{1}{2}k_B T$ of translational kinetic energy in each of the three directions of a Cartesian coordinate system, where $k_B = 1.38 \cdot 10^{-23}$ J/K is Boltzmann's constant. So the specific heat per molecule is $\frac{3}{2}k_B$. For a kmol ($6.02 \cdot 10^{26}$) of molecules instead of one, k_B becomes the “universal gas constant” $R_u = 8.31$ kJ/kmol K. Hence for a

$$\text{monatomic ideal gas: } \bar{C}_v = \frac{3}{2}R_u = 12.5 \text{ kJ/kmol K} \quad (9.63)$$

on a kmol basis. This is very accurate for all the 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. Classical physics, in particular the “equipartition theorem,” would then predict $\frac{7}{2}k_B$ as the specific heat per molecule; $\frac{3}{2}k_B$ of kinetic energy for each atom, plus $\frac{1}{2}k_B$ of potential energy in the internal vibration of the pairs of atoms towards and away from each other. However, experimental values do not at all agree. (And it is even worse than it looks. The $\frac{7}{2}k_B$ assumes an analysis in terms of the dynamics of simplistic atoms with all their mass in their nuclei, and the vibrations between the pairs of atoms modeled as a harmonic oscillator. As Maxwell noted, if you really take classical theory at face value, things get much worse still, since the individual internal part of the atoms, in particular the electrons, would have to absorb their own thermal energy too.)

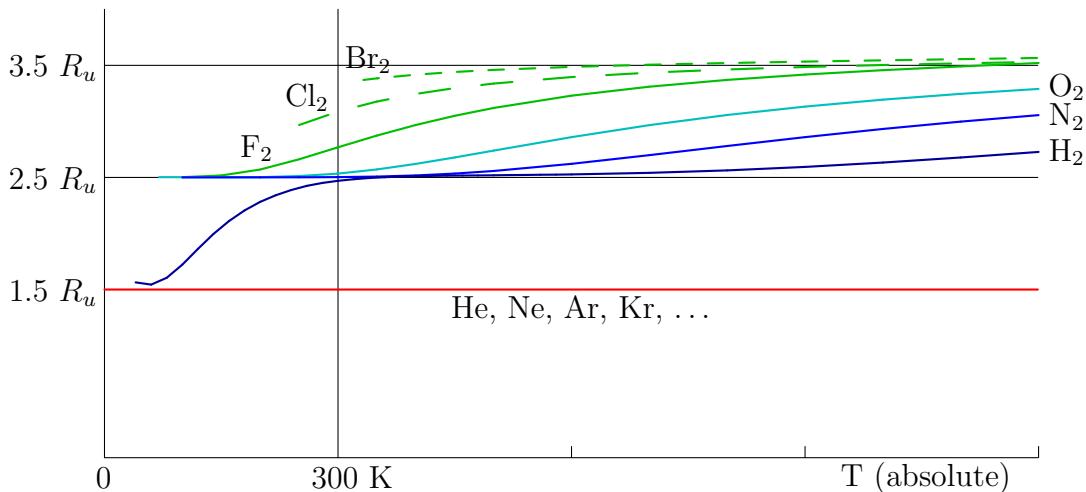


Figure 9.15: Specific heat at constant volume of gases. Temperatures from absolute zero to 1,200 K. Data from NIST-JANAF and AIP.

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 9.15. 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, as well as the harmonic oscillation energy, are quantized. For hydrogen at low temperatures, the typical available thermal energy $\frac{1}{2}k_B T$ is not enough to reach the next level for either. No energy can therefore be put into rotation of the molecule, nor in increased 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 motion are “frozen out.”

At normal temperatures, there is enough thermal energy to reach nonzero angular momentum states, but not higher vibrational ones, and the specific heat becomes

$$\text{typical diatomic ideal gas: } \bar{C}_v = \frac{5}{2}R_u = 20.8 \text{ kJ/kmol K.} \quad (9.64)$$

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 9.15. 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 from the solution of the harmonic oscillator derived in chapter 2.6. The energy levels of an harmonic oscillator are apart by an amount $\hbar\omega$, where ω is the angular frequency. Modeled as a simple spring-mass system, $\omega = \sqrt{c/m}$, where c is the equivalent spring stiffness and m the equivalent mass. So light atoms that are bound tightly will require a lot of energy to reach the second vibrational state. Hydrogen is much lighter than nitrogen or oxygen, explaining the higher temperature before vibration become important for it. The molecular 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 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 9.15 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 $\frac{3}{2}R_u$ per atom in kinetic energy and $\frac{3}{2}R_u$ in potential energy:

$$\text{law of Dulong and Petit: } \bar{C}_v = 3R_u = 25 \text{ kJ/kmol K.} \quad (9.65)$$

Not only is this 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.

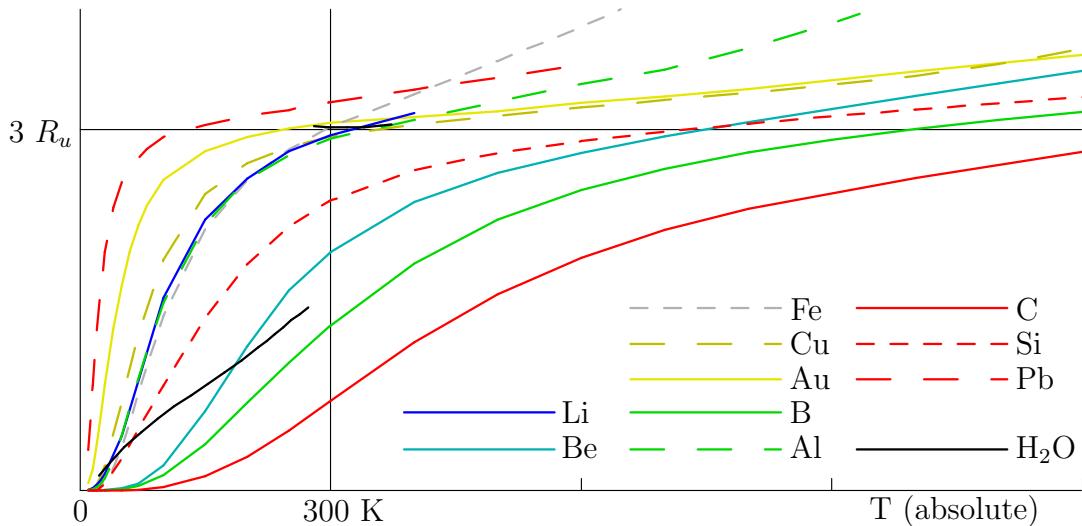


Figure 9.16: Specific heat at constant pressure of solids. Temperatures from absolute zero to 1,200 K. Carbon is diamond; graphite is similar. Water is ice and liquid. Data from NIST-JANAF, CRC, AIP, Rohsenow et al.

Actually, it turns out, figure 9.16, 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 be 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 1860, 1250, and 1000 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 $\frac{5}{2}R_u$ instead of $\frac{7}{2}R_u$. 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, $k_B T$, 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 $k_B T_D$ 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 9.14.6 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 $\frac{3}{2}k_B T$ 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 5.11: 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 9.16.

Chapter 10

Electromagnetism

The main objective of this chapter is to discuss electromagnetic effects. However, these effects are closely tied to more advanced concepts in angular momentum and relativity, so these will be discussed first.

10.1 All About Angular Momentum

The quantum mechanics of angular momentum is fascinating. It is also very basic to much of quantum mechanics, so you may want to browse through this section to get an idea of what is there.

In chapter 4.4, it was already mentioned that angular momentum comes in two basic kinds: orbital angular momentum, which is a result of the motion of particles, and the “built-in” angular momentum called spin.

The eigenfunctions of orbital angular momentum are the so called “spherical harmonics” of chapter 3.1, and they show that the orbital angular momentum in any arbitrarily chosen direction, taken as the z -direction from now on, comes in whole multiples m of Planck’s constant \hbar :

$$L_z = m\hbar \quad \text{with } m \text{ an integer for orbital angular momentum}$$

Integers are whole numbers, such as $0, \pm 1, \pm 2, \pm 3, \dots$. The square orbital angular momentum $L^2 = L_x^2 + L_y^2 + L_z^2$ comes in values

$$L^2 = l(l+1)\hbar^2 \quad \text{with } l \geq 0, \text{ and for orbital angular momentum } l \text{ is an integer.}$$

The numbers l and m are called the azimuthal and magnetic quantum numbers.

When spin angular momentum is included, it is conventional to still write L_z as $m\hbar$ and L^2 as $l(l+1)\hbar^2$, there is nothing wrong with that, but then m and l are no longer necessarily integers. The spin of common particles, such as electrons, neutrons, and protons, instead has $m = \pm \frac{1}{2}$ and $l = \frac{1}{2}$. But while m

and l can be half integers, this section will find that they can never be anything more arbitrary than that, regardless of what sort of angular momentum it is. A particle with, say, spin $\frac{1}{3}\hbar$ cannot not exist according to the theory.

In order to have a consistent notation, from now on every angular momentum eigenstate with quantum numbers l and m will be indicated as $|l\ m\rangle$ whether it is a spherical harmonic Y_l^m , a particle spin state, or a combination of angular momenta from more than one source.

10.1.1 The fundamental commutation relations

Analyzing non-orbital 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 [17, 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.”

$$[\hat{L}_x, \hat{L}_y] = i\hbar\hat{L}_z \quad [\hat{L}_y, \hat{L}_z] = i\hbar\hat{L}_x \quad [\hat{L}_z, \hat{L}_x] = i\hbar\hat{L}_y \quad (10.1)$$

They can be derived for orbital angular momentum (see chapter 3.4.4), but must be *postulated* to also apply to spin angular momentum {A.86}.

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 \hat{L}_x , \hat{L}_y , and \hat{L}_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 \hat{L}_x , \hat{L}_y , and \hat{L}_z have in common are states $|0\ 0\rangle$ of no angular momentum at all {A.87}. 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 \hat{L}_x would turn an eigenfunction of say \hat{L}_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.

10.1.2 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 $|l\ m\rangle$ eigenstates with definite values of square angular momentum L^2 and z -angular momentum L_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 $|l\ m\rangle$ to exist is that the angular momentum operators \hat{L}^2 and \hat{L}_z commute. They do; using the commutator manipulations of chapter 3.4.4), it is easily found that:

$$[\hat{L}^2, \hat{L}_x] = [\hat{L}^2, \hat{L}_y] = [\hat{L}^2, \hat{L}_z] = 0 \quad \text{where } \hat{L}^2 = \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2$$

So mathematics says that eigenstates $|l\ m\rangle$ of \hat{L}_z and \hat{L}^2 exist satisfying

$$\hat{L}_z|l\ m\rangle = L_z|l\ m\rangle \quad \text{where by definition } L_z = m\hbar \tag{10.2}$$

$$\hat{L}^2|l\ m\rangle = L^2|l\ m\rangle \quad \text{where by definition } L^2 = l(l+1)\hbar^2 \text{ and } l \geq 0 \tag{10.3}$$

and that are complete in the sense that any state can be described in terms of these $|l\ m\rangle$.

Unfortunately the eigenstates $|l\ m\rangle$, except for $|0\ 0\rangle$ states, do not satisfy relations like (10.2) for \hat{L}_x or \hat{L}_y . The problem is that \hat{L}_x and \hat{L}_y do not commute with \hat{L}_z . But \hat{L}_x and \hat{L}_y do commute with \hat{L}^2 , and you might wonder if that is still worth something. To find out, multiply, say, the zero commutator $[\hat{L}^2, \hat{L}_x]$ by $|l\ m\rangle$:

$$[\hat{L}^2, \hat{L}_x]|l\ m\rangle = (\hat{L}^2\hat{L}_x - \hat{L}_x\hat{L}^2)|l\ m\rangle = 0$$

Now take the second term to the right hand side of the equation, noting that $\hat{L}^2|l m\rangle = L^2|l m\rangle$ with L^2 just a number that can be moved up-front, to get:

$$\hat{L}^2(\hat{L}_x|l m\rangle) = L^2(\hat{L}_x|l m\rangle)$$

Looking a bit closer at this equation, it shows that the combination $\hat{L}_x|l m\rangle$ satisfies the same eigenvalue problem for \hat{L}^2 as $|l m\rangle$ itself. In other words, the multiplication by \hat{L}_x does not affect the square angular momentum L^2 at all.

To be picky, that is not quite true if $\hat{L}_x|l m\rangle$ 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 $|l m\rangle$ an eigenstate of \hat{L}_x with eigenvalue zero in addition to an eigenstate of \hat{L}_z {A.87}). Except for that trivial case, \hat{L}_x does not affect square angular momentum. And neither does \hat{L}_y or any combination of the two.

Angular momentum in the z -direction is affected by \hat{L}_x and by \hat{L}_y , since they do not commute with \hat{L}_z like they do with \hat{L}^2 . Nor is it possible to find any linear combination of \hat{L}_x and \hat{L}_y that does commute with \hat{L}_z . What is the next best thing? Well, it *is* possible to find two combinations, to wit

$$\hat{L}^+ \equiv \hat{L}_x + i\hat{L}_y \quad \text{and} \quad \hat{L}^- \equiv \hat{L}_x - i\hat{L}_y, \quad (10.4)$$

that satisfy the “commutator eigenvalue problems”:

$$[\hat{L}_z, \hat{L}^+] = \hbar\hat{L}^+ \quad \text{and} \quad [\hat{L}_z, \hat{L}^-] = -\hbar\hat{L}^-.$$

These two turn out to be quite remarkable operators.

Like \hat{L}_x and \hat{L}_y , their combinations \hat{L}^+ and \hat{L}^- leave L^2 alone. To examine what the operator \hat{L}^+ does with the linear momentum in the z -direction, multiply its commutator relation above by an eigenstate $|l m\rangle$:

$$(\hat{L}_z\hat{L}^+ - \hat{L}^+\hat{L}_z)|l m\rangle = \hbar\hat{L}^+|l m\rangle$$

Or, taking the second term to the right hand side of the equation and noting that by definition $\hat{L}_z|l m\rangle = m\hbar|l m\rangle$,

$$\hat{L}_z(\hat{L}^+|l m\rangle) = (m+1)\hbar(\hat{L}^+|l m\rangle)$$

That is a stunning result, as it shows that $\hat{L}^+|lm\rangle$ is an eigenstate with z angular momentum $L_z = (m+1)\hbar$ instead of $m\hbar$. In other words, \hat{L}^+ adds exactly one unit \hbar to the z -angular momentum, turning an $|l m\rangle$ state into a $|l m+1\rangle$ one!

If you apply \hat{L}^+ another time, you get a state of still higher z -angular momentum $|l m+2\rangle$, and so on, like the rungs on a ladder. This is graphically illustrated for some examples in figures 10.1 and 10.2. The process eventually

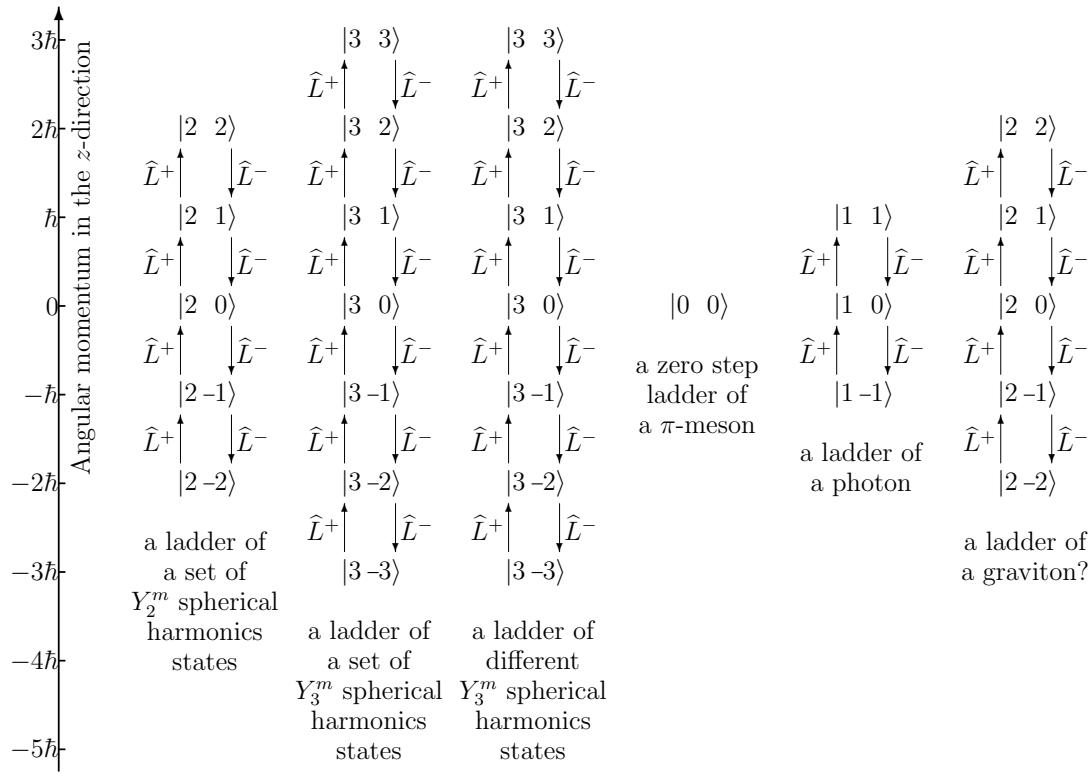


Figure 10.1: Example bosonic ladders.

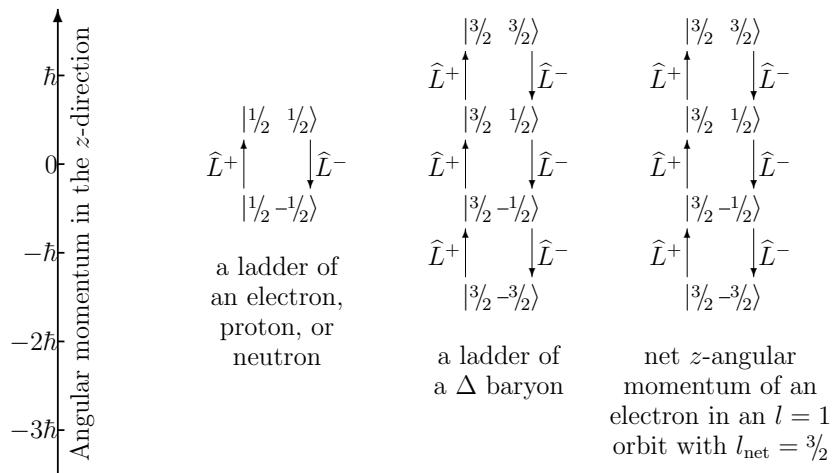


Figure 10.2: Example fermionic ladders.

comes to an halt at some top rung $m = m_{\max}$ where $\hat{L}^+|l m_{\max}\rangle = 0$. 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 {A.88}.

The second “ladder operator” \hat{L}^- 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. \hat{L}^- provides the second stile to the ladders, and must terminate at some bottom rung m_{\min} .

10.1.3 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 l 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 $-l$ to $+l$ in unit increments. In other words, the bosonic and fermionic example ladders in figures 10.1 and 10.2 are representative of all that is possible.

To start, in order for a ladder to end at a top rung m_{\max} , $\hat{L}^+|l m\rangle$ has to be zero for $m = m_{\max}$. More specifically, its magnitude $|\hat{L}^+|l m\rangle|$ must be zero. The square magnitude is given by the inner product with itself:

$$|\hat{L}^+|l m\rangle|^2 = \langle \hat{L}^+|l m\rangle | \hat{L}^+|l m\rangle \rangle = 0.$$

Now because of the complex conjugate that is used in the left hand side of an inner product, (see chapter 1.3), $\hat{L}^+ = \hat{L}_x + i\hat{L}_y$ goes to the other side of the product as $\hat{L}^- = \hat{L}_x - i\hat{L}_y$, and you must have

$$|\hat{L}^+|l m\rangle|^2 = \langle |l m\rangle | \hat{L}^- \hat{L}^+ |l m\rangle \rangle$$

That operator product can be multiplied out:

$$\hat{L}^- \hat{L}^+ \equiv (\hat{L}_x - i\hat{L}_y)(\hat{L}_x + i\hat{L}_y) = \hat{L}_x^2 + \hat{L}_y^2 + i(\hat{L}_x \hat{L}_y - \hat{L}_y \hat{L}_x),$$

but $\hat{L}_x^2 + \hat{L}_y^2$ is the square angular momentum \hat{L}^2 except for \hat{L}_z^2 , and the term within the parentheses is the commutator $[\hat{L}_x, \hat{L}_y]$ which is according to the fundamental commutation relations equal to $i\hbar\hat{L}_z$, so

$$\hat{L}^- \hat{L}^+ = \hat{L}^2 - \hat{L}_z^2 - \hbar\hat{L}_z \tag{10.5}$$

The effect of each of the operators in the left hand side on a state $|l m\rangle$ is known and the inner product can be figured out:

$$|\hat{L}^+|l m\rangle|^2 = l(l+1)\hbar^2 - m^2\hbar^2 - m\hbar^2 \tag{10.6}$$

The question where angular momentum ladders end can now be answered:

$$l(l+1)\hbar^2 - m_{\max}^2 \hbar^2 - m_{\max} \hbar^2 = 0$$

There are two possible solutions to this quadratic equation for m_{\max} , to wit $m_{\max} = l$ or $-m_{\max} = l + 1$. The second solution is impossible since it already would have the square z -angular momentum exceed the total square angular momentum. So unavoidably,

$$m_{\max} = l.$$

That is one of the things this section was supposed to show.

The lowest rung on the ladder goes the same way; you get

$$\hat{L}^+ \hat{L}^- = \hat{L}^2 - \hat{L}_z^2 + \hbar \hat{L}_z \quad (10.7)$$

and then

$$|\hat{L}^-|l m\rangle|^2 = l(l+1)\hbar^2 - m^2 \hbar^2 + m \hbar^2 \quad (10.8)$$

and the only acceptable solution for the lowest rung on the ladders is

$$m_{\min} = -l.$$

It is nice and symmetric; ladders run from $m = -l$ up to $m = l$, as the examples in figures 10.1 and 10.2 already showed.

And in fact, it is more than that; it also limits what the quantum numbers l 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:

$$\text{total number of steps} = m_{\max} - m_{\min} = 2l$$

But the number of steps is a whole number, and so the azimuthal quantum l must either be a nonnegative integer, such as $0, 1, 2, \dots$, or half of one, such as $\frac{1}{2}, \frac{3}{2}, \dots$. Integer l 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 l is a half-integer, then so are the corresponding values of m , since m starts from $-l$ and increases in unit steps. See again figures 10.1 and 10.2 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 10.1 with \hat{L}^+ and then come down the second one using \hat{L}^- . The reason is that the states $|l m\rangle$ are eigenstates of the operators $\hat{L}^- \hat{L}^+$, (10.5), and $\hat{L}^+ \hat{L}^-$, (10.7), so going up with \hat{L}^+ and then down again with \hat{L}^- , 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.

10.1.4 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 4.5.6 that this state takes the form

$$|0\ 0\rangle_{ab} = \frac{|1/2\ 1/2\rangle_a |1/2\ -1/2\rangle_b - |1/2\ -1/2\rangle_a |1/2\ 1/2\rangle_b}{\sqrt{2}}$$

(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 $|1/2\ -1/2\rangle_a$ 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 \hat{L}^+ and \hat{L}^- . 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 $|l\ m\rangle$. What you have seen so far is that $\hat{L}^+|l\ m\rangle$ 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, $\hat{L}^+|l\ m\rangle$ is some multiple of a suitably normalized eigenstate $|l\ m+1\rangle$;

$$\hat{L}^+|l\ m\rangle = C|l\ m+1\rangle$$

where the number C is the multiple. What *is* that multiple? Well, from the magnitude of $\hat{L}^+|l\ m\rangle$, derived earlier in (10.6) you know that its square magnitude is

$$|C|^2 = l(l + 1)\hbar^2 - m^2\hbar^2 - m\hbar^2.$$

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 $|l\ m\rangle$ 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 {A.89}:

$$\hat{L}^+|l m\rangle = \hbar\sqrt{l(l+1) - m(1+m)} |l m+1\rangle \quad (10.9)$$

$$\hat{L}^-|l m\rangle = \hbar\sqrt{l(l+1) + m(1-m)} |l m-1\rangle \quad (10.10)$$

10.1.5 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 $1/2$. 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 4.5.6, 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 $|1/2 1/2\rangle$ and $|1/2 -1/2\rangle$, there are four different combined “product” states:

$$|1/2 1/2\rangle_a |1/2 1/2\rangle_b, |1/2 1/2\rangle_a |1/2 -1/2\rangle_b, |1/2 -1/2\rangle_a |1/2 1/2\rangle_b, \text{ and } |1/2 -1/2\rangle_a |1/2 -1/2\rangle_b.$$

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 $|1/2 1/2\rangle_a |1/2 1/2\rangle_b$ state follows from

$$\begin{aligned} (\hat{L}_{za} + \hat{L}_{zb}) |1/2 1/2\rangle_a |1/2 1/2\rangle_b &= \frac{1}{2}\hbar |1/2 1/2\rangle_a |1/2 1/2\rangle_b + |1/2 1/2\rangle_a \frac{1}{2}\hbar |1/2 1/2\rangle_b \\ &= \hbar |1/2 1/2\rangle_a |1/2 1/2\rangle_b \end{aligned}$$

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 \hat{L}_{za} only acts on particle a , and \hat{L}_{zb} only on particle b {A.90}. 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 $|1/2 1/2\rangle_a |1/2 1/2\rangle_b$ anyway, there is a trick: multiply it with the combined step-up operator

$$\hat{L}_{ab}^+ = \hat{L}_a^+ + \hat{L}_b^+$$

Each part returns zero: \hat{L}_a^+ because particle a is at the top of its ladder and \hat{L}_b^+ because particle b is. So the combined state $|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b$ must be at the top of the ladder too; there is no higher rung. That must mean $l_{ab} = m_{ab} = 1$; the combined state must be a $|1\ 1\rangle$ state. It can be *defined* it as *the* combination $|1\ 1\rangle$ state:

$$|1\ 1\rangle_{ab} \equiv |1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b \quad (10.11)$$

You could just as well have defined $|1\ 1\rangle_{ab}$ as $-|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b$ or $i|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b$, 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 $|1\ 1\rangle_{ab} = |1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b$ by \hat{L}_{ab}^- : that will go one step down the combined states ladder and produce a combination state $|1\ 0\rangle_{ab}$:

$$\hat{L}_{ab}^-|1\ 1\rangle_{ab} = \hbar\sqrt{1(1+1) + 1(1-1)}|1\ 0\rangle_{ab} = \hat{L}_a^-|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b + \hat{L}_b^-|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b$$

or

$$\hbar\sqrt{2}|1\ 0\rangle_{ab} = \hbar|1/2\ -1/2\rangle_a|1/2\ 1/2\rangle_b + \hbar|1/2\ 1/2\rangle_a|1/2\ -1/2\rangle_b$$

where the effects of the ladder-down operators were taken from (10.10). (Note that this requires that the individual particle spin states are normalized consistent with the ladder operators.) The second triplet state is therefore:

$$|1\ 0\rangle_{ab} \equiv \sqrt{1/2}|1/2\ 1/2\rangle_a|1/2\ -1/2\rangle_b + \sqrt{1/2}|1/2\ -1/2\rangle_a|1/2\ 1/2\rangle_b \quad (10.12)$$

But this gives only *one* $|l\ m\rangle$ combination state for the *two* product states $|1/2\ 1/2\rangle_a|1/2\ -1/2\rangle_b$ and $|1/2\ -1/2\rangle_a|1/2\ 1/2\rangle_b$ with zero net z -momentum. If you want to describe unequal combinations of them, like $|1/2\ 1/2\rangle_a|1/2\ -1/2\rangle_b$ by itself, it cannot be just a multiple of $|1\ 0\rangle_{ab}$. This suggests that there may be another $|l\ 0\rangle_{ab}$ 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 :

$$C_1|1/2\ 1/2\rangle_a|1/2\ -1/2\rangle_b + C_2|1/2\ -1/2\rangle_a|1/2\ 1/2\rangle_b$$

For this combination to step up to zero,

$$\begin{aligned} (\hat{L}_a^+ + \hat{L}_b^+) & (C_1|1/2\ 1/2\rangle_a|1/2\ -1/2\rangle_b + C_2|1/2\ -1/2\rangle_a|1/2\ 1/2\rangle_b) \\ &= \hbar C_1|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b + \hbar C_2|1/2\ 1/2\rangle_a|1/2\ 1/2\rangle_b \end{aligned}$$

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

$$|0\ 0\rangle_{ab} = \sqrt{\frac{1}{2}} |1/2\ 1/2\rangle_a |1/2\ -1/2\rangle_b - \sqrt{\frac{1}{2}} |1/2\ -1/2\rangle_a |1/2\ 1/2\rangle_b. \quad (10.13)$$

To find the remaining triplet state, just apply \hat{L}_{ab}^- once more, to $|1\ 0\rangle_{ab}$ above. It gives:

$$|1\ -1\rangle_{ab} = |1/2\ -1/2\rangle_a |1/2\ -1/2\rangle_b \quad (10.14)$$

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.

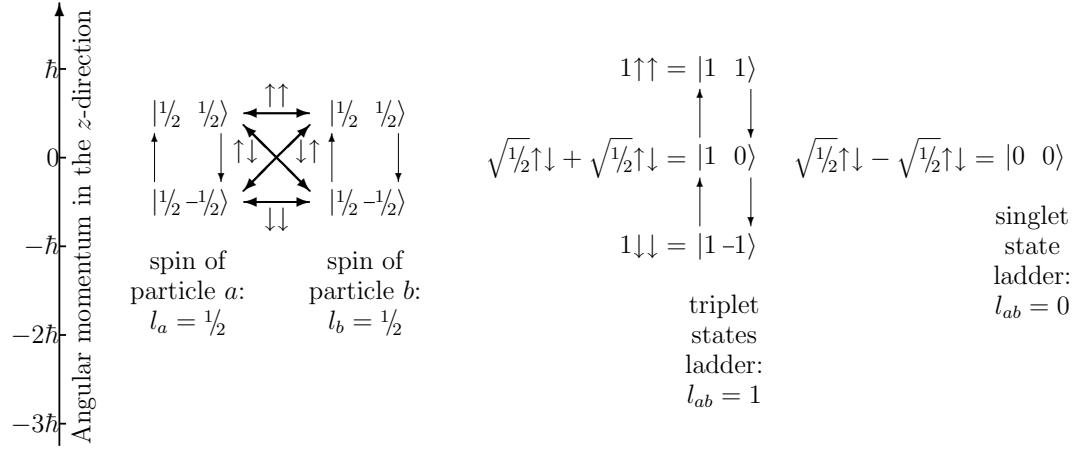


Figure 10.3: Triplet and singlet states in terms of ladders

Figure 10.3 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 $|1/2\ 1/2\rangle_a |1/2\ -1/2\rangle_b$ cannot be shown as a rung on a ladder. In fact, from adding (10.12) and (10.13) it is seen that

$$|1/2\ 1/2\rangle_a |1/2\ -1/2\rangle_b = \sqrt{\frac{1}{2}} |1\ 0\rangle_{ab} + \sqrt{\frac{1}{2}} |0\ 0\rangle_{ab}$$

which makes it a combination of the middle rungs of the triplet and singlet ladders, rather than a single rung.

10.1.6 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 10.3. In particular, you need to know the coefficients that multiply the product states in those combinations.

			$\begin{matrix} 1 \\ \hline 1 \end{matrix}$	
$\begin{matrix} 0 \\ \hline 0 \end{matrix}$	$\begin{matrix} 0 \\ \hline 0 \end{matrix}$	$\begin{matrix} 1 \\ \hline 1 \end{matrix}$	$ 1/2\ 1/2\rangle_a 1/2\ 1/2\rangle_b$	
$\begin{matrix} \frac{1}{2} \\ \hline -\frac{1}{2} \end{matrix}$	$\begin{matrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{matrix}$	$\begin{matrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{matrix}$	$ 1/2\ 1/2\rangle_a 1/2\ -1/2\rangle_b$	
$\begin{matrix} 1 \\ \hline 1 \end{matrix}$	$\begin{matrix} 1 \\ \hline 1 \end{matrix}$	$\begin{matrix} 1/2 & -1/2 \\ 1/2 & -1/2 \end{matrix}$	$ 1/2\ -1/2\rangle_a 1/2\ -1/2\rangle_b$	

Figure 10.4: Clebsch-Gordan coefficients of two spin one half particles.

These coefficients are called “Clebsch-Gordan” coefficients. Figure 10.4 shows the ones from figure 10.3 tabulated. 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 {A.91}. 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 $l = 1$ (i.e. $2\hbar^2$) value. The z -momentum will always be zero.

$l_a = 1, l_b = \frac{1}{2}$ $ 1\ 1\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} \\ -\sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} \end{pmatrix} 1\ 0\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} \\ -\sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} \end{pmatrix} 1\ -1\rangle_a 1/2\ 1/2\rangle_b$	$l_a = \frac{3}{2}, l_b = \frac{1}{2}$ $ 2\ 2\rangle_{ab}$ $ 3/2\ 3/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{3}{4}} & \sqrt{\frac{1}{4}} \\ -\sqrt{\frac{1}{4}} & \sqrt{\frac{3}{4}} \end{pmatrix} 3/2\ 3/2\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{3}{4}} & \sqrt{\frac{1}{4}} \\ -\sqrt{\frac{1}{4}} & \sqrt{\frac{3}{4}} \end{pmatrix} 3/2\ 1/2\rangle_a 1/2\ 1/2\rangle_b$	$l_a = \frac{1}{2}, l_b = \frac{1}{2}$ $ 1\ 0\rangle_{ab}$ $ 1/2\ 1/2\rangle_a 1/2\ 1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \end{pmatrix} 1\ -1\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \end{pmatrix} 1\ 0\rangle_a 1/2\ 1/2\rangle_b$
$l_a = \frac{3}{2}, l_b = \frac{1}{2}$ $ 1\ 1\rangle_{ab}$ $ 3/2\ 3/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} \\ -\sqrt{\frac{2}{5}} & \sqrt{\frac{3}{5}} \end{pmatrix} 2\ 1\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} \\ -\sqrt{\frac{2}{5}} & \sqrt{\frac{3}{5}} \end{pmatrix} 2\ 0\rangle_a 1/2\ 1/2\rangle_b$	$l_a = \frac{5}{2}, l_b = \frac{1}{2}$ $ 2\ 2\rangle_{ab}$ $ 5/2\ 3/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{4}{5}} & \sqrt{\frac{1}{5}} \\ -\sqrt{\frac{1}{5}} & \sqrt{\frac{4}{5}} \end{pmatrix} 2\ 2\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{4}{5}} & \sqrt{\frac{1}{5}} \\ -\sqrt{\frac{1}{5}} & \sqrt{\frac{4}{5}} \end{pmatrix} 2\ 1\rangle_a 1/2\ 1/2\rangle_b$	$l_a = \frac{1}{2}, l_b = \frac{1}{2}$ $ 0\ 0\rangle_{ab}$ $ 1/2\ 1/2\rangle_a 1/2\ 1/2\rangle_b$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 1\ 1\rangle_{ab}$
$l_a = \frac{1}{2}, l_b = \frac{1}{2}$ $ 1\ -1\rangle_{ab}$ $ 1/2\ -1/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 1\ 0\rangle_{ab}$	$l_a = \frac{3}{2}, l_b = \frac{1}{2}$ $ 2\ 1\rangle_{ab}$ $ 3/2\ 1/2\rangle_a 1/2\ 1/2\rangle_b$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 2\ 0\rangle_{ab}$	$l_a = \frac{1}{2}, l_b = \frac{1}{2}$ $ 1\ 0\rangle_{ab}$ $ 1/2\ 1/2\rangle_a 1/2\ 1/2\rangle_b$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 1\ 1\rangle_{ab}$
$l_a = \frac{1}{2}, l_b = \frac{1}{2}$ $ 2\ -1\rangle_{ab}$ $ 3/2\ -1/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{1}{4}} & \sqrt{\frac{3}{4}} \\ -\sqrt{\frac{3}{4}} & \sqrt{\frac{1}{4}} \end{pmatrix} 3/2\ -1/2\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{1}{4}} & \sqrt{\frac{3}{4}} \\ -\sqrt{\frac{3}{4}} & \sqrt{\frac{1}{4}} \end{pmatrix} 3/2\ -3/2\rangle_a 1/2\ 1/2\rangle_b$	$l_a = \frac{3}{2}, l_b = \frac{1}{2}$ $ 3/2\ -1/2\rangle_{ab}$ $ 5/2\ -1/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{2}{5}} & \sqrt{\frac{3}{5}} \\ -\sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} \end{pmatrix} 2\ 0\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{2}{5}} & \sqrt{\frac{3}{5}} \\ -\sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} \end{pmatrix} 2\ -1\rangle_a 1/2\ 1/2\rangle_b$	$l_a = \frac{5}{2}, l_b = \frac{1}{2}$ $ 2\ -2\rangle_{ab}$ $ 5/2\ -3/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} \sqrt{\frac{1}{5}} & \sqrt{\frac{4}{5}} \\ -\sqrt{\frac{4}{5}} & \sqrt{\frac{1}{5}} \end{pmatrix} 2\ -1\rangle_a 1/2\ -1/2\rangle_b$ $- \begin{pmatrix} \sqrt{\frac{1}{5}} & \sqrt{\frac{4}{5}} \\ -\sqrt{\frac{4}{5}} & \sqrt{\frac{1}{5}} \end{pmatrix} 2\ -2\rangle_a 1/2\ 1/2\rangle_b$
$l_a = \frac{3}{2}, l_b = \frac{1}{2}$ $ 2\ -2\rangle_{ab}$ $ 3/2\ -3/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 3/2\ -3/2\rangle_a 1/2\ -1/2\rangle_b$	$l_a = \frac{5}{2}, l_b = \frac{1}{2}$ $ 2\ -2\rangle_{ab}$ $ 5/2\ -3/2\rangle_a 1/2\ -1/2\rangle_b$ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 2\ -2\rangle_a 1/2\ 1/2\rangle_b$	

Figure 10.5: Clebsch-Gordan coefficients for a spin-one-half second particle.

How about the Clebsch-Gordan coefficients to combine other ladders than the spins of two spin $1/2$ 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 {A.92} and outputs the tables in human-readable form {A.93}, like figures 10.5 and 10.6.

From the figures you may note that when two states with total angular momentum quantum numbers l_a and l_b are combined, the combinations have total angular quantum numbers ranging from $|l_a + l_b|$ to $|l_a - l_b|$. This is similar to the fact that when in classical mechanics two angular momentum vectors are combined, the combined total angular momentum L_{ab} is at most $L_a + L_b$ and at least $|L_a - L_b|$. (The so-called “triangle inequality” for combining vectors.) But of course, l is not quite a proportional measure of L unless L is large; in fact, $L = \sqrt{l(l+1)}\hbar$ {A.94}.

10.1.7 Some important results

This section gives some results that are used frequently in quantum analysis, but usually not explicitly stated.

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 $s = 1/2$ 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:

$$j = l + \frac{1}{2} : \quad |j m_j\rangle = \sqrt{\frac{j+m_j}{2j}} Y_l^{m_j-\frac{1}{2}} \uparrow + \sqrt{\frac{j-m_j}{2j}} Y_l^{m_j+\frac{1}{2}} \downarrow$$

$ 3/2, -3/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{2/5} & \sqrt{3/5} \\ -\sqrt{3/5} & \sqrt{2/5} \end{bmatrix} 3/2, -1/2\rangle_a 1, 0\rangle_b$	$ 1/2, 1/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/6} & \sqrt{8/15} & \sqrt{3/10} \\ -\sqrt{1/3} & -\sqrt{1/15} & \sqrt{3/5} \\ \sqrt{1/6} & -\sqrt{8/15} & \sqrt{3/10} \end{bmatrix} 3/2, 1/2\rangle_a 1, -1\rangle_b$	$ 3/2, 3/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/5} & \sqrt{2/5} \\ -\sqrt{2/5} & \sqrt{3/5} \end{bmatrix} 3/2, 3/2\rangle_a 1, 0\rangle_b$	$ 5/2, 1/2\rangle_{ab}$ $\begin{bmatrix} 1 & 3/2, 3/2\rangle_a 1, 1\rangle_b \\ 5/2, 3/2\rangle_{ab} & l_a = 3/2, l_b = 1 \end{bmatrix}$
$ 3/2, -3/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{2/5} & \sqrt{3/5} \\ -\sqrt{3/5} & \sqrt{2/5} \end{bmatrix} 3/2, -3/2\rangle_a 1, 0\rangle_b$	$ 3/2, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/10} & \sqrt{1/2} & \sqrt{1/5} \\ -\sqrt{2/5} & 0 & \sqrt{3/5} \\ \sqrt{3/10} & -\sqrt{1/2} & \sqrt{1/5} \end{bmatrix} 2, 0\rangle_a 1, -1\rangle_b$	$ 2, 1\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/5} & \sqrt{1/3} & \sqrt{1/15} \\ \sqrt{3/10} & \sqrt{1/6} & \sqrt{8/15} \\ \sqrt{1/10} & -\sqrt{1/2} & \sqrt{2/5} \end{bmatrix} 2, 1\rangle_a 1, 0\rangle_b$	$ 2, 2\rangle_{ab}$ $\begin{bmatrix} 1 & 2, 2\rangle_a 1, 1\rangle_b \\ 3, 2\rangle_{ab} & l_a = 2, l_b = 1 \end{bmatrix}$
$ 1/2, -1/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{2/5} & \sqrt{3/5} \\ -\sqrt{3/5} & \sqrt{2/5} \end{bmatrix} 1/2, -1/2\rangle_a 1, 0\rangle_b$	$ 1, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/10} & \sqrt{1/2} & \sqrt{1/5} \\ -\sqrt{2/5} & 0 & \sqrt{3/5} \\ \sqrt{3/10} & -\sqrt{1/2} & \sqrt{1/5} \end{bmatrix} 1, 0\rangle_a 1, -1\rangle_b$	$ 1, 1\rangle_{ab}$ $\begin{bmatrix} \sqrt{2/3} & \sqrt{1/3} \\ -\sqrt{1/3} & \sqrt{2/3} \end{bmatrix} 1, 1\rangle_a 1, 0\rangle_b$	
$ 3/2, -1/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/6} & \sqrt{8/15} & \sqrt{3/10} \\ -\sqrt{1/3} & -\sqrt{1/15} & \sqrt{3/5} \\ \sqrt{1/6} & -\sqrt{8/15} & \sqrt{3/10} \end{bmatrix} 3/2, -1/2\rangle_a 1, 1\rangle_b$	$ 2, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/5} & \sqrt{1/3} & \sqrt{1/15} \\ \sqrt{3/10} & \sqrt{1/6} & \sqrt{8/15} \\ \sqrt{1/10} & -\sqrt{1/2} & \sqrt{2/5} \end{bmatrix} 2, 0\rangle_a 1, 1\rangle_b$	$ 3, 1\rangle_{ab}$ $\begin{bmatrix} \sqrt{2/3} & \sqrt{1/3} \\ -\sqrt{1/3} & \sqrt{2/3} \end{bmatrix} 3, 1\rangle_a 1, 1\rangle_b$	
$ 3/2, 1/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/2} & \sqrt{8/15} & \sqrt{3/10} \\ -\sqrt{1/3} & -\sqrt{1/15} & \sqrt{3/5} \\ \sqrt{1/2} & -\sqrt{8/15} & \sqrt{3/10} \end{bmatrix} 3/2, 1/2\rangle_a 1, 1\rangle_b$	$ 3, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/5} & \sqrt{1/3} & \sqrt{1/15} \\ \sqrt{3/10} & \sqrt{1/6} & \sqrt{8/15} \\ \sqrt{1/10} & -\sqrt{1/2} & \sqrt{2/5} \end{bmatrix} 3, 0\rangle_a 1, 1\rangle_b$	$ 2, 1\rangle_{ab}$ $\begin{bmatrix} \sqrt{2/3} & \sqrt{1/3} \\ -\sqrt{1/3} & \sqrt{2/3} \end{bmatrix} 2, 1\rangle_a 1, 1\rangle_b$	
$ 1/2, 1/2\rangle_{ab}$ $\begin{bmatrix} 1 & 3/2, -3/2\rangle_a 1, -1\rangle_b \\ 1/2, 1/2\rangle_{ab} & \end{bmatrix}$	$ 1, 1\rangle_{ab}$ $\begin{bmatrix} \sqrt{3/10} & \sqrt{1/2} & \sqrt{1/5} \\ -\sqrt{2/5} & 0 & \sqrt{3/5} \\ \sqrt{3/10} & -\sqrt{1/2} & \sqrt{1/5} \end{bmatrix} 1, 1\rangle_a 1, -1\rangle_b$	$ 2, 1\rangle_{ab}$ $\begin{bmatrix} 1 & 1, 1\rangle_a 1, 1\rangle_b \\ 2, 1\rangle_{ab} & l_a = 1, l_b = 1 \end{bmatrix}$	
$ 3/2, 1/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/10} & \sqrt{1/2} & \sqrt{2/5} \\ -\sqrt{3/10} & -\sqrt{1/6} & \sqrt{8/15} \\ \sqrt{3/5} & -\sqrt{1/3} & \sqrt{1/15} \end{bmatrix} 3/2, 1/2\rangle_a 1, 1\rangle_b$	$ 0, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/3} & \sqrt{1/2} & \sqrt{1/6} \\ -\sqrt{1/3} & 0 & \sqrt{2/3} \\ \sqrt{1/3} & -\sqrt{1/2} & \sqrt{1/6} \end{bmatrix} 0, 0\rangle_a 1, -1\rangle_b$	$ 1, 1\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{bmatrix} 1, 1\rangle_a 1, 0\rangle_b$	
$ 3/2, -1/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/10} & \sqrt{1/2} & \sqrt{2/5} \\ -\sqrt{3/10} & -\sqrt{1/6} & \sqrt{8/15} \\ \sqrt{3/5} & -\sqrt{1/3} & \sqrt{1/15} \end{bmatrix} 3/2, -1/2\rangle_a 1, 1\rangle_b$	$ 1, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/3} & \sqrt{1/2} & \sqrt{1/6} \\ -\sqrt{1/3} & 0 & \sqrt{2/3} \\ \sqrt{1/3} & -\sqrt{1/2} & \sqrt{1/6} \end{bmatrix} 1, 0\rangle_a 1, 0\rangle_b$	$ 1, 1\rangle_{ab}$ $\begin{bmatrix} 1 & 1, 0\rangle_a 1, 1\rangle_b \\ 1, 1\rangle_{ab} & \end{bmatrix}$	
$ 1/2, -1/2\rangle_{ab}$ $\begin{bmatrix} 1 & 2, -2\rangle_a 1, -1\rangle_b \\ 1/2, -1/2\rangle_{ab} & \end{bmatrix}$	$ 1, -1\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{bmatrix} 1, -1\rangle_a 1, -1\rangle_b$	$ 2, -1\rangle_{ab}$ $\begin{bmatrix} 1 & 1, -1\rangle_a 1, -1\rangle_b \\ 2, -1\rangle_{ab} & \end{bmatrix}$	$ 1, -1\rangle_{ab}$ $\begin{bmatrix} 1 & 1, -1\rangle_a 1, -1\rangle_b \\ 1, -1\rangle_{ab} & \end{bmatrix}$
$ 3/2, -3/2\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/3} & \sqrt{2/3} \\ -\sqrt{2/3} & \sqrt{1/3} \end{bmatrix} 3/2, -3/2\rangle_a 1, 0\rangle_b$	$ 1, 0\rangle_{ab}$ $\begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ -\sqrt{1/2} & \sqrt{1/2} \end{bmatrix} 1, 0\rangle_a 1, -1\rangle_b$	$ 2, -2\rangle_{ab}$ $\begin{bmatrix} 1 & 1, 0\rangle_a 1, 1\rangle_b \\ 2, -2\rangle_{ab} & l_a = 1, l_b = 1 \end{bmatrix}$	$ 1, -1\rangle_{ab}$ $\begin{bmatrix} 1 & 1, -1\rangle_a 1, -1\rangle_b \\ 1, -1\rangle_{ab} & \end{bmatrix}$

Figure 10.6: Clebsch-Gordan coefficients for a spin-one second particle.

If the net spin is $j = l - \frac{1}{2}$, assuming that $l > 0$, that becomes

$$j = l - \frac{1}{2} : \quad |j m_j\rangle = -\sqrt{\frac{j+1-m_j}{2j+2}} Y_l^{m_j-\frac{1}{2}} \uparrow + \sqrt{\frac{j+1+m_j}{2j+2}} Y_l^{m_j+\frac{1}{2}} \downarrow$$

Note that if the net angular momentum is determinate, 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:

$$\langle l_{ab} m_{ab} || l_a m_a \rangle |l_b m_b\rangle = (-1)^{l_a + l_b - l_{ab}} \langle l_{ab} m_{ab} || l_b m_b \rangle |l_a m_a\rangle$$

For $l_a = l_b = \frac{1}{2}$, this verifies that the triplet states $l_{ab} = 1$ are symmetric, and the singlet state $l_{ab} = 0$ is antisymmetric. More generally, states with the maximum net angular momentum $l_{ab} = l_a + l_b$ 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

$$\langle l_{ab} m_{ab} || l_a m_a \rangle |l_b m_b\rangle = (-1)^{l_a - l_{ab} + m_b} \sqrt{\frac{2l_{ab} + 1}{2l_a + 1}} \langle l_a m_a || l_{ab} m_{ab} \rangle |l_b - m_b\rangle$$

This is of interest in figuring out what states produce zero net angular momentum, $l_{ab} = m_{ab} = 0$. In that case, the right hand side is zero unless $l_b = l_a$ and $m_b = -m_a$; and then $\langle l_a m_a || 0 0 \rangle |l_a m_a\rangle = 1$. 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.

10.1.8 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 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 10.1, {A.95}. In it, I is the number of fermions in the shell. Further j^P is the net angular momentum of the single-particle states

		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	1													
$\frac{7}{2}$	1				1														
	3		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							
$\frac{11}{2}$	1						1												
	3		1	1	1	2	2	1	2	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																
$\frac{7}{2}$	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								
$\frac{11}{2}$	2	1	1	1		1		1		1		1								
	4	2		3	1	4	2	4	2	4	2	3	1	2	1	1				
	6	3		4	3	6	3	7	4	6	4	5	2	4	2	2	1	1	1	

Table 10.1: Possible combined angular momentum of identical fermions in shells of single-particle states that differ in magnetic quantum number.

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 $j^P = \frac{1}{2}$. For a single particle, the “combined” momentum j is simply the single particle momentum j^P , explaining the single 1 in the $\frac{1}{2}$ 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 $\frac{1}{2}, \frac{3}{2}, \dots, -\frac{1}{2}$. All the ten states in this set have net angular momentum $j = j^P = \frac{1}{2}$.

Next assume that there are two particles in the same $j^P = \frac{1}{2}$ single-particle states. Then if both particles would be in the $m^P = \frac{1}{2}$ single-particle state, their combined angular momentum in the z -direction m would be $2 \times \frac{1}{2} = 9$. 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 $m^P = \frac{1}{2}$ 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 $m^P = \frac{1}{2}$ state, the second one can at most be in the $m^P = \frac{3}{2}$ 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, $\dots, -8$, 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 $j^P = \frac{1}{2}$ 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 cancellation 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 $j^P = \frac{1}{2}$ 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 $j^P = \frac{1}{2}$ 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 $\frac{5}{2}, \frac{3}{2}, \dots, -\frac{5}{2}$. So there are a total of 12 independent combination states that have net angular momentum $j = \frac{5}{2}$.

Note that a shell has $2j^p + 1$ different single-particle states, because the magnetic quantum number m^p can have the values $j^p, j^p-1, \dots, -j^p$. Therefore a shell can accommodate up to $2j^p + 1$ fermions according to the exclusion principle. However, the table only lists combined angular momentum values for up to $j^p + \frac{1}{2}$ 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 $j^p = \frac{1}{2}$ 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 10.2. 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.

10.1.9 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 not by a capital L , but by a capital S . Similarly, the azimuthal quantum number is then indicated by s instead of l . 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 \hat{S}_x and \hat{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 \hat{S}^+ and \hat{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 (10.9) and (10.10), you have:

$$\hat{S}^+ \uparrow = (\hat{S}_x + i\hat{S}_y)\uparrow = 0 \quad \hat{S}^- \uparrow = (\hat{S}_x - i\hat{S}_y)\uparrow = \hbar\downarrow$$

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
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		1									
	8	1		1		1		1		1		1								
	9		1		1		1		1		1		1							
	2	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		1								
	6	2	2	1	3	1	3	1	2	1	1	1		1						
	7	1	3	1	3	2	3	2	3	1	2	1	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	2	1		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					
	5	2	1		1	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					
	5	2	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		
	7	2	1		1	1		1		1		1		1		1				
	8	2	1		1	1		1		1		1		1		1				
5	2	1		1	1	1		1		1		1		1		1				

Table 10.2: Possible combined angular momentum of identical bosons.

By adding or subtracting the two equations, you find the effects of \hat{S}_x and \hat{S}_y on the spin-up state:

$$\hat{S}_x \uparrow = \frac{1}{2}\hbar \downarrow \quad \hat{S}_y \uparrow = \frac{1}{2}i\hbar \downarrow$$

It works the same way for the spin-down state $\downarrow = |1/2 - 1/2\rangle$:

$$\hat{S}_x \downarrow = \frac{1}{2}\hbar \uparrow \quad \hat{S}_y \downarrow = -\frac{1}{2}i\hbar \uparrow$$

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 \hat{S}_x will turn it into:

$$\hat{S}_x (a\uparrow + b\downarrow) = a \left(0\uparrow + \frac{1}{2}\hbar \downarrow \right) + b \left(\frac{1}{2}\hbar \uparrow + 0\downarrow \right)$$

while the operator \hat{S}_y turns it into

$$\hat{S}_y (a\uparrow + b\downarrow) = a \left(0\uparrow + \frac{1}{2}\hbar i\downarrow \right) + b \left(-\frac{1}{2}\hbar i\uparrow + 0\downarrow \right)$$

And of course, since \uparrow and \downarrow are the eigenstates of \hat{S}_z ,

$$\hat{S}_z (a\uparrow + b\downarrow) = a \left(\frac{1}{2}\hbar \uparrow + 0\downarrow \right) + b \left(0\uparrow - \frac{1}{2}\hbar \downarrow \right)$$

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”:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (10.15)$$

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 \hat{S}_x and \hat{S}_y -operators in terms of the eigenstates \uparrow and \downarrow of the \hat{S}_z operator. You can use the techniques of linear algebra, or you can guess. For example, if you guess $a = b = 1$,

$$\hat{S}_x \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2}\hbar \sigma_x \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2}\hbar \begin{pmatrix} 0 \times 1 + 1 \times 1 \\ 1 \times 1 + 0 \times 1 \end{pmatrix} = \frac{1}{2}\hbar \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

so $a = b = 1$ is an eigenstate of \hat{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:

$$\rightarrow = \frac{1}{\sqrt{2}}\uparrow + \frac{1}{\sqrt{2}}\downarrow \quad \leftarrow = -\frac{i}{\sqrt{2}}\uparrow + \frac{i}{\sqrt{2}}\downarrow \quad \otimes = \frac{1}{\sqrt{2}}\uparrow + \frac{i}{\sqrt{2}}\downarrow \quad \odot = \frac{1}{\sqrt{2}}\uparrow - \frac{i}{\sqrt{2}}\downarrow \quad (10.16)$$

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 a 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 (10.16). 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 $\hat{S}_y \pm i\hat{S}_z$ and the y -direction one $\hat{S}_z \pm i\hat{S}_x$, 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 -axes, the x or y axis in the system you are looking at now, {A.96}.

10.1.10 General spin matrices

The arguments that produced the Pauli spin matrices for a system with spin $\frac{1}{2}$ work equally well for systems with larger square angular momentum.

In particular, from the definition of the ladder operators

$$\hat{L}^+ \equiv \hat{L}_x + i\hat{L}_y \quad \hat{L}^- \equiv \hat{L}_x - i\hat{L}_y$$

it follows by taking the sum, respectively difference, that

$$\hat{L}_x = \frac{1}{2}\hat{L}^+ + \frac{1}{2}\hat{L}^- \quad \hat{L}_y = -i\frac{1}{2}\hat{L}^+ + i\frac{1}{2}\hat{L}^- \quad (10.17)$$

Therefore, the effect of either \hat{L}_x or \hat{L}_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 (10.9) and (10.10).

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 \hat{L}_x is real symmetric while \hat{L}_y is purely imaginary, equal to i times a real skew-symmetric matrix. Be sure to tell all your 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.

10.2 The Relativistic Dirac Equation

Relativity threw up some road blocks when quantum mechanics was first formulated, especially for the electrically charged 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.

For zero spin particles, including relativity appears to be simple. The classical kinetic energy Hamiltonian for a particle in free space,

$$H = \frac{1}{2m} \sum_{i=1}^3 \hat{p}_i^2 \quad \hat{p}_i = \frac{\hbar}{i} \frac{\partial}{\partial r_i}$$

can be replaced by Einstein's relativistic expression

$$H = \sqrt{(m_0 c^2)^2 + \sum_{i=1}^3 (\hat{p}_i c)^2}$$

where m_0 is the rest mass of the particle and $m_0 c^2$ 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:

$$\left[(m_0 c^2)^2 + \sum_{i=1}^3 (\hat{p}_i c)^2 \right] \psi = E^2 \psi$$

This is the “Klein-Gordon” relativistic version of the Hamiltonian eigenvalue problem, and with a bit of knowledge of partial differential equations, you can check that the unsteady version, chapter 6.1, obeys the speed of light as the maximum propagation speed, as you would expect, chapter 13.5.

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 $1/2$ elementary particles like quarks, it turned out,) the square root produces a simple linear combination of the individual square root terms:

$$\sqrt{(m_0c^2)^2 + \sum_{i=1}^3 (\hat{p}_i c)^2} = \alpha_0 m_0 c^2 + \sum_{i=1}^3 \alpha_i \hat{p}_i c \quad (10.18)$$

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 10.1.9. The square root cannot be eliminated with matrices smaller than 4×4 in actual size. A derivation is in note {A.98}.

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 10.6, 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 $E = mc^2$, {A.4}.

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.

10.3 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 chapter 12.2 for some of the ideas.)

In classical electromagnetics, the force on a particle with charge q in a field with electric strength \vec{E} and magnetic strength \vec{B} is given by the Lorentz force law

$$m \frac{d\vec{v}}{dt} = q (\vec{E} + \vec{v} \times \vec{B}) \quad (10.19)$$

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

$$H = \frac{1}{2m} (\hat{\vec{p}} - q\vec{A})^2 + q\varphi \quad (10.20)$$

where $\varphi = V/q$ is the “electric potential” per unit charge, and \vec{A} is the “magnetic vector potential” per unit charge. And this simplest guess is in fact right.

The relationship between the vector potential \vec{A} and the magnetic field strength \vec{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{E} and \vec{B} will be constant on the small quantum scales. That means that the derivatives of φ will be constant, (since \vec{E} is the negative gradient of φ), and presumably the same for the derivatives of \vec{A} .

Now according to chapter 6.1.7, the evolution of the expectation value of position is found as

$$\frac{d\langle \hat{\vec{r}} \rangle}{dt} = \left\langle \frac{i}{\hbar} [H, \vec{r}] \right\rangle$$

Working out the commutator with the Hamiltonian above, {A.99}, you get,

$$\frac{d\langle \hat{\vec{r}} \rangle}{dt} = \frac{1}{m} \langle \hat{\vec{p}} - q\vec{A} \rangle$$

This is unexpected; it shows that $\hat{\vec{p}}$, i.e. $\hbar\nabla/i$, is no longer the operator of the normal momentum $m\vec{v}$ when there is a magnetic field; $\hat{\vec{p}} - q\vec{A}$ gives the normal

momentum. The momentum represented by $\hat{\vec{p}}$ by itself is called “canonical” momentum to distinguish it from normal momentum:

The canonical momentum $\hbar\nabla/\mathbf{i}$ 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, {A.4.10})

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 $\hat{\vec{p}} - q\vec{A}$, then

$$m \frac{d\langle \vec{v} \rangle}{dt} = \frac{d\langle \hat{\vec{p}} - q\vec{A} \rangle}{dt} = \left\langle \frac{i}{\hbar} [H, \hat{\vec{p}} - q\vec{A}] \right\rangle - q \left\langle \frac{\partial \vec{A}}{\partial t} \right\rangle$$

The objective is now to ensure that the right hand side is the correct Lorentz force (10.19) for the assumed Hamiltonian, by a suitable definition of \vec{B} in terms of \vec{A} .

After a lot of grinding down commutators, {A.99}, it turns out that indeed the Lorentz force is obtained,

$$m \frac{d\langle \vec{v} \rangle}{dt} = q (\vec{E} + \langle \vec{v} \rangle \times \vec{B})$$

provided that:

$$\boxed{\vec{E} = -\nabla\varphi - \frac{\partial \vec{A}}{\partial t} \quad \vec{B} = \nabla \times \vec{A}} \quad (10.21)$$

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{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{B} in its actual workings.

10.4 Maxwell's Equations [Descriptive]

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{B} "derived" (guessed) there, (10.21). 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

$$\text{Maxwell's second equation: } \nabla \cdot \vec{B} = 0 \quad (10.22)$$

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{E} , but its divergence is *not* zero:

$$\text{Maxwell's first equation: } \nabla \cdot \vec{E} = \frac{\rho}{\epsilon_0} \quad (10.23)$$

where ρ is the electric charge per unit volume that is present and the constant $\epsilon_0 = 8.85 \times 10^{-12} \text{ C}^2/\text{J m}$ is called the permittivity of space.

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 10.7. By symmetry, the electric field at all points on the spherical surface is radial, and everywhere has the same magnitude $E = |\vec{E}|$; figure 10.7 shows it for eight selected points.

Now watch what happens if you integrate both sides of Maxwell's first equation (10.23) 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

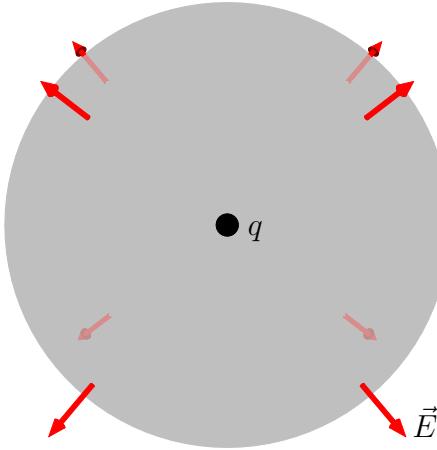


Figure 10.7: Relationship of Maxwell's first equation to Coulomb's law.

calculus says that the divergence of any vector, \vec{E} in this case, integrated over the *volume* of the sphere, equals the radial electric field E integrated over the *surface* of the sphere. Since E is constant on the surface, and the surface of a sphere is just $4\pi r^2$, the right hand side integrates to $4\pi r^2 E$. So in total, you get for the integrated first Maxwell's equation that $4\pi r^2 E = q/\epsilon_0$. Take the $4\pi r^2$ to the other side and there you have the Coulomb electric field of a point charge:

$$\text{Coulomb's law: } E = \frac{q}{4\pi r^2 \epsilon_0} \quad (10.24)$$

Multiply by $-e$ and you have the electrostatic force on an electron in that field according to the Lorentz equation (10.19). 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.

Of course, you can also integrate Maxwell's first equation over more general regions than a sphere centered around a charge. For example figure 10.8 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 inte-

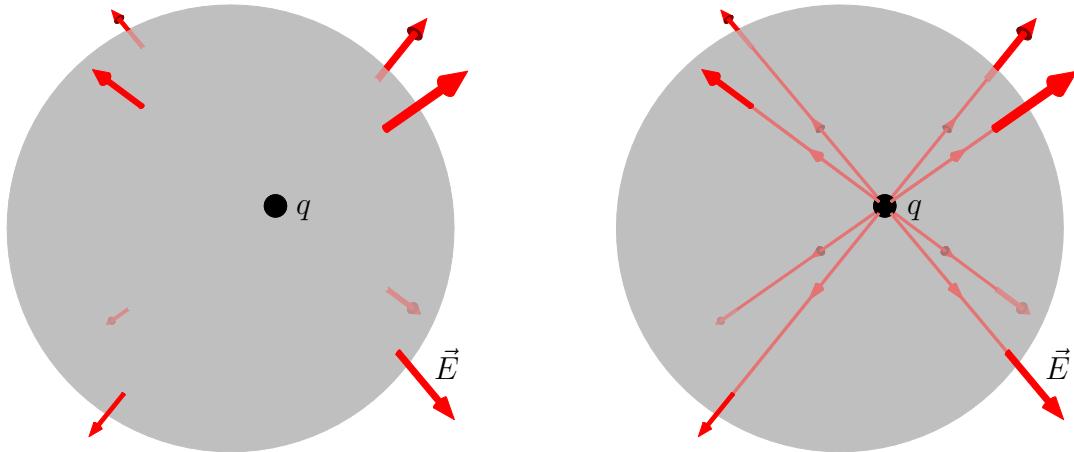


Figure 10.8: Maxwell's first equation for a more arbitrary region. The figure to the right includes the field lines through the selected points.

grate 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 precision, 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 10.8 shows the field lines through the selected points; a single charge has radial field lines.

Assume that you draw the field lines densely, more like figure 10.9 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, {A.100}. 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-

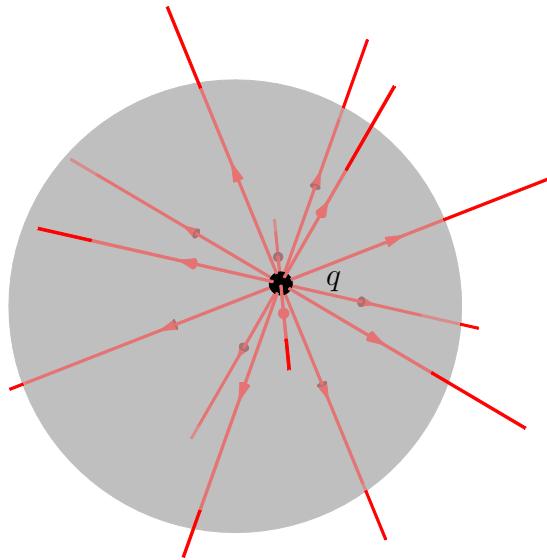


Figure 10.9: The net number of field lines leaving a region is a measure for the net charge inside that region.

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.

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 10.10 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 10.7. 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:

$$\text{Maxwell's third equation: } \nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad (10.25)$$

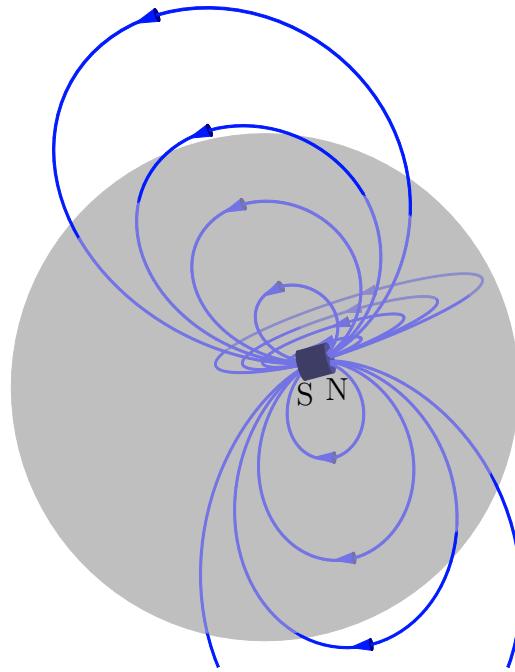


Figure 10.10: Since magnetic monopoles do not exist, the net number of magnetic field lines leaving a region is always zero.

The “curl”, $\nabla \times$, is also often indicated as “rot”.

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 10.11. 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 10.11 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 10.11. 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 10.11. You better not make that a big bulb unless

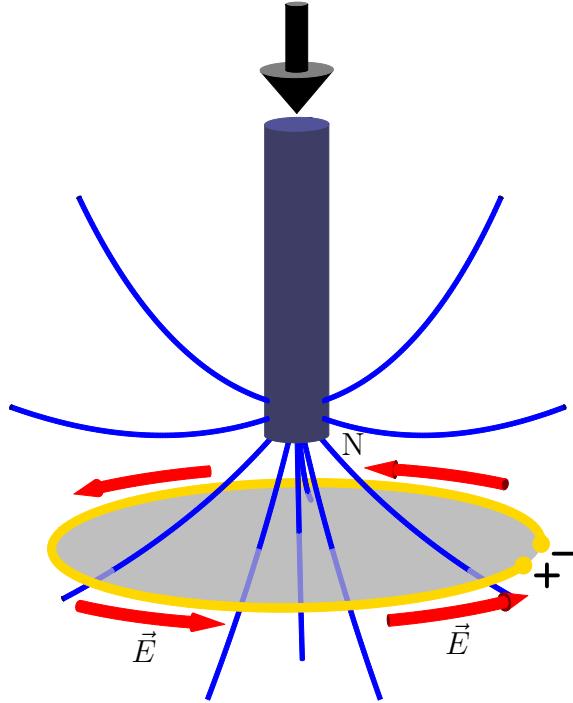


Figure 10.11: Electric power generation.

you wrap the wire around a lot more times to form a spool, but anyway. {A.101}.

Maxwell's fourth and final equation is a similar expression for the curl of the magnetic field:

$$\text{Maxwell's fourth equation: } c^2 \nabla \times \vec{B} = \vec{j} + \frac{\partial \vec{E}}{\partial t} \quad (10.26)$$

where \vec{j} 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{B} by a factor c to get the speed of light to show up equally in the equations for the curl of \vec{E} and the curl of \vec{B} , but then the Lorentz force law must be adjusted too.)

The big difference from the third equation is the appearance of the current density \vec{j} . So, there are two ways to create a circulatory magnetic field, as shown in figure 10.12: (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 10.11.

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

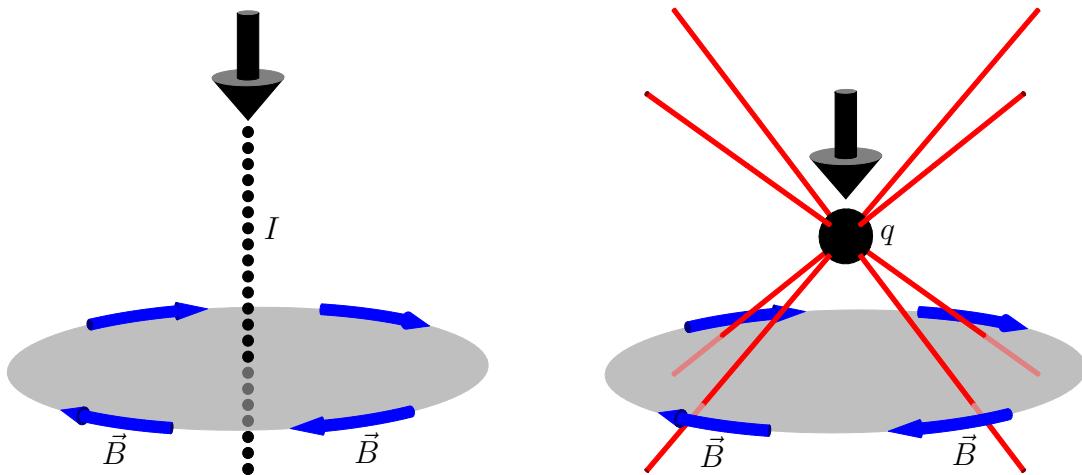


Figure 10.12: Two ways to generate a magnetic field: using a current (left) or using a varying electric field (right).

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:”

$$\text{Maxwell's continuity equation: } \nabla \cdot \vec{j} = -\frac{\partial \rho}{\partial t} \quad (10.27)$$

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{j} will be zero. In that case, the solutions of Maxwell's equations are simply combinations of “traveling waves.” A traveling wave takes the form

$$\vec{E} = \hat{k}E_0 \cos(\omega(t - y/c) - \phi) \quad \vec{B} = \hat{i}\frac{1}{c}E_0 \cos(\omega(t - y/c) - \phi) \quad (10.28)$$

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 $\hat{k}E_0$ of the electric field of the wave. 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 ϕ 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.

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{E} in one of Maxwell's equations is no longer the same as the messed-up \vec{E} in another, and the same for \vec{B} . So physicists rename one messed-up \vec{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.

10.5 Example Static Electromagnetic Fields

In this section, some basic solutions of Maxwell's equations are described. They will be of interest in chapter 12.1.6 for understanding relativistic effects on the

hydrogen atom (though certainly not essential). They are also of considerable practical importance for a lot of non-quantum 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 10.3 and 10.4. 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{j}_m in addition to an electric current density \vec{j} . 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 $c\vec{B}$ and \vec{j}_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 $c\vec{B}$ instead of \vec{B} in SI units, but no such luck.

10.5.1 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

$$\text{point charge at the origin: } \rho(\vec{r}) = q\delta^3(\vec{r}) \quad (10.29)$$

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 10.9 in the previous subsection. According to Coulomb's law, the electric field of a point charge is

$$\text{electric field of a point charge: } \vec{E} = \frac{q}{4\pi\epsilon_0 r^2} \hat{r} \quad (10.30)$$

where r is the distance from the charge, \hat{r} is the unit vector pointing straight away from the charge, and $\epsilon_0 = 8.85 \cdot 10^{-12} \text{ C}^2/\text{J m}$ is the permittivity of space. Now for static electric charges the electric field is minus the gradient of a potential φ ,

$$\vec{E} = -\nabla\varphi \quad \nabla \equiv \hat{i}\frac{\partial}{\partial x} + \hat{j}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$$

Physical constants:

$$\epsilon_0 = 8.854\,187\,817 \dots \text{ } 10^{-12} \text{C}^2/\text{Nm}^2 \quad c = 299\,792\,458 \approx 3\,10^8 \text{ m/s}$$

Lorentz force law:

$$\vec{F} = q \left(\vec{E} + \frac{\vec{v}}{c} \times c\vec{B} \right) + \frac{q_m}{c} \left(c\vec{B} - \frac{\vec{v}}{c} \times \vec{E} \right)$$

Maxwell's equations:

$$\begin{aligned} \nabla \cdot \vec{E} &= \frac{1}{\epsilon_0} \rho & \nabla \cdot c\vec{B} &= \frac{1}{\epsilon_0} \frac{\rho_m}{c} \\ \nabla \times \vec{E} &= -\frac{1}{c} \frac{\partial c\vec{B}}{\partial t} - \frac{1}{\epsilon_0 c} \frac{\vec{j}_m}{c} & \nabla \times c\vec{B} &= \frac{1}{c} \frac{\partial \vec{E}}{\partial t} + \frac{1}{\epsilon_0 c} \vec{j} \\ \nabla \cdot \vec{j} + \frac{\partial \vec{\rho}}{\partial t} &= 0 & \nabla \cdot \vec{j}_m + \frac{\partial \vec{\rho}_m}{\partial t} &= 0 \end{aligned}$$

Existence of a potential:

$$\vec{E} = -\nabla \varphi \quad \text{iff} \quad \nabla \times \vec{E} = 0 \quad \vec{B} = -\nabla \varphi_m \quad \text{iff} \quad \nabla \times \vec{B} = 0$$

Point charge at the origin:

$$\varphi = \frac{q}{4\pi\epsilon_0 r} \quad \vec{E} = \frac{q}{4\pi\epsilon_0 r^3} \vec{r} \quad c\varphi_m = \frac{q_m}{4\pi\epsilon_0 c r} \quad c\vec{B} = \frac{q_m}{4\pi\epsilon_0 c r^3} \vec{r}$$

Point charge at the origin in 2D:

$$\varphi = \frac{q'}{2\pi\epsilon_0} \ln \frac{1}{r} \quad \vec{E} = \frac{q'}{2\pi\epsilon_0 r^2} \vec{r} \quad c\varphi_m = \frac{q'_m}{2\pi\epsilon_0 c} \ln \frac{1}{r} \quad c\vec{B} = \frac{q'_m}{2\pi\epsilon_0 c r^2} \vec{r}$$

Charge dipoles:

$$\begin{aligned} \varphi &= \frac{q}{4\pi\epsilon_0} \left[\frac{1}{|\vec{r} - \vec{r}_\oplus|} - \frac{1}{|\vec{r} - \vec{r}_\ominus|} \right] & c\varphi_m &= \frac{q_m}{4\pi\epsilon_0 c} \left[\frac{1}{|\vec{r} - \vec{r}_\oplus|} - \frac{1}{|\vec{r} - \vec{r}_\ominus|} \right] \\ \vec{E} &= \frac{q}{4\pi\epsilon_0} \left[\frac{\vec{r} - \vec{r}_\oplus}{|\vec{r} - \vec{r}_\oplus|^3} - \frac{\vec{r} - \vec{r}_\ominus}{|\vec{r} - \vec{r}_\ominus|^3} \right] & c\vec{B} &= \frac{q_m}{4\pi\epsilon_0 c} \left[\frac{\vec{r} - \vec{r}_\oplus}{|\vec{r} - \vec{r}_\oplus|^3} - \frac{\vec{r} - \vec{r}_\ominus}{|\vec{r} - \vec{r}_\ominus|^3} \right] \\ \vec{\varphi} &= q(\vec{r}_\oplus - \vec{r}_\ominus) \quad E_{\text{ext}} = -\vec{\varphi} \cdot \vec{E}_{\text{ext}} & \vec{\mu} &= q_m(\vec{r}_\oplus - \vec{r}_\ominus) \quad E_{\text{ext}} = -\vec{\mu} \cdot \vec{B}_{\text{ext}} \end{aligned}$$

Charge dipoles in 2D:

$$\begin{aligned} \varphi &= \frac{q'}{2\pi\epsilon_0} \left[\ln \frac{1}{|\vec{r} - \vec{r}_\oplus|} - \ln \frac{1}{|\vec{r} - \vec{r}_\ominus|} \right] & c\varphi_m &= \frac{q'_m}{2\pi\epsilon_0 c} \left[\ln \frac{1}{|\vec{r} - \vec{r}_\oplus|} - \ln \frac{1}{|\vec{r} - \vec{r}_\ominus|} \right] \\ \vec{E} &= \frac{q'}{2\pi\epsilon_0} \left[\frac{\vec{r} - \vec{r}_\oplus}{|\vec{r} - \vec{r}_\oplus|^2} - \frac{\vec{r} - \vec{r}_\ominus}{|\vec{r} - \vec{r}_\ominus|^2} \right] & c\vec{B} &= \frac{q'_m}{2\pi\epsilon_0 c} \left[\frac{\vec{r} - \vec{r}_\oplus}{|\vec{r} - \vec{r}_\oplus|^2} - \frac{\vec{r} - \vec{r}_\ominus}{|\vec{r} - \vec{r}_\ominus|^2} \right] \\ \vec{\varphi}' &= q'(\vec{r}_\oplus - \vec{r}_\ominus) \quad E'_{\text{ext}} = -\vec{\varphi}' \cdot \vec{E}_{\text{ext}} & \vec{\mu}' &= q'_m(\vec{r}_\oplus - \vec{r}_\ominus) \quad E'_{\text{ext}} = -\vec{\mu}' \cdot \vec{B}_{\text{ext}} \end{aligned}$$

Table 10.3: Electromagnetics I: Fundamental equations and basic solutions.

Distributed charges:

$$\begin{aligned}
 \varphi &= \frac{1}{4\pi\epsilon_0} \int_{\text{all } \vec{r}} \frac{1}{|\vec{r} - \vec{r}'|} \rho(\vec{r}') d^3\vec{r}' & c\varphi_m &= \frac{1}{4\pi\epsilon_0} \int_{\text{all } \vec{r}} \frac{1}{|\vec{r} - \vec{r}'|} \frac{\rho_m(\vec{r}')}{c} d^3\vec{r}' \\
 \vec{E} &= \frac{1}{4\pi\epsilon_0} \int_{\text{all } \vec{r}} \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3} \rho(\vec{r}') d^3\vec{r}' & c\vec{B} &= \frac{1}{4\pi\epsilon_0} \int_{\text{all } \vec{r}} \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3} \frac{\rho_m(\vec{r}')}{c} d^3\vec{r}' \\
 \varphi &\sim \frac{q}{4\pi\epsilon_0 r} + \frac{1}{4\pi\epsilon_0 r^3} \vec{\varphi} \cdot \vec{r} & c\varphi_m &\sim \frac{q_m}{4\pi\epsilon_0 c r} + \frac{1}{4\pi\epsilon_0 cr^3} \vec{\mu} \cdot \vec{r} \\
 \vec{E} &\sim \frac{q}{4\pi\epsilon_0 r^3} + \frac{1}{4\pi\epsilon_0} \frac{3(\vec{\varphi} \cdot \vec{r})\vec{r} - \vec{\varphi}r^2}{r^5} & c\vec{B} &\sim \frac{q_m}{4\pi\epsilon_0 c r^3} + \frac{1}{4\pi\epsilon_0} \frac{3(\vec{\mu} \cdot \vec{r})\vec{r} - \vec{\mu}r^2}{cr^5} \\
 q &= \int \rho(\vec{r}) d^3\vec{r} & \vec{\varphi} &= \int \vec{r}\rho(\vec{r}) d^3\vec{r} & q_m &= \int \rho_m(\vec{r}) d^3\vec{r} & \vec{\mu} &= \int \vec{r}\rho_m(\vec{r}) d^3\vec{r}
 \end{aligned}$$

Ideal charge dipoles:

$$\begin{aligned}
 \varphi &= \frac{1}{4\pi\epsilon_0} \frac{\vec{\varphi} \cdot \vec{r}}{r^3} & c\varphi_m &= \frac{1}{4\pi\epsilon_0} \frac{\vec{\mu} \cdot \vec{r}}{cr^3} \\
 \vec{E} &= \frac{1}{4\pi\epsilon_0} \frac{3(\vec{\varphi} \cdot \vec{r})\vec{r} - \vec{\varphi}r^2}{r^5} - \frac{\vec{\varphi}}{3\epsilon_0} \delta^3(\vec{r}) & c\vec{B} &= \frac{1}{4\pi\epsilon_0} \frac{3(\vec{\mu} \cdot \vec{r})\vec{r} - \vec{\mu}r^2}{cr^5} - \frac{\vec{\mu}}{3\epsilon_0 c} \delta^3(\vec{r})
 \end{aligned}$$

Biot-Savart law for current densities and currents:

$$\begin{aligned}
 \vec{E} &= \frac{1}{4\pi\epsilon_0 c} \int_{\text{all } \vec{r}} \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3} \times \frac{\vec{j}_m(\vec{r}')}{c} d^3\vec{r}' & c\vec{B} &= -\frac{1}{4\pi\epsilon_0 c} \int_{\text{all } \vec{r}} \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3} \times \vec{j}(\vec{r}') d^3\vec{r}' \\
 \vec{E} &= \frac{1}{4\pi\epsilon_0 c} \int_{\text{all } \vec{r}} \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3} \times \frac{I_m(\vec{r}')}{c} d\vec{r}' & c\vec{B} &= -\frac{1}{4\pi\epsilon_0 c} \int_{\text{all } \vec{r}} \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|^3} \times I(\vec{r}') d\vec{r}'
 \end{aligned}$$

2D field due to a straight current along the z -axis:

$$\varphi = \frac{I_m}{2\pi\epsilon_0 c^2} \theta \quad \vec{E} = -\frac{I_m}{2\pi\epsilon_0 c^2} \frac{1}{r} \hat{i}_\theta \quad c\varphi_m = -\frac{I}{2\pi\epsilon_0 c} \theta \quad c\vec{B} = \frac{I}{2\pi\epsilon_0 c} \frac{1}{r} \hat{i}_\theta$$

Current dipole moment:

$$\begin{aligned}
 \vec{\varphi} &= -\frac{1}{2c} \int_{\text{all } \vec{r}} \vec{r} \times \frac{\vec{j}_m(\vec{r}')}{c} d^3\vec{r}' & \vec{\mu} &= \frac{1}{2} \int_{\text{all } \vec{r}} \vec{r} \times \vec{j}(\vec{r}') d^3\vec{r}' = \frac{q_c}{2m_c} \vec{L} \\
 \vec{M} &= \vec{\varphi} \times \vec{E}_{\text{ext}} & E_{\text{ext}} &= -\vec{\varphi} \cdot \vec{E}_{\text{ext}} & \vec{M} &= \vec{\mu} \times \vec{B}_{\text{ext}} & E_{\text{ext}} &= -\vec{\mu} \cdot \vec{B}_{\text{ext}}
 \end{aligned}$$

Ideal current dipoles:

$$\begin{aligned}
 \varphi &= \frac{1}{4\pi\epsilon_0} \frac{\vec{\varphi} \cdot \vec{r}}{r^3} & c\varphi_m &= \frac{1}{4\pi\epsilon_0} \frac{\vec{\mu} \cdot \vec{r}}{cr^3} \\
 \vec{E} &= \frac{1}{4\pi\epsilon_0} \frac{3(\vec{\varphi} \cdot \vec{r})\vec{r} - \vec{\varphi}r^2}{r^5} + \frac{2\vec{\varphi}}{3\epsilon_0} \delta^3(\vec{r}) & c\vec{B} &= \frac{1}{4\pi\epsilon_0} \frac{3(\vec{\mu} \cdot \vec{r})\vec{r} - \vec{\mu}r^2}{cr^5} + \frac{2\vec{\mu}}{3\epsilon_0 c} \delta^3(\vec{r})
 \end{aligned}$$

Table 10.4: Electromagnetics II: Electromagnetostatic solutions.

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

$$\boxed{\text{electric potential of a point charge: } \varphi = \frac{q}{4\pi\epsilon_0 r}} \quad (10.31)$$

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

$$\text{spherical charge around the origin: } \rho = \begin{cases} \frac{q}{\frac{4}{3}\pi\varepsilon^3} & \text{if } r \leq \varepsilon \\ 0 & \text{if } r > \varepsilon \end{cases} \quad (10.32)$$

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.

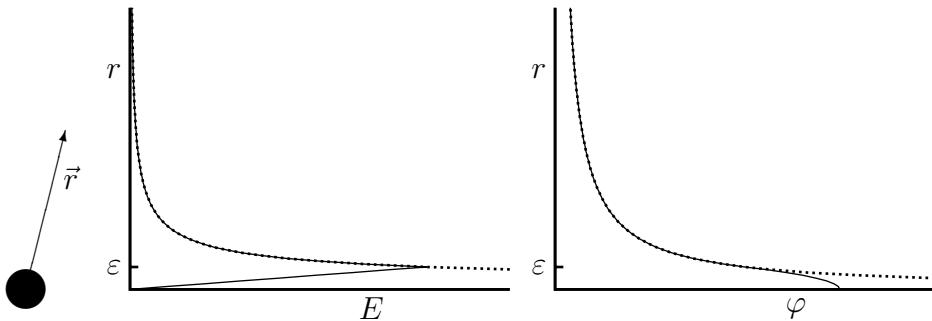


Figure 10.13: Electric field and potential of a charge that is distributed uniformly within a small sphere. The dotted lines indicate the values for a point charge.

Figure 10.13 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

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 10.3 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.

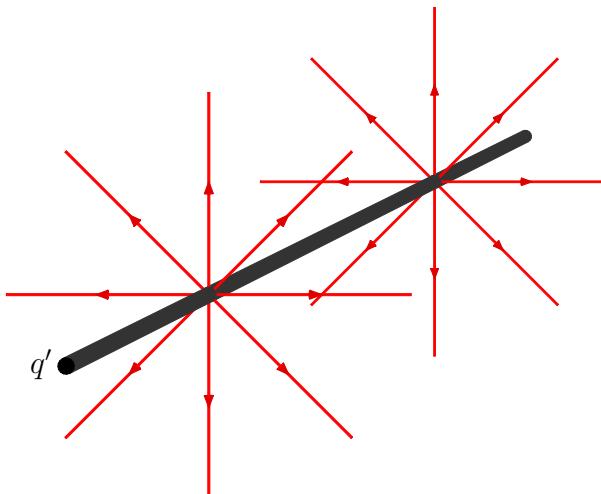


Figure 10.14: Electric field of a two-dimensional line charge.

A homogeneous distribution of charges along an infinite straight line is called a line charge. As shown in figure 10.14, 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 10.3; 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.

10.5.2 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.

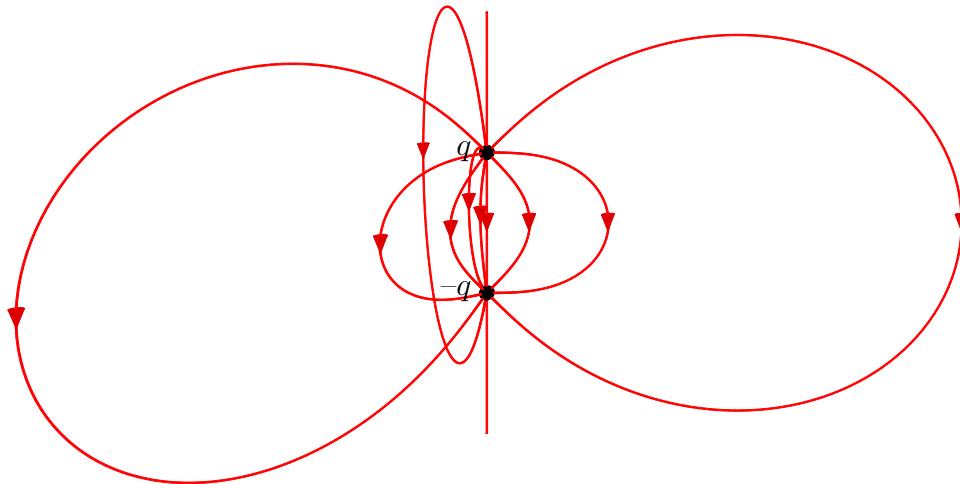


Figure 10.15: Field lines of a vertical electric dipole.

Figure 10.15 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” $\vec{\rho}$ is defined as the product of the charge strength q times the connecting vector from negative to positive charge:

$$\text{electric dipole moment: } \vec{\rho} = q(\vec{r}_+ - \vec{r}_-) \quad (10.33)$$

where \vec{r}_+ and \vec{r}_- 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:

$$\text{potential of an electric dipole: } \varphi = \frac{q}{4\pi\epsilon_0} \frac{1}{|\vec{r} - \vec{r}_+|} - \frac{q}{4\pi\epsilon_0} \frac{1}{|\vec{r} - \vec{r}_-|} \quad (10.34)$$

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

$$\text{field of an electric dipole: } \vec{E} = \frac{q}{4\pi\epsilon_0} \frac{\vec{r} - \vec{r}_\oplus}{|\vec{r} - \vec{r}_\oplus|^3} - \frac{q}{4\pi\epsilon_0} \frac{\vec{r} - \vec{r}_\ominus}{|\vec{r} - \vec{r}_\ominus|^3} \quad (10.35)$$

To obtain that result from taking 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:

$$\frac{\partial |\vec{r} - \vec{r}_0|^n}{\partial r_i} = n|\vec{r} - \vec{r}_0|^{n-2}(r_i - r_{0,i}) \quad \nabla_{\vec{r}} |\vec{r} - \vec{r}_0|^n = n|\vec{r} - \vec{r}_0|^{n-2}(\vec{r} - \vec{r}_0) \quad (10.36)$$

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 10.15.

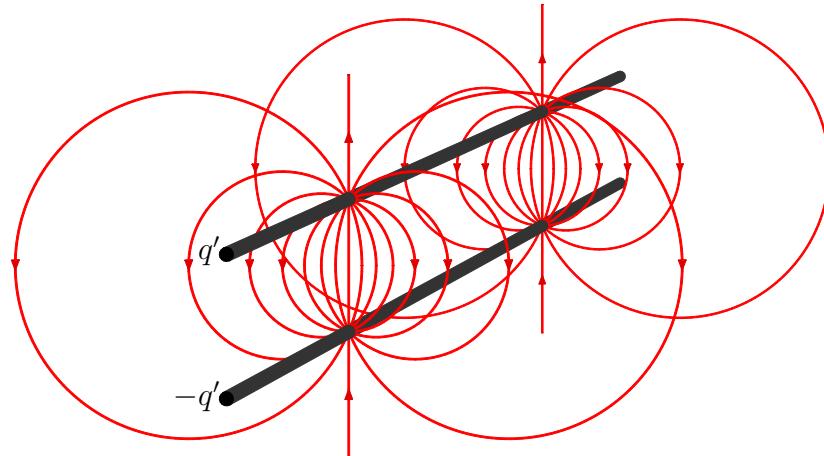


Figure 10.16: Electric field of a two-dimensional dipole.

Corresponding expressions can also be written down in two dimensions, for opposite charges distributed along parallel straight lines. Figure 10.16 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

$$\text{magnetic dipole moment: } \vec{\mu} = q_m(\vec{r}_\oplus - \vec{r}_\ominus) \quad (10.37)$$

remains finite. So you can think of an ideal magnetic dipole as infinitely strong magnetic poles infinitely close together.

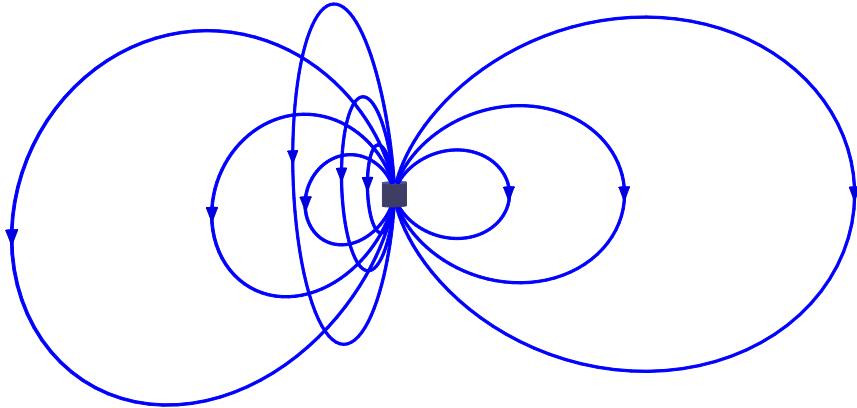


Figure 10.17: Field of an ideal magnetic dipole.

The field lines of a vertical ideal magnetic dipole are shown in figure 10.17. Their egg shape is in spherical coordinates described by, {A.102},

$$r = r_{\max} \sin^2 \theta \quad \phi = \text{constant} \quad (10.38)$$

To find the magnetic field itself, start with the magnetic potential of a non-ideal 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

$$\varphi_m = f(\vec{r} - \vec{r}_\oplus) - f(\vec{r}) \quad f(\vec{r}) = \frac{q_m}{4\pi\epsilon_0 c^2} \frac{1}{|\vec{r}|}$$

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

$$f(\vec{r} - \vec{r}_\oplus) - f(\vec{r}) \sim -\vec{r}_\oplus \cdot \nabla f \quad \text{for } \vec{r}_\oplus \rightarrow 0$$

From this the magnetic potential of an ideal dipole at the origin can be found by using the expression (10.36) 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

$$\text{potential of an ideal magnetic dipole: } \varphi_m = \frac{1}{4\pi\epsilon_0 c^2} \frac{\vec{\mu} \cdot \vec{r}}{r^3} \quad (10.39)$$

The corresponding magnetic field can be found as minus the gradient of the potential, using again (10.36) and the fact that the gradient of $\vec{\mu} \cdot \vec{r}$ is just $\vec{\mu}$:

$$\vec{B} = \frac{1}{4\pi\epsilon_0 c^2} \frac{3(\vec{\mu} \cdot \vec{r})\vec{r} - \vec{\mu}r^2}{r^5} \quad (10.40)$$

Similar expressions can be written down for ideal electric dipoles and in two-dimensions. They are listed in tables 10.3 and 10.4. (The delta functions will be discussed in the next subsection.)

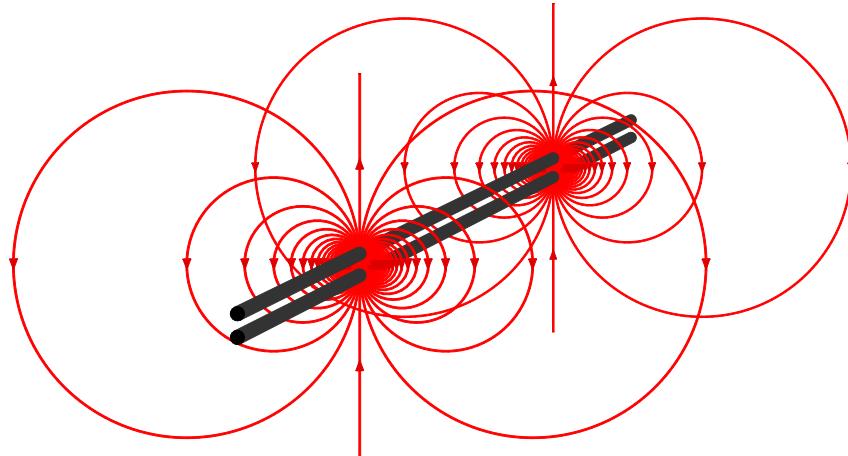


Figure 10.18: Electric field of an almost ideal two-dimensional dipole.

Figure 10.18 shows an *almost* ideal two-dimensional electric dipole. The spacing between the charges has been reduced significantly compared to that in figure 10.16, 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.

10.5.3 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,

$$q_{\text{region}} = \int_{\text{region}} \rho(\vec{r}) d^3\vec{r} \quad (10.41)$$

As far as the potential is concerned, each little piece $\rho(\vec{r}) d^3\vec{r}$ 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,

$$\varphi(\vec{r}) = \frac{1}{4\pi\epsilon_0} \int_{\text{all } \underline{\vec{r}}} \frac{1}{|\vec{r} - \underline{\vec{r}}|} \rho(\underline{\vec{r}}) d^3\underline{\vec{r}} \quad (10.42)$$

The electric field and similar expression for magnetic charge distributions and in two dimensions may be found in table 10.4

Note that when the integral expression for the potential is differentiated to find the electric field, as in table 10.4, 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

$$E_x = -\frac{\partial \varphi}{\partial x} = -\frac{1}{4\pi\epsilon_0} \int_{\text{all } \underline{\vec{r}}} \frac{1}{|\vec{r} - \underline{\vec{r}}|} \frac{\partial \rho(\underline{\vec{r}})}{\partial \underline{x}} d^3\underline{\vec{r}}$$

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, \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:

$$\varphi = \frac{1}{4\pi\epsilon_0} \int_{\text{all } \underline{\vec{r}}} \left[\frac{1}{|\vec{r}|} + \frac{\vec{r}}{|\vec{r}|^3} \cdot \underline{\vec{r}} + \dots \right] \rho(\underline{\vec{r}}) d^3\underline{\vec{r}}$$

where (10.36) was used to evaluate the gradient of $1/|\vec{r} - \underline{\vec{r}}|$ 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

$$\varphi = \frac{q}{4\pi\epsilon_0 r} + \frac{1}{4\pi\epsilon_0} \frac{\vec{\phi} \cdot \vec{r}}{r^3} + \dots \quad q \equiv \int_{\text{all } \underline{\vec{r}}} \rho(\underline{\vec{r}}) d^3\underline{\vec{r}} \quad \vec{\phi} \equiv \int_{\text{all } \underline{\vec{r}}} \underline{\vec{r}} \rho(\underline{\vec{r}}) d^3\underline{\vec{r}} \quad (10.43)$$

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 (10.43) 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 $1/\varepsilon |\vec{r} - \underline{\vec{r}}|^2$, 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 note {A.102} and shown in table 10.4. 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.

10.5.4 Solution of the Poisson equation

The previous subsections stumbled onto the solution of an important mathematical problem, the Poisson equation. The Poisson equation is

$$\nabla^2 \varphi = f \quad (10.44)$$

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 (10.42):

$$\nabla^2 \varphi = f \implies \varphi = \int_{\text{all } \vec{r}} G(\vec{r} - \vec{r}) f(\vec{r}) d^3 \vec{r} \quad G(\vec{r}) = -\frac{1}{4\pi |\vec{r}|} \quad (10.45)$$

The function $G(\vec{r} - \vec{r})$ 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 \vec{r} . The integral solution of the Poisson equation can therefore be understood as dividing function f up into spikes $f(\vec{r}) d^3 \vec{r}$; 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,

$$\nabla^2 G(\vec{r}) = \delta^3(\vec{r}) \quad G(\vec{r}) = -\frac{1}{4\pi |\vec{r}|} \quad (10.46)$$

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 10.13, 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:

$$\varphi = -\frac{1}{4\pi r} \int_{\text{all } \vec{r}} f(\vec{r}) d^3\vec{r} - \frac{\vec{r}}{4\pi r^3} \cdot \int_{\text{all } \vec{r}} \vec{r} f(\vec{r}) d^3\vec{r} + \dots \quad (10.47)$$

(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.)

10.5.5 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 ;

$$I \equiv q'v \quad (10.48)$$

The current density \vec{j} 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.

As shown in figure 10.19, 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 $2\pi rB$; 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

$$\text{magnetic field of an infinite straight wire: } B = \frac{I}{2\pi\epsilon_0 c^2 r} \quad (10.49)$$

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 10.20 shows the field produced by a single wire loop, in vacuum. To find the

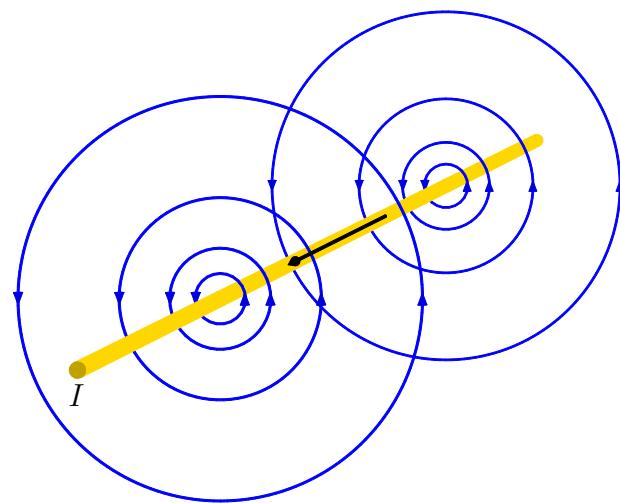


Figure 10.19: Magnetic field lines around an infinite straight electric wire.

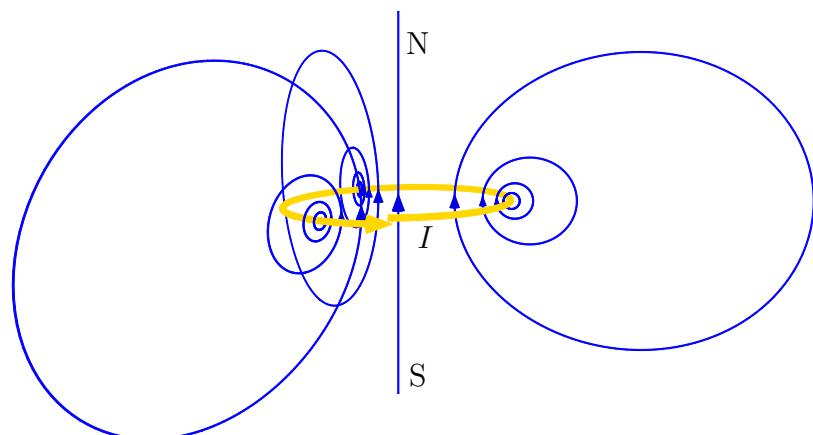


Figure 10.20: An electromagnet consisting of a single wire loop. The generated magnetic field lines are in blue.

fields produced by curved wires, use the so-called “Biot-Savart law” listed in table 10.4 and derived in {A.102}. You need it when you end up writing a book on quantum mechanics and have to plot the field.

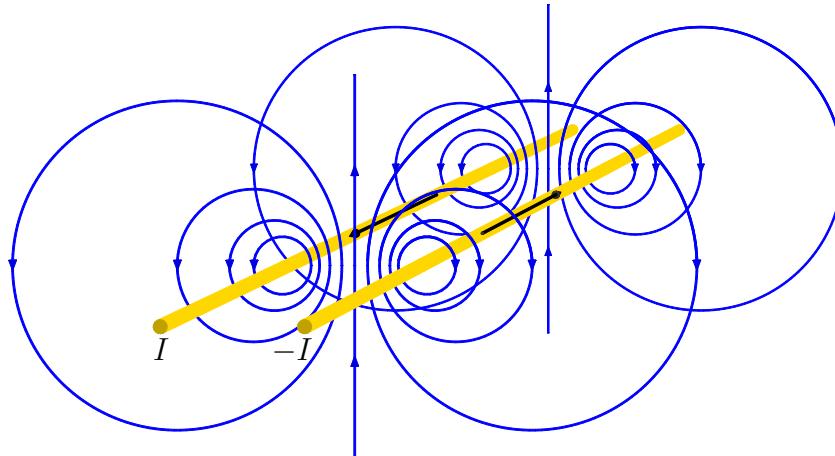


Figure 10.21: A current dipole.

Of course, while figure 10.20 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 10.21, 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 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, {A.102}.)

10.5.6 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{B} experiences a force pushing it

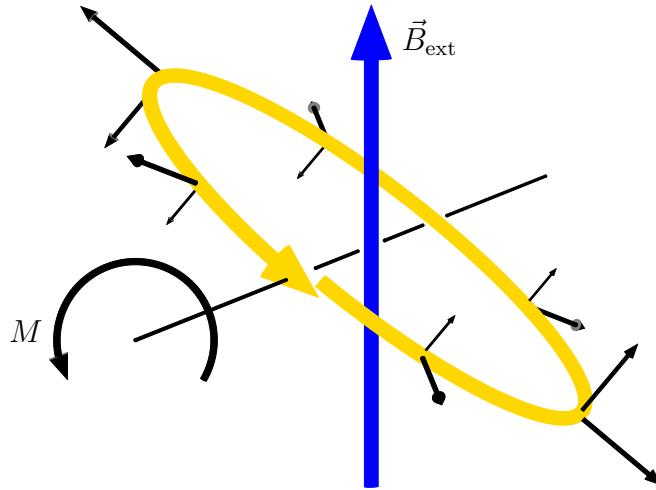


Figure 10.22: Electric motor using a single wire loop. The Lorentz forces (black vectors) exerted by the external magnetic field on the electric current carriers in the wire produce a net moment M on the loop. The self-induced magnetic field of the wire and the corresponding radial forces are not shown.

sideways equal to

$$\vec{F} = q\vec{v} \times \vec{B}$$

Consider the wire loop in an external magnetic field sketched in figure 10.22. 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 10.23, the component of the magnetic field parallel to the loop equals $B_{\text{ext}} \sin \theta$. It causes a sideways force on each element $rd\phi$ of the wire equal to

$$dF = \underbrace{q'rd\phi v}_{dq} \underbrace{B_{\text{ext}} \sin \theta \sin \phi}_{\vec{v} \times \vec{B}_{\text{parallel}}}$$

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

$$dM = \underbrace{r \sin \phi}_{\text{arm}} \underbrace{q'rd\phi v B_{\text{ext}} \sin \theta \sin \phi}_{\text{force}}$$

produces

$$M = \pi r^2 q' v B_{\text{ext}} \sin \theta$$

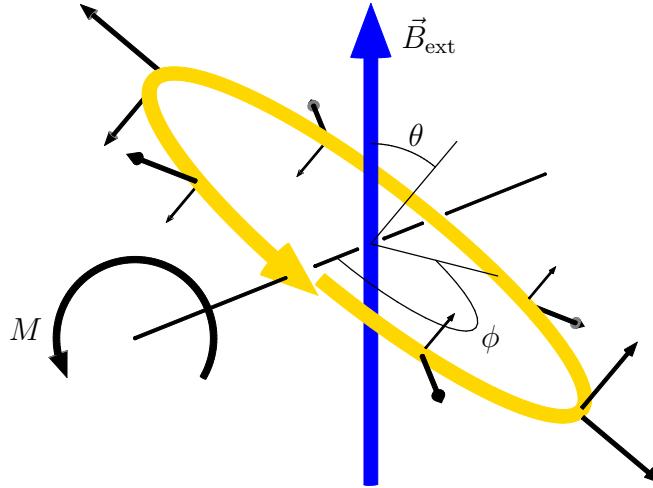


Figure 10.23: Variables for the computation of the moment on a wire loop in a magnetic field.

If the work $Md\theta$ done by this moment is formulated as a change in energy of the loop in the magnetic field, that energy is

$$E_{\text{ext}} = -\pi r^2 q' v B_{\text{ext}} \cos \theta$$

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

$$\vec{M} = \vec{\mu} \times \vec{B}_{\text{ext}} \quad E_{\text{ext}} = -\vec{\mu} \cdot \vec{B}_{\text{ext}}$$

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_c/q_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

$$\vec{\mu} = \frac{q_c}{2m_c} \vec{L} \tag{10.50}$$

Usually the current carrier is an electron, so $q_c = -e$ and $m_c = m_e$.

These results apply to any arbitrary current distribution, not just a circular wire loop. Formulae are in table 10.4 and general derivations in note {A.102}.

10.6 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{B} affects a quantum system, like say an electron in an hydrogen atom.

Well, if the Hamiltonian (10.20) for a charged particle is written out and cleaned up, {A.103}, it is seen that a constant magnetic field adds two terms. The most important of the two is

$$H_{BL} = -\frac{q}{2m}\vec{B} \cdot \hat{\vec{L}} \quad (10.51)$$

where q is the charge of the particle, m its mass, \vec{B} the external magnetic field, assumed to be constant on the scale of the atom, and $\hat{\vec{L}}$ 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 10.12 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

$$\vec{\mu}_L \equiv \frac{q}{2m}\vec{L}. \quad (10.52)$$

In terms of this magnetic dipole moment, the energy is

$$H_{BL} = -\vec{\mu}_L \cdot \vec{B}. \quad (10.53)$$

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,

$$\gamma_L = \frac{q}{2m} \quad (10.54)$$

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”

$$\mu_B = \frac{e\hbar}{2m_e} \approx 9.274 \cdot 10^{-24} \text{ J/T} \quad (10.55)$$

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{B} , and $\hat{\vec{L}}$, 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, {A.104}. The energy that an electron picks up in a magnetic field due to its inherent spin is:

$$H_{BS} = -g_e \frac{q}{2m_e} \vec{B} \cdot \hat{\vec{S}} \quad g_e \approx 2 \quad q = -e \quad (10.56)$$

(This section uses again S rather than L 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,” [34, p. 222].

It should be noted that really the g -factor of an electron is about 0.1% larger than 2 because of interaction with the quantized electromagnetic field ignored in the Dirac equation. This quantized electromagnetic field, whose particle is the photon, has a ground state energy that is nonzero even in vacuum, much like a harmonic oscillator has a nonzero ground state energy. 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, [12], is called the “anomalous magnetic moment,” [17, 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, [12], 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 $\frac{1}{2}$ 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

$$\gamma_s = g \frac{q}{2m} \quad (10.57)$$

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”

$$\mu_N = \frac{e\hbar}{2m_p} \approx 5.050\,78 \cdot 10^{-27} \text{ J/T} \quad (10.58)$$

that is based on proton charge and mass.

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

$$H_{BD} = \frac{q^2}{8m} (\vec{B} \times \hat{\vec{r}})^2 \quad (10.59)$$

Note that a system, like an atom, minimizes this contribution by staying away from magnetic fields: it is positive and proportional to 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:

$$\mu_B = \frac{e\hbar}{2m_e} \approx 5.788 \cdot 10^{-5} \text{ eV/T} \quad \frac{e^2 a_0^2}{8m_e} = 6.1565 \cdot 10^{-11} \text{ eV/T}^2$$

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.

10.7 Stern-Gerlach Apparatus [Descriptive]

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 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.

10.8 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.

10.8.1 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 $1/2$. 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 {A.105}. (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.

10.8.2 The Hamiltonian

The magnetic fields will be assumed to be of the form

$$\vec{B} = B_0 \hat{k} + B_1 (\hat{i} \cos \omega t - \hat{j} \sin \omega t) \quad (10.60)$$

where B_0 is the Tesla-strength primary magnetic field, 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, 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 $1/2$ nucleus in the center of this magnetic field. As discussed in section 10.6, 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

$$H = -\vec{\mu} \cdot \vec{B}$$

where $\vec{\mu}$ is called the magnetic dipole strength of the nucleus. This dipole strength is proportional to its spin angular momentum $\hat{\vec{S}}$:

$$\vec{\mu} = \gamma \hat{\vec{S}}$$

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:

$$H = -\gamma \hat{\vec{S}} \cdot \vec{B} \quad (10.61)$$

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 10.1.9,

$$\hat{S}_x = \frac{\hbar}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \hat{S}_y = \frac{\hbar}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \hat{S}_z = \frac{\hbar}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (10.62)$$

Substitution of these expressions for the spin, and (10.60) for the magnetic field into (10.61) gives after cleaning up the final Hamiltonian:

$$H = -\frac{\hbar}{2} \begin{pmatrix} \omega_0 & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_0 \end{pmatrix} \quad \omega_0 = \gamma B_0 \quad \omega_1 = \gamma B_1 \quad (10.63)$$

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.

10.8.3 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 $i\hbar\dot{\Psi} = H\Psi$, so if $\Psi = a\uparrow + b\downarrow$,

$$i\hbar \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} = -\frac{\hbar}{2} \begin{pmatrix} \omega_0 & 0 \\ 0 & -\omega_0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

The solution for the coefficients a and b of the spin-up and -down states is:

$$a = a_0 e^{i\omega_0 t/2} \quad b = b_0 e^{-i\omega_0 t/2}$$

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”,

$$|a| = |a_0| \equiv \cos(\alpha/2) \quad |b| = |b_0| \equiv \sin(\alpha/2)$$

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 be found as the inner products $\langle \Psi | \hat{S}_x \Psi \rangle$ and $\langle \Psi | \hat{S}_y \Psi \rangle$, chapter 3.3.3. Substituting the representation in terms of spinors and Pauli spin matrices, and cleaning up using the Euler formula (1.5), you get

$$\langle S_x \rangle = \frac{\hbar}{2} \sin \alpha \cos(\omega_0 t + \phi) \quad \langle S_y \rangle = -\frac{\hbar}{2} \sin \alpha \sin(\omega_0 t + \phi)$$

where ϕ 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.

The component in the x, y -plane is $\frac{1}{2}\hbar \sin \alpha$, and this component rotates around the z -axis, as shown in figure 10.24, causing the end point of the expectation angular momentum vector to sweep out a circular path around the magnetic field \vec{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.

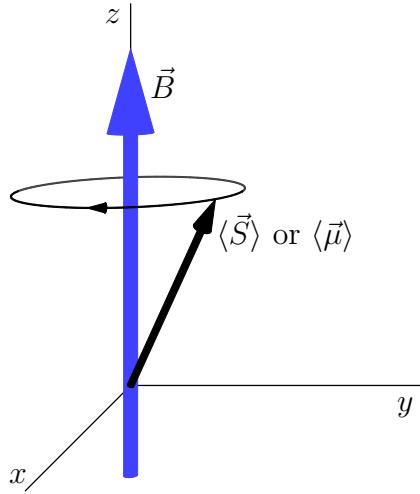


Figure 10.24: Larmor precession of the expectation spin (or magnetic moment) vector around the magnetic field.

10.8.4 Effect of the perturbation

In the presence of the perturbing magnetic field, the unsteady Schrödinger equation $i\hbar\dot{\Psi} = H\Psi$ becomes

$$i\hbar \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} = -\frac{\hbar}{2} \begin{pmatrix} \omega_0 & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \quad (10.64)$$

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 {A.106}. The analysis brings in an additional quantity that will be called the “resonance factor”

$$f = \sqrt{\frac{\omega_1^2}{(\omega - \omega_0)^2 + \omega_1^2}} \quad (10.65)$$

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:

$$a = \left[a_0 \left(\cos \left(\frac{\omega_1 t}{2f} \right) - if \frac{\omega - \omega_0}{\omega_1} \sin \left(\frac{\omega_1 t}{2f} \right) \right) + b_0 if \sin \left(\frac{\omega_1 t}{2f} \right) \right] e^{i\omega t/2} \quad (10.66)$$

$$b = \left[b_0 \left(\cos \left(\frac{\omega_1 t}{2f} \right) + if \frac{\omega - \omega_0}{\omega_1} \sin \left(\frac{\omega_1 t}{2f} \right) \right) + a_0 if \sin \left(\frac{\omega_1 t}{2f} \right) \right] e^{-i\omega t/2} \quad (10.67)$$

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

$$|b|^2 = f^2 \sin^2 \left(\frac{\omega_1 t}{2f} \right) \quad (10.68)$$

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.

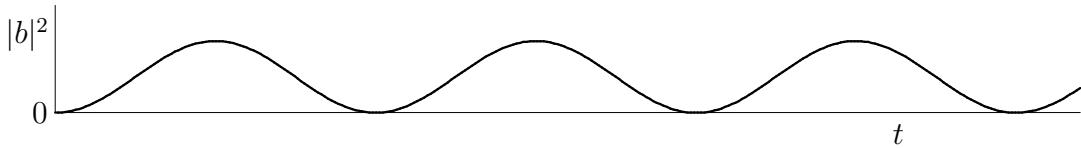


Figure 10.25: Probability of being able to find the nuclei at elevated energy versus time for a given perturbation frequency ω .

Continuing the perturbation beyond that time is bad news; it decreases the probability of elevated states again. As figure 10.25 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 sources differ about the precise definition of this frequency, but the one that seems to be most logical is ω_1/f .

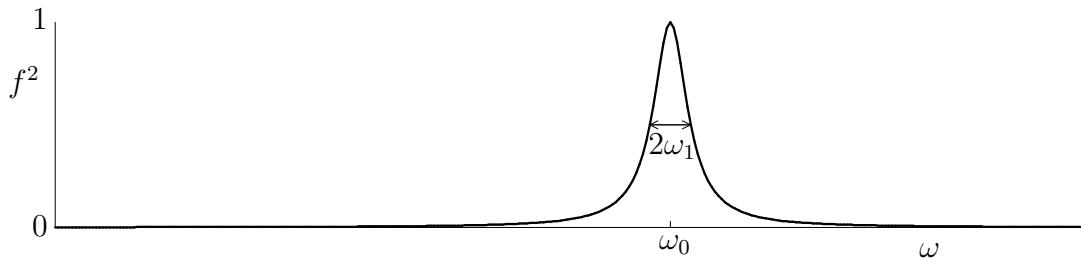


Figure 10.26: Maximum probability of finding the nuclei at elevated energy.

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 10.26. 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 10.26 must be narrow, and since its width is proportional to $\omega_1 = \gamma B_1$, that means the perturbing magnetic field must be very weak compared to the primary magnetic field.

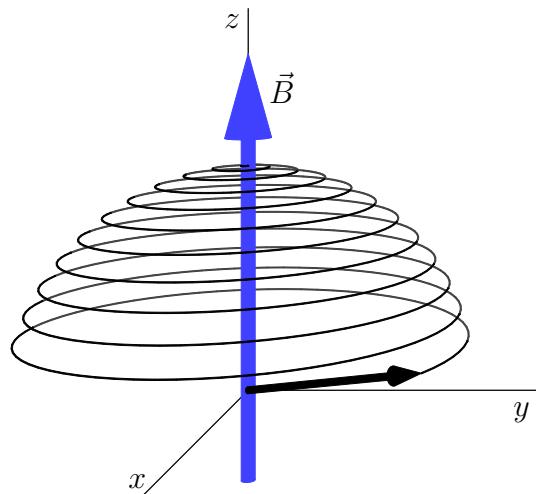


Figure 10.27: A perturbing magnetic field, rotating at precisely the Larmor frequency, causes the expectation spin vector to come cascading down out of the ground state.

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 10.27, 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 excited 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

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 11

Nuclei [Unfinished Draft]

This chapter has not been finished. I have not been able to look at it for a year, and I will presumably not be able to work much or any on it in the foreseeable future either. Since I think some parts of it are already of interest, I am posting it as is. The reader beware, much of it has been poorly proofread, if at all.

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 them with their positive electric charge Ze . But then there is nuclear energy. Here the nuclei call the shots. They are discussed in this section.

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.

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). This chapter explains some important approximate quantum models that have been developed that way, ones that work very well to explain that enormous mountain of data. A primary reference for this chapter was the popular book by Krane, [21]. That book is particularly recommended if you want an understandable description of how the experimental evidence led physicists to formulate these theoretical models.

11.1 Fundamental Concepts

Nuclei consist of protons and neutrons. Protons and neutrons are therefore called “nucleons.” Nucleons attract each other with the “nuclear force,” also called “residual strong force,” thus keeping the nucleus together. The nuclear force is strong, but it is also very short range, extending over no more than a couple of femtometers. (A fm equals 10^{-15} m and is sometimes called a fermi.) The strength of the force is about the same regardless of the type of nucleons involved, charged protons and/or uncharged neutrons. That is called “charge independence.” More restrictively, but even more accurately, the nuclear force is the same if you replace protons by neutrons and vice-versa. That is called “charge symmetry.”

The nuclear force is not a fundamental one. It is just a residual of the “color force” between the “quarks” of which protons and neutrons consist. That is much like the Van der Waals/London force between molecules is a residual of the electromagnetic force between the electrons and nuclei of which molecules exist, {A.69}. However, the theory of the color force, “quantum chromodynamics,” is well beyond the scope of this book. It is also not really important for nanotechnology. In fact, it is not really that important for nuclear engineering either because the details of the theory are uncertain, and numerical solution is intractable, [12].

11.2 The Simplest Nuclei

The simplest nucleus is the hydrogen one, just a single proton. It is trivial, 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 , equal to $1.602\,18\,10^{-19}$ C, the same as that of an electron but opposite in sign (positive). It has the same spin as an electron, $\frac{1}{2}$. And like an electron, a proton has a magnetic dipole strength; in other words, it acts as a little electromagnet. However, the proton is about 2000 times heavier than the electron. Also, the magnetic dipole strength of a proton is much less than that of an electron; the magnetic strength is roughly less by the same factor that the proton is heavier than the electron, chapter 10.6.

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 the neutron disintegrates in on average about 10 minutes, emitting an electron and an antineutrino and turning into a proton. That is called “beta decay.” Like a proton, a neutron has spin $\frac{1}{2}$. And despite the zero net charge, it too has a magnetic dipole strength. Its magnetic strength is about two thirds of that of a proton, and in the direction opposite to its spin rather than parallel to it. The explanation

for the magnetic properties is that the three quarks that make up a neutron do have charge.

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 electron-proton hydrogen-atom has been critical for deducing the quantum structure of atoms, so the 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 solved in chapter 3.2. For the hydrogen atom, spectroscopic information about the excited states, such as the Balmer series, provided a gold mine of information. Unfortunately, it turns out that the deuteron is so weakly bound that it has no excited states. If you try to excite it, it falls apart. Experimentally, the net binding energy is about 2 MeV, where a MeV is the energy that an electron would pick up in a one-million voltage difference.

The experimental size of the deuteron can be used to estimate its kinetic energy following the ideas of Heisenberg's uncertainty principle. If that is done, [21, pp. 81-83], it turns out that the kinetic energy is on the order of 30 MeV. Since the net binding energy is only 2 MeV, the expectation potential energy must be only slightly larger than the kinetic energy. Therefore, the kinetic energy is very nearly enough to tear the deuteron apart.

Still, it is hard to complain about the weak binding of the deuteron. If it would not bind at all, life as we know it would not exist. The formation of nuclei heavier than hydrogen, including carbon, begins with deuterium. Two MeV out of 30 made all the difference.

As a further complication, spin has a major effect on the force between the proton and neutron. In the hydrogen atom, that effect exists but is extremely small. In particular, in the hydrogen atom ground state, the electron and proton combine in the singlet state of zero spin; however, the triplet state of unit spin has only very slightly higher energy, chapter 12.1.6. In case of the deuteron, however, 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 or two neutrons does not exist, even though two neutrons or two protons attract each other almost the same as the proton and neutron in the deuteron. A symmetric spatial state, appropriate for a ground state, {A.25 and A.26}, needs in the case of two protons or two neutrons an asymmetric singlet spin state state to satisfy the antisymmetrization requirement, chapter 4.6. But only the triplet state is bound.

There is a further complication. Spin provides directionality. The nuclear force between the proton and neutron in the deuteron is not well aligned with the line connecting them. As a result, the Hamiltonian does not commute with orbital angular momentum, making orbital angular momentum uncertain. Experimentally, it is deduced that the deuteron has about a 95% probability

of being found in an $l = 0$ state in which the orbital angular momentum is zero and a 5% probability of being found in an $l = 2$ state in which it is not. Despite being the simplest nontrivial nucleus, even the deuteron is complicated and incompletely understood.

11.3 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 4.9. 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. To indicate the number of neutrons in a nucleus, the convention is to follow the element 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 also contains 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, the difference in chemical prop-

erties 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 6500 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 the helium-3 atom is a fermion while the helium-4 atom is a boson. Protons and neutrons have spin $\frac{1}{2}$, as do electrons, so the difference of one neutron switches the net atom spin between integer and half integer. That turns a bosonic atom into a fermionic one or vice-versa. At extremely low temperatures it makes a difference, chapter 9.

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_1H , hydrogen-2 by ^2_1H , helium-3 by ^3_2He , and helium-4 by ^4_2He .

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 be found by subtracting the atomic number from the mass number,

$$N = A - Z \quad (11.1)$$

so this book will leave it out. It may also be noted that the atomic number is 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 by heart.

Isotopes have the same chemical symbol and atomic number, just a different mass number. However, deuterium is often indicated by chemical symbol D instead of H, because 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_1H . 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 not. Figure 11.1 shows the stable nuclei as green squares. The leftmost green square in the bottom row 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 are helium-3 and helium-4 respectively.

More generally, isotopes are found on the same horizontal line in the figure. Selected values of the atomic number Z are shown as labeled horizontal lines.

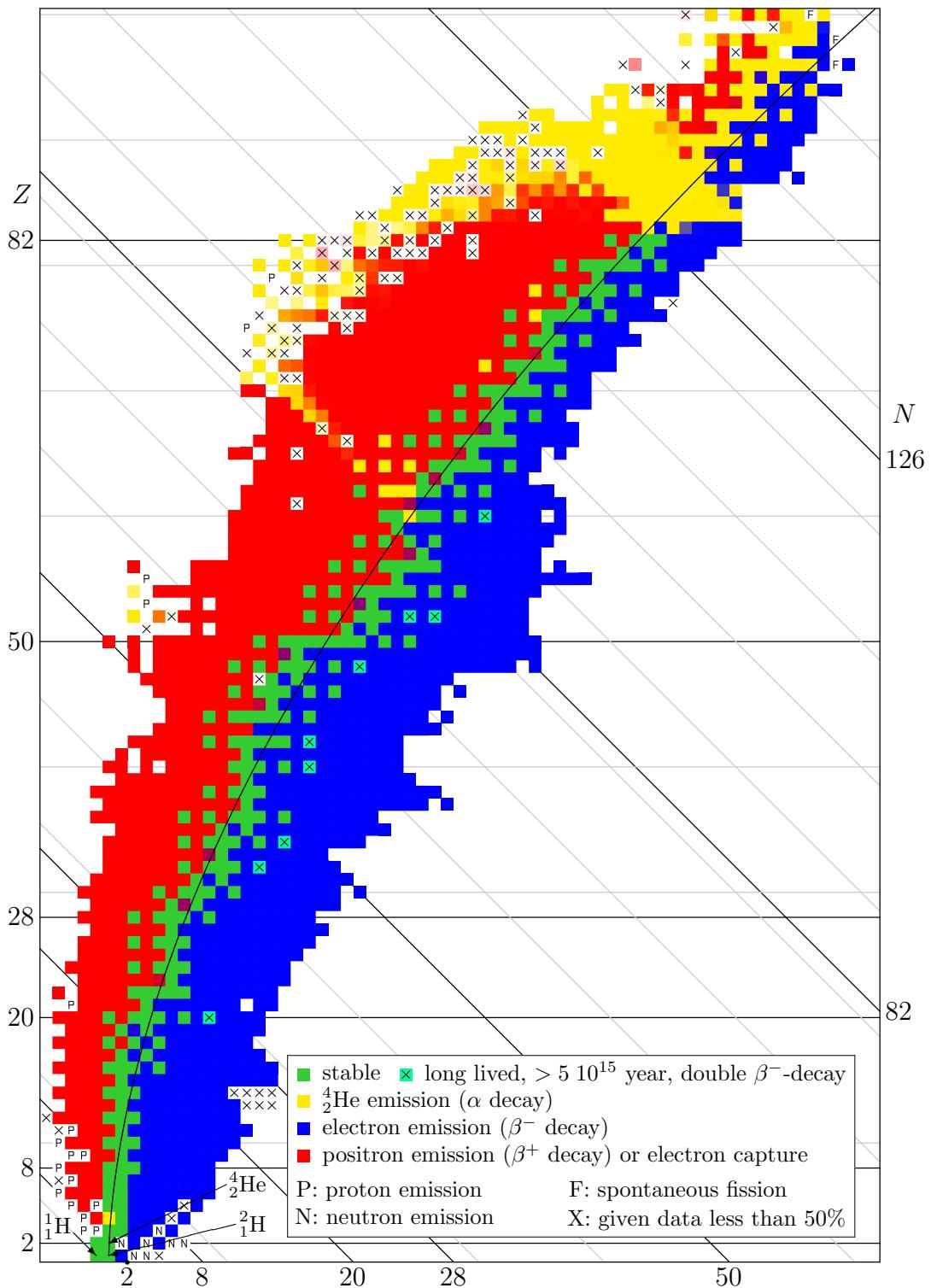


Figure 11.1: Nuclear decay modes.

Similar, selected values of the neutron number N are shown as labeled lines sloping down at 45 degrees. Nuclei with the same number of neutrons are called “isotones.” How clever, to replace the p in isotopes with an n.

The horizontal position of each square in figure 11.1 indicates the “neutron excess” $N - Z$. For example, hydrogen-2 and helium-4 both have neutron excess zero, they have equal numbers of protons and neutrons. So they are at the same horizontal position 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 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 ${}^2_1\text{H}$ in figure 11.1. That is ${}^3_1\text{H}$, hydrogen-3 or tritium. It is unstable. After on average about twelve years, it will emit an electron. The electron carries away one unit of negative charge; that turns a neutron with zero net charge into a positively charged proton. So hydrogen-3, with one proton and two neutrons, changes into helium-3, with two protons and one neutron. The mass number has stayed the same but the atomic number has increased one unit. In terms of figure 11.1, the nucleus has changed into one that is one place up and two places to the left.

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 emitted radiation was simply 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 minuscule 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 11.1 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 instead, they can turn them 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 neutron.

However, a nucleus has a much easier way to get rid of one unit of net positive charge: it can swipe an electron from the atom. This is called “electron capture” (EC). 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. It is also called K-capture or L-capture, depending on the electron shell from which the electron is swiped.

Of course, “inverse beta decay” is not inverse beta decay, because in beta decay the emitted electron does not go into an empty atomic orbit, and an antineutrino is emitted instead of a neutrino absorbed.

The term “beta-plus decay” (β^+ -decay) usually refers to positron emission, but NUBASE 2003 uses it 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 11.1. 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.

In either beta-minus or beta-plus decay, the mass number A does not change. Nuclei with the same mass number are called “isobars.” Yes, this conflicts with the established usage of the word isobar for constant pressure line, but in this case physicists have blown it. There is not likely to be any resulting confusion unless there is a nuclear winter.

There are a variety of other ways in which nuclei may decay. As shown in figure 11.1, if the number of protons or neutrons is really excessive, the nucleus may just kick the bums out instead of convert them.

Similarly, heavy nuclei that are weakened by Coulomb repulsions tend to just throw some nucleons out. Commonly, a ${}^4_2\text{He}$ 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 figure 11.1.

If nuclei are really oversized, they may just disintegrate completely; that is called spontaneous fission.

Another process, “gamma decay,” is much like spontaneous decay of excited electron levels in atoms. In gamma decay an excited nucleus transitions to a lower energy state and emits the released energy as very energetic electromagnetic radiation. The nucleus remains of the same type: there is no change in the number of protons or neutrons. Nuclear emissions are commonly associated with additional gamma radiation, since the emission tends to leave the nucleus in an excited state. Gamma decay as a separate process is often referred to as an “isomeric transition” (IT) or “internal transition.”

A second way that a 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 not part of the nucleus completely out of the atom than “*internal conversion*” (IC)? Internal conversion is usually

included in the term isomeric transition.

Figure 11.1 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 lists that contribution as unknown and only mentions that beta-minus decay is negligible.

RE	RaA	RaB	RaC	RaC1	RaC2	RaD	RaE	RaF
$^{222}_{86}\text{Rn}$	$^{218}_{84}\text{Po}$	$^{214}_{82}\text{Pb}$	$^{214}_{83}\text{Bi}$	$^{214}_{84}\text{Po}$	$^{210}_{81}\text{Tl}$	$^{210}_{82}\text{Pb}$	$^{210}_{83}\text{Bi}$	$^{210}_{84}\text{Po}$

Table 11.1: Alternate names for nuclei.

Since so many outsiders know what nuclear symbols mean, physicists prefer to use obsolete names to confuse them. Table 11.1 has a list of names used. The abbreviations refer to historical names for decay products of radium (radium emanation, radium A, etc.)

If you look closer at which nuclei are stable or not, it is seen that stability tends to be enhanced if the number of protons and/or neutrons is even and reduced if it is odd. Physicists therefore speak of even-even nuclei, even-odd nuclei, etcetera. Note that if the mass number A is odd, the nucleus is either even-odd or odd-even. If the mass number is even, the nucleus is either even-even or odd-odd. Odd mass number nuclei tend to be easier to analyze.

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 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 are called isotones, and nuclei with the same total number of nucleons are called isobars.

- o— For an example notation, consider ${}^4_2\text{He}$. It indicates a helium atom nucleus consisting of 4 nucleons, the left superscript, of which 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. This nucleus is called helium-4, where the 4 is again the number of nucleons.
 - o— Nuclei can decay by various mechanisms. To promote confusion, emission of helium-4 nuclei is called alpha decay. Emission of electrons is called beta decay, or β decay, or beta-minus decay, or β^- decay, or negatron emission, or negaton emission, but never electron emission. Emission of positrons (positons) may be called beta-plus decay, or β^+ decay, but beta-plus decay might be used to also indicate electron capture, depending on who used the term. Electron capture may also be called K-capture or L-capture or even inverse beta decay, though it is not. Emission of electromagnetic radiation is called gamma decay or γ decay. 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.
 - o— No, this is not a story made up by this book to put physicists in a bad light.
 - o— Odd mass numbers correspond to even-odd or odd-even nuclei. Even mass numbers correspond to either even-even nuclei, which tend to have relatively high stability, or to odd-odd ones, which tend to have relatively low stability.
-

11.4 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 11.1:

$$\text{magic numbers: } 2, 8, 20, 28, 50, 82, 126, \dots \quad (11.2)$$

These 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 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 the 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 presence will further be apparent throughout the figures in this chapter.

11.5 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 what they are and the related units.

11.5.1 Decay rate

If a large number I of unstable nuclei of the same kind are examined, then the number of nuclei that decays during an infinitesimally small time interval dt is given by

$$\boxed{dI = -\lambda I dt} \quad (11.3)$$

where λ is a constant often called the “specific decay rate” of the nucleus. If the amount of nuclei at time zero is I_0 , then at an arbitrary later time t it is

$$I = I_0 e^{-\lambda t} \quad (11.4)$$

The reciprocal of the specific decay rate has units of time, and so it is called the “lifetime” of the nucleus:

$$\boxed{\tau \equiv \frac{1}{\lambda}} \quad (11.5)$$

However, it is not really a lifetime except in some average mathematical sense. Also, if more than one decay process occurs,

Add specific decay rates, not lifetimes.

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.

A physically more meaningful quantity than lifetime is the time for about half the nuclei in a given sample to disappear. This time is called the “half-life” $\tau_{1/2}$. From the exponential expression (11.4) above, it follows that the half-life is shorter than the lifetime by a factor $\ln 2$:

$$\boxed{\tau_{1/2} = \tau \ln 2} \quad (11.6)$$

Note that $\ln 2$ is less than one.

For example tritium, ${}^3\text{H}$ or T, has a half life of 12.32 years. If you have a collection of tritium nuclei, after 12.32 years, only half will be left. After 24.64 years, only a quarter will remain, after a century only 0.4%, and after a millennium only $4 \cdot 10^{-23}\%$. Some tritium is continuously being created in the atmosphere by cosmic rays, but because of the geologically short half-life, there is no big accumulation. The total amount of tritium remains negligible.

11.5.2 Other definitions

You probably think that having three different names for essentially the same single quantity, the specific decay rate λ , 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. 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.

How about calling the product λI the “decay rate” or “disintegration rate” or simply the “activity?” How about “mean lifetime” instead of lifetime?

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 of course equal $3.7 \cdot 10^{10}$ decays per second. If you add 3 and 7 you get 10, not? You also have 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 Coulomb per kg. Of course, a better unit than that is needed, so the “roentgen” or “röntgen” R is defined to $2.58 \cdot 10^{-4}$ C/kg. It is important if you are made of air or want to compute how far the radiation moves in air.

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_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.

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).

11.6 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, {A.4}. 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,

$$\boxed{m_{\text{nucleus}} = Zm_p + Nm_n - \frac{E_B}{c^2}} \quad (11.7)$$

where

$$m_p = 1.672\,621\,10^{-27} \text{ kg} \quad m_n = 1.674\,927\,10^{-27} \text{ kg}$$

are the mass of a lone proton respectively a lone neutron at rest, and E_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 2.2245 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 of them are approximately equal to the number of nucleons, (within a percent or so.) The current official definition is that a carbon-12, $^{12}_6\text{C}$, 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, [24]:

$$\boxed{m_{\text{nucleus}} = m_{\text{atom}} - Zm_e + A_e Z^{2.39} + B_e Z^{5.35}} \quad (11.8)$$

$$m_e = 0.510\,998\,910 \text{ MeV} \quad A_e = 1.443\,81 \cdot 10^{-5} \text{ MeV} \quad B_e = 1.554\,68 \cdot 10^{-12} \text{ MeV}$$

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.

11.7 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_B can be found from the mass of the nucleus. The specific binding energy is defined as the binding energy per nucleon, E_B/A . Figure 11.2 shows the specific binding energy of the nuclei with known masses. The highest specific binding energy is 8.8 MeV, and occurs for $^{62}_{28}\text{Ni}$ nickel. Nickel has 28 protons, a magic number. However, nonmagic $^{58}_{26}\text{Fe}$ and $^{56}_{26}\text{Fe}$ 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 11.2 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_2He nucleus, the alpha particle. It has a stunning 7.07 MeV binding energy per nucleon, exceeding its immediate neighbors by far.

The ^8_4Be beryllium nucleus is not bad either, also with 7.07 MeV per nucleon, almost exactly as high as $^4_2\text{He}_2$, though admittedly that is achieved using eight nucleons instead of only four. But clearly, ^8_4Be 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 $^8_4\text{Be}_4$ is not. It falls apart in about 67 as ($67 \cdot 10^{-18}$ 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_4Be 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

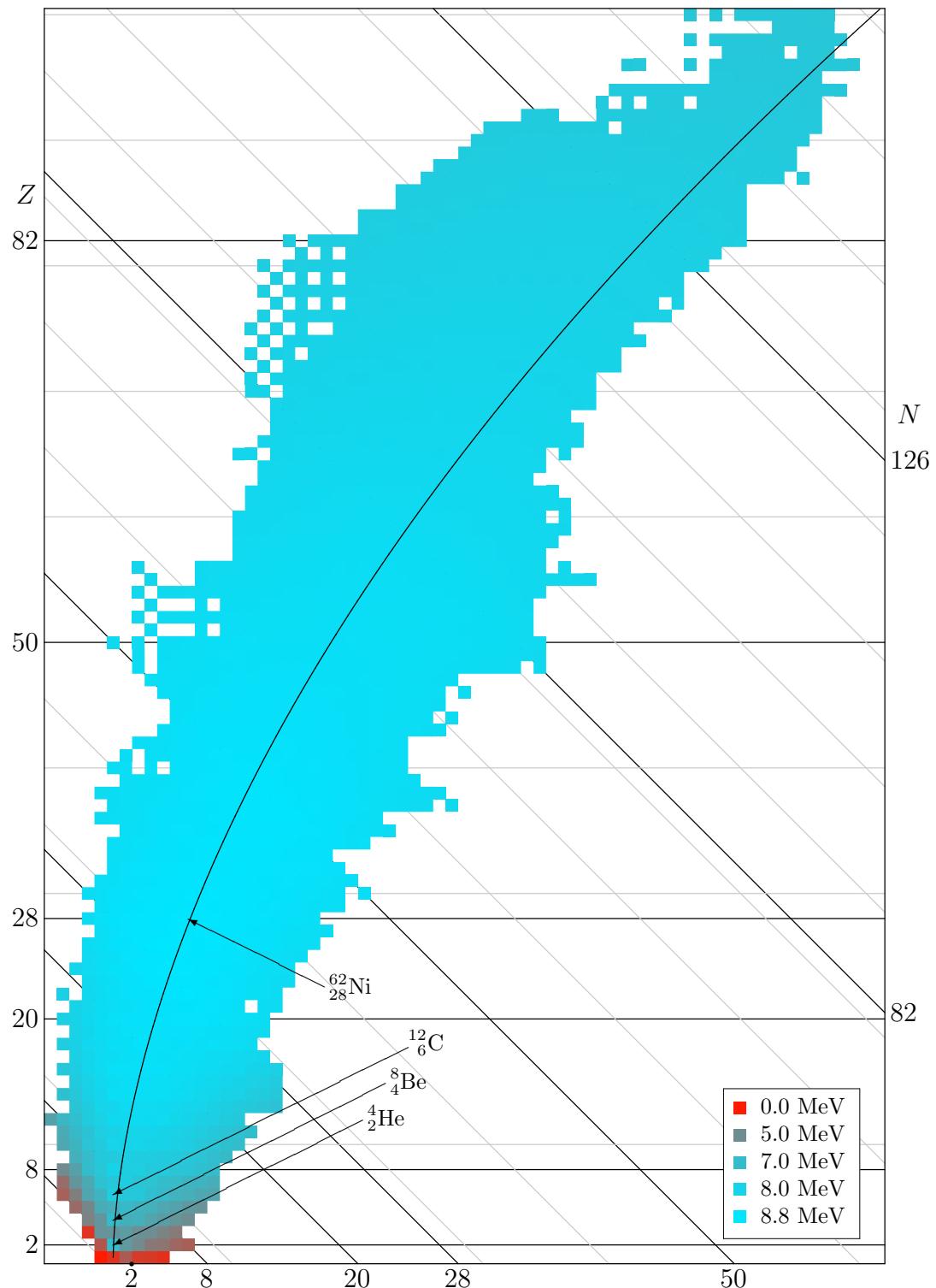


Figure 11.2: Binding energy per nucleon.

out, the carbon of life is formed in the interior of aging stars when ^8_4Be captures a third alpha particle, to produce $^{12}_6\text{C}$, 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_4Be . The process is far too slow to have occurred in the Big Bang, however.

For $^{12}_6\text{C}_6$ carbon, the superior number of nucleons has become big enough to 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.

11.8 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 11.3 and 11.4 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 11.3 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 11.4 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. This is discussed further in section 11.12.

Figures 11.5 and 11.6 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.

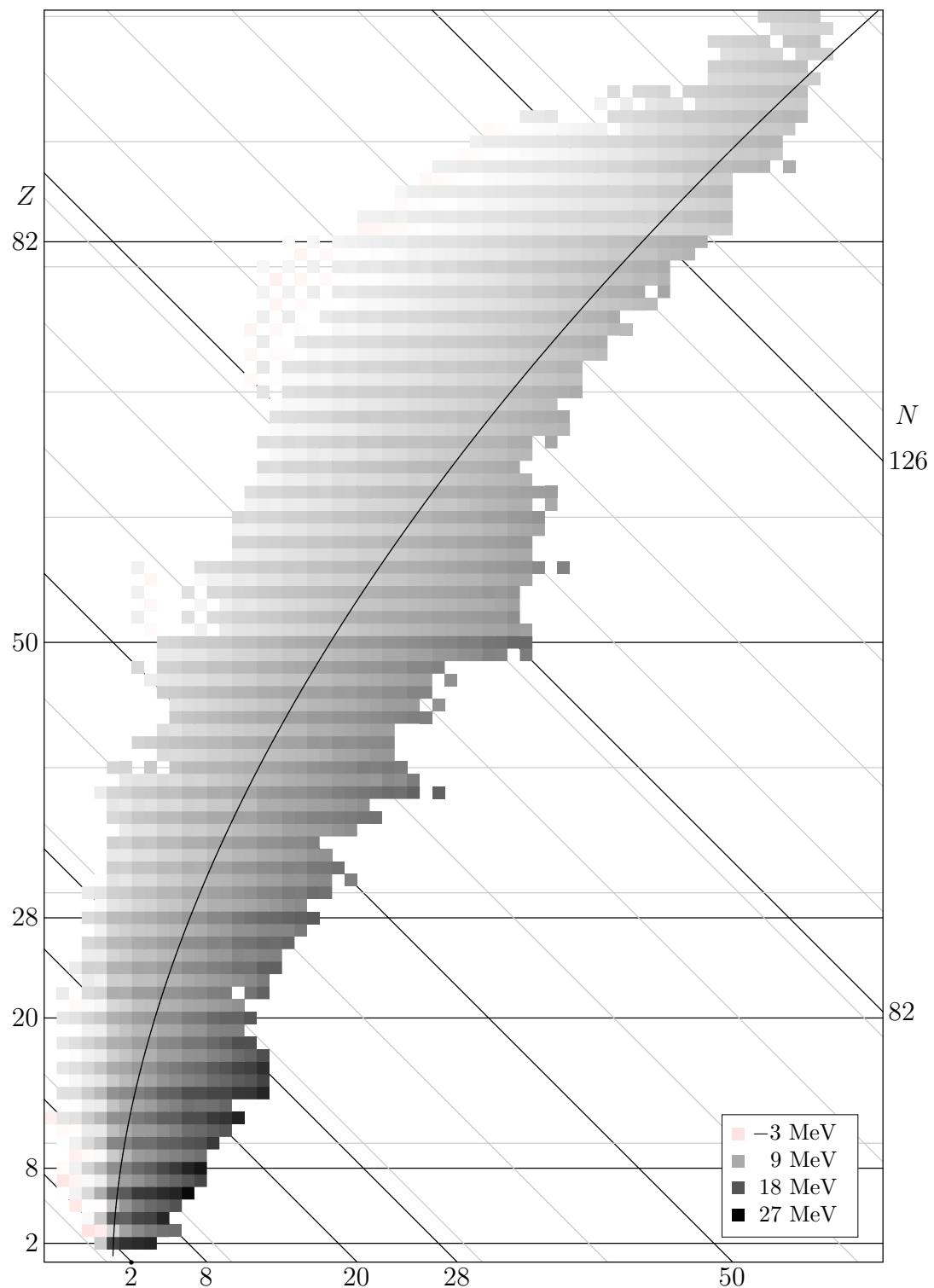


Figure 11.3: Proton separation energy.

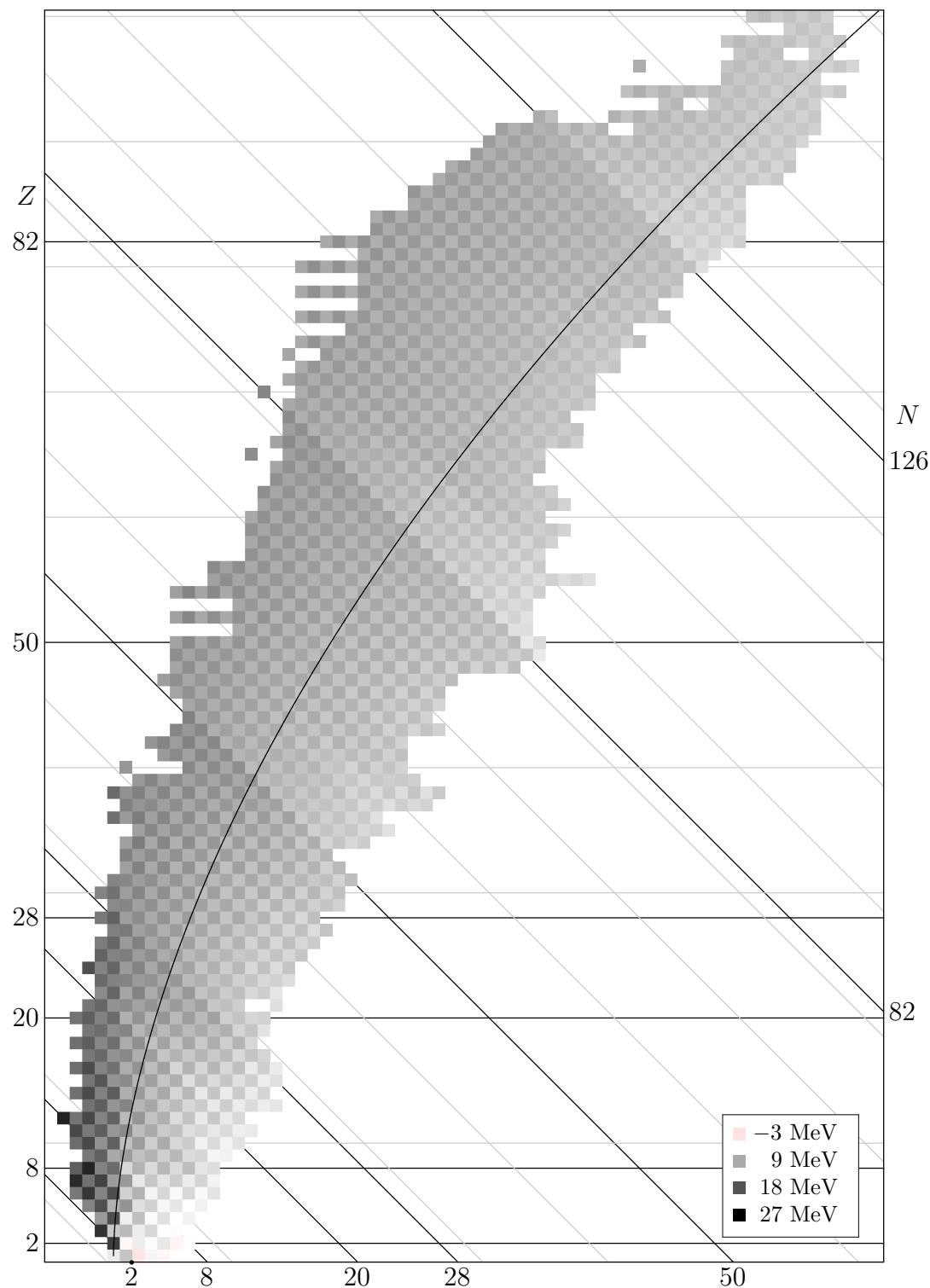


Figure 11.4: Neutron separation energy.

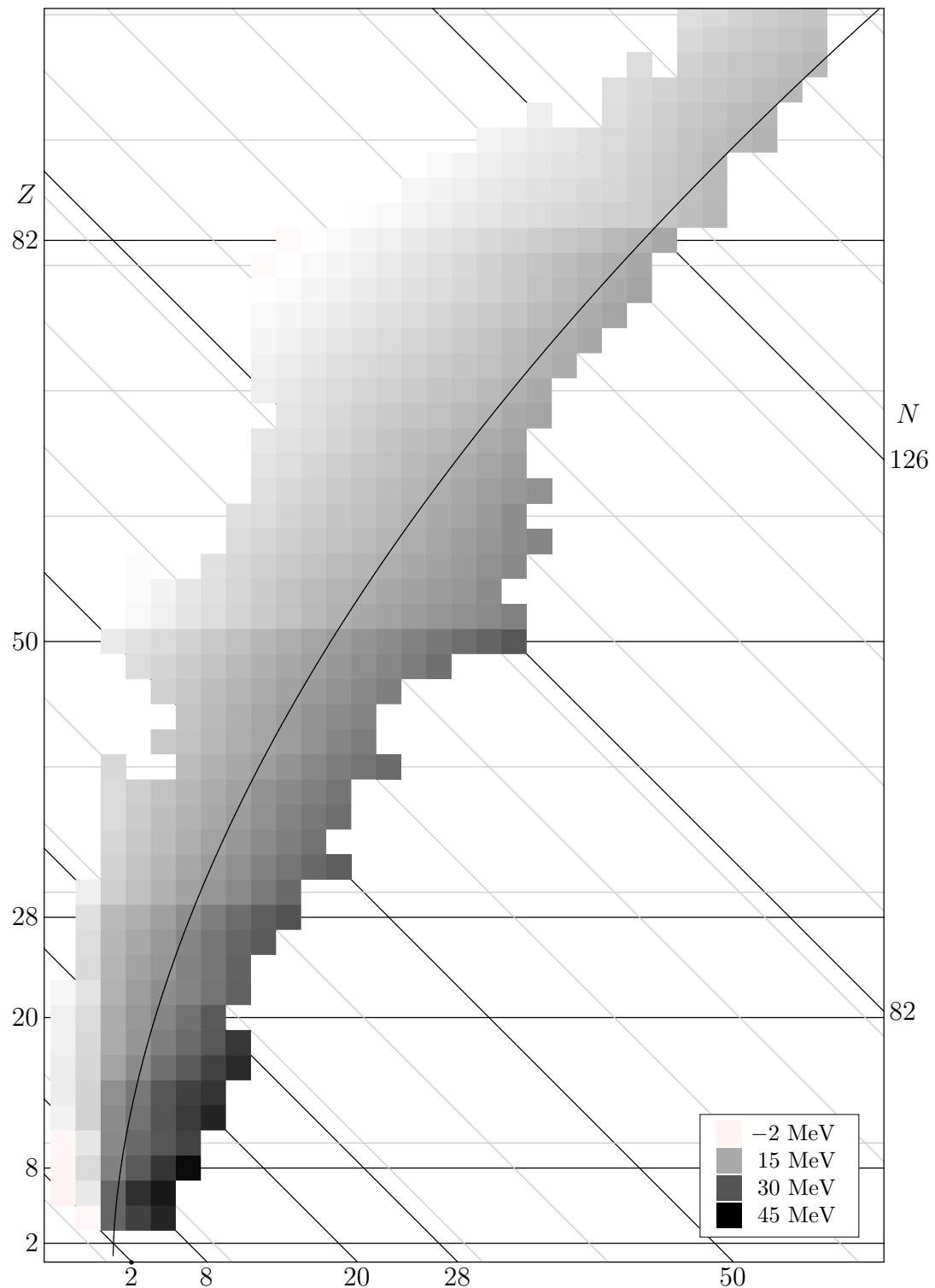


Figure 11.5: Proton pair separation energy.

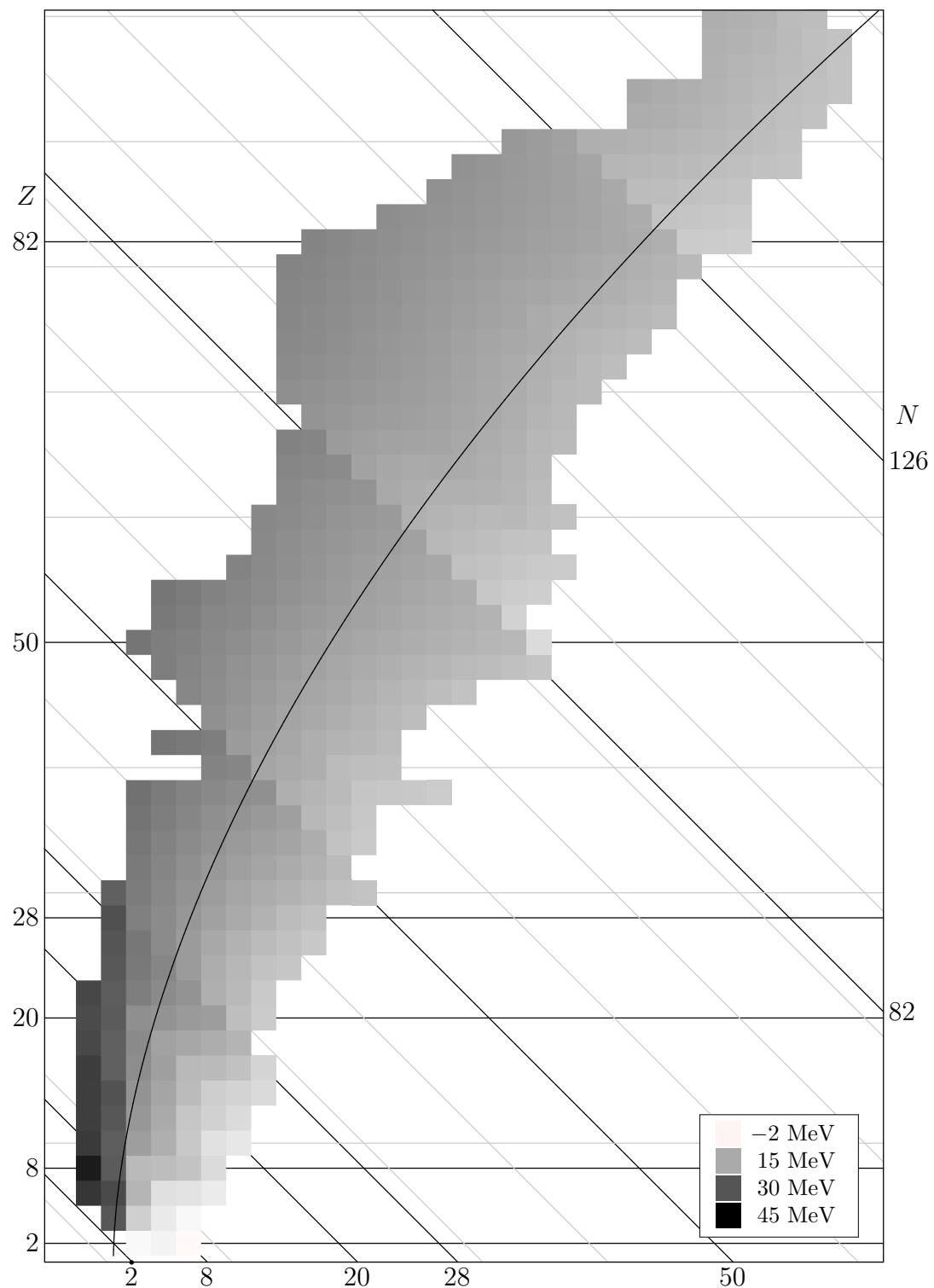


Figure 11.6: Neutron pair separation energy.

11.9 Nuclear Forces

The purpose of this section is to examine the fundamental nature of the nuclear forces. While the nuclear forces are incompletely understood, important qualitative properties can be deduced. Detailed quantitative nuclear models can also be developed, but they will not be used in this book.

It is helpful to first examine electromagnetic interactions more closely. Consider the electromagnetic interaction between two charged particles like the proton and electron in a hydrogen atom. (For discussion purposes, the proton will here be regarded as a point charge like the electron, though in truth it is a composite of three charged quarks.) It has been routinely assumed in this book that the proton and electron interact through a Coulomb potential that is inversely proportional to the distance between the two. That assumption is fine for analyzing the hydrogen atom, but it runs into major problems if there is nontrivial motion between the charges. In particular, an interaction given by the distance between the charges implies that the hydrogen electron will immediately notice it if the proton moves. But instantaneous interaction over a distance is impossible according to relativity, {A.4}. No observable effect can move faster than the speed of light.

In classical electrodynamics, this problem is fixed by assuming that the electron and proton do not interact directly. Each interacts instead with the electromagnetic field at its own location. By changing the electromagnetic field, each charged particle can however indirectly affect the other charged particle at a slightly later time. It is only under conditions in which the physics is quasi-steady that a direct Coulomb potential interaction can be used as an *approximation*. In particular, the time scale of the motion of the charged particles must be slow compared to the time that it takes the electromagnetic waves to move from one particle to the other at the speed of light.

The quantum theory of electromagnetic interactions is called “quantum electrodynamics.” Its ideas trace back to the early days of quantum mechanics, but there were some major problems with computations coming up with infinite values. A final theory 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 path-integral 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.)

In quantum electrodynamics, the electromagnetic field is quantized into discrete photons. In a typical electromagnetic interaction between a proton and an electron, the proton might create a photon, and the electron might subsequently absorb that photon. The creation of the photon by the proton is a three-particle interaction: the proton before the emission, the emitted photon, and the proton

after the emission. Similarly the absorption of the photon by the electron is a three-particle interaction: the electron before the absorption, the photon before the absorption, and the electron after the absorption. Note that the complete interaction can be described as an exchange of a virtual photon between proton and electron. (A particle is considered a virtual one if it disappears again after it is created.)

It might seem that since both the electron and the proton have distributed wave functions, they should also be able to interact directly. In particular, it might seem plausible that at a given point in space time there could be a nontrivial interaction between the electron and proton wave functions at that point. However, that would be a four particle interaction: the electron and proton before the interaction and the electron and proton after it. Having four particles at exactly the same point in space-time should have zero probability compared to having only three. So a direct proton-electron interaction has zero probability of occurring compared to the three-particle interactions of photon creation and absorption.

Following the ideas of quantum electrodynamics and pioneering work of Sheldon Glashow, Steven Weinberg and Abdus Salam in 1967 independently developed a particle exchange model for the weak force that is responsible for beta decay. All three received the Nobel prize for that work in 1979.

In weak interactions, the exchanged particles are not photons, but a set of three particles: the negatively charged W^- , (think W for weak force), the positively charged W^+ , and the 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 and the gluons of color force interactions. In fact, they have gigantic rest masses: the W^- has a rest mass of about 80 GeV and the Z^0 about 90. Compare that with the rest mass of a proton or neutron, less than 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. It meets the requirements of physics well.

The W^\pm are responsible for interactions like beta decay, while the Z^0 only produce interactions that do not involve the exchange of charge, such as neutrino scattering. In beta-minus decay, the neutron, (or more precisely a down quark in the neutron), first emits a W^- and turns into a proton, and that W^- then decays into an electron and electron antineutrino. Therefore, the decay process consists of two successive three-particle interactions: first neutron, proton, and W^- , then W^- , electron and antineutrino.

The reason it took long for this mechanism to be identified is that the W^- disintegrates almost instantly, within a distance on the order of 10^{-3} femtometer (fm). That is a negligible distance compared to the multi-femtometer size of a

nucleus. So initially it was assumed that the neutron produced the electron and antineutrino directly. That idea gave rise to the Fermi theory of beta decay. The Fermi theory is very accurate even though it ignores the W^- , section 11.19.5.

The short lifetime of the virtual W^- has to do with the question how a massive particle can suddenly appear out of nothing in the first place. For example, consider the beta decay of a lone neutron. How can this neutron, worth less than a single GeV in rest mass energy, suddenly pop up a gigantic 80 GeV W^- ? What happened to energy conservation??

Well, what happened to energy conservation is a mixture of special relativity and the uncertainty inherent in quantum mechanics. The creation of particles out of pure energy is allowed by special relativity. In particular, special relativity gives the total energy E of a particle as, {A.4} (A.7),

$$E^2 = p^2 c^2 + (mc^2)^2$$

where c is the speed of light, mc^2 is the rest mass energy of the particle, and p the momentum of the particle, which is related to its kinetic energy.

Quantum mechanics replaces the momentum \vec{p} by the operator $\hbar \nabla / i$, producing the so-called relativistic Klein-Gordon eigenvalue problem

$$-\hbar^2 c^2 \nabla^2 \psi + (mc^2)^2 \psi = E^2 \psi$$

A one GeV neutron does not have 80 GeV of energy lying around, but the Klein-Gordon equation has a nontrivial solution for $E = 0$:

$$\psi = C \frac{e^{-mc^2 r / \hbar c}}{r}$$

where C is an arbitrary constant. The radial coordinate r can here be identified with the distance from the neutron, or rather down quark, that anchors the W^- . Just plug this solution in the eigenvalue problem to check it, using the expression for the Laplacian in spherical coordinates found in any mathematical handbook, [28, p. 126]. (Actually, that is not quite right. The Laplacian of the given solution produces a delta function at $r = 0$ that would have to be cancelled by a corresponding spike of attraction at $r = 0$ between the down quark and the W^- .)

The bottom line is that according to the Klein-Gordon solution the wave function of the W^- is proportional to $e^{-mc^2 r / \hbar c} / r$. The exponential is negligibly small unless the distance from the down quark is no more than of order $\hbar c / mc^2$. With $\hbar c$ about 200 MeV fm and the W^- rest mass energy mc^2 equal to 80 GeV, that works out to around 0.0025 fm. It is the distance mentioned earlier.

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. First note that the Heisenberg uncertainty relation reads:

$$\sigma_{p_x} \sigma_x \leq \frac{1}{2}\hbar$$

where σ_{p_x} and σ_x are the standard deviations in momentum and position in the x -direction, respectively. Next note that relativity considers ct pretty much like the fourth position coordinate in addition to x , y , and z , {A.4}. Similarly, it adds E/c as in essence the fourth momentum vector component. Following that idea, the uncertainty relation seems to become

$$\sigma_E \sigma_t \leq \frac{1}{2}\hbar$$

Unfortunately, space and time are not really the same thing. So it is not quite clear what sense to make of, for example, the standard deviation in time. Careful analytical arguments like those of Mandelshtam and Tamm do succeed in giving a meaningful definition of it, but its usefulness is limited.

Ignore it. Careful analytical arguments are for wimps! Take out your pen and cross out “ σ_t .” Write in “any time difference you want.” Cross out “ σ_E ” and write in “any energy difference you want.” As long as you are at it anyway, also cross out “ \leq ” and write in “=.” This can be justified because both are mathematical symbols, hence equivalent. And inequalities are so vague anyway. You have now obtained the popular version of the Heisenberg energy-time uncertainty relation:

any energy difference you want \times any time difference you want = $\frac{1}{2}\hbar$

(11.9)

This is an extremely powerful equation that can explain anything in physics involving two quantities that have dimensions of energy and time. Be sure, however, to only publicize the cases in which it gives the right answer.

To apply it to beta decay, replace “any energy difference you want” with “the rest mass energy of the W^- .” Replace “any time difference you want” with “the lifetime of the W^- .” Note that if the W^- was not a quantum particle with uncertainty in position, it would travel with about speed $\frac{1}{2}c$ while it exists, because it cannot travel with a speed less than zero nor more than the speed of light. Put it all together, and it shows that the W^- would travel about 0.001 fm if it had a position to begin with. The result has been derived using only elementary school mathematics. Tell all your friends you heard it here first.

Returning to the solution of the Klein-Gordon equation,

$$\psi = \frac{C}{r} e^{-mc^2 r/\hbar c}$$

note that if the particle involved is a photon, which has a rest mass of zero, the exponential drops out. Then a $1/r$ amplitude of interacting with the anchored

photon field results. The radial dependence of the quasi-steady Coulomb potential is $1/r$ too. That suggests that forces mediated by a particle with a nonzero mass might be describable by a quasi-steady potential of the form

$$V = \frac{C}{r} e^{-mc^2 r/\hbar c}$$

A potential like this is called a Yukawa potential.

Even more important than the weak force is the nuclear force that keeps the nucleons together. The nuclear force has a considerably larger range of up to a couple of femtometer. It is in final analysis a consequence of the color force between the quarks that make up the nucleons. However, deriving the nuclear force directly in terms of the color force would be nontrivial. For one, the color force is mediated by massless gluons. Therefore the color force has an infinite range like the electromagnetic one, not a finite one like the nuclear force.

Worse, while the electromagnetic force becomes at least negligibly small at large distances, the color force does not. As a result, you are not going to find isolated quarks like you can find isolated electrons or isolated protons. Quarks are always bound together. They clump together in quark-antiquark pairs called “mesons,” or in quark or antiquark triplets called “baryons,” including the proton and neutron. Since quarks have spin $\frac{1}{2}$, mesons are bosons, while baryons are fermions. If you keep pulling the two quarks of a meson apart, you will not end up with two separate quarks. Instead you will eventually put in enough energy to create a new quark-antiquark pair in between the two, and that spoils the quark separation that you thought you had achieved.

Still, because of the fact that quarks clump together, it is possible to describe the interaction between nucleons in terms of the exchange of mesons. In particular, the primary interaction between nucleons is believed to be due to the exchange of pi-mesons, or pions for short. There are three of them, the charged π^+ and π^- pions and the uncharged π^0 . The π^\pm have a mass of about 140 MeV and the π^0 135 MeV. That makes the typical range of the force $\hbar c/m_\pi c^2$ about 1.4 fm, which fits the known range of the nuclear force well.

The potential that is produced by the exchange of single pions is also found to depend on the nucleon spins. The complete potential is called the “one-pion exchange potential,” or “OPEP” for short:

$$V = 4g_\pi^2 \left(\frac{m_\pi}{m_p} \right)^2 m_\pi c^2 \hat{T}_1 \cdot \hat{T}_2 \left[\frac{(\hat{\vec{S}}_1 \cdot \vec{r})(\hat{\vec{S}}_2 \cdot \vec{r})}{\hbar^2 r^2} \left(1 + 3\frac{d}{r} + 3\frac{d^2}{r^2} \right) - \frac{\hat{\vec{S}}_1 \cdot \hat{\vec{S}}_2}{\hbar^2} \left(\frac{d}{r} + \frac{d^2}{r^2} \right) \right] \frac{e^{-r/d}}{r/d} \quad (11.10)$$

where g_π^2 is a nondimensional constant whose empirical value is about 15, m_p the nucleon mass, $d = \hbar c/m_\pi c^2$ is the typical range of the force, and \vec{r} is the

connecting vector from nucleon 1 to nucleon 2. The dot product $\hat{\vec{T}}_1 \cdot \hat{\vec{T}}_2$ involves the so-called “isospin” of the nucleons, which will be discussed in section 11.18. For now, it should suffice that the dot product is a nondimensional number that, along with the spin state, is related to the symmetry of the spatial state of the nucleons through the antisymmetrization requirement.

The strong spin dependence of nuclear forces evident in the OPEP is reflected in the fact that the triplet spin state of the deuteron is bound, but the singlet state is not. Furthermore, the OPEP does not just depend on the length r of the connecting vector between nucleons \vec{r} , but also on the vector \vec{r} itself on account of the dot products with nucleon spins. Now classically, the force is found by differentiating the potential with respect to \vec{r} . If you do that with a potential that only depends on the distance between the particles r , you get a force that is along the connecting line between the particles. Such a force does not produce a net moment, so orbital angular momentum is preserved. However, a potential that depends explicitly on \vec{r} , rather than just r , produces a force that is not aligned with the connecting line. Such a force produces a moment, and orbital angular momentum is then no longer preserved. In quantum mechanics, the Hamiltonian no longer commutes with the operator of orbital angular momentum. Therefore orbital angular momentum becomes uncertain in energy eigenstates. That is seen in the deuteron, in which in the ground state the orbital angular momentum has about a 95% probability of being zero and a 5% probability of having azimuthal quantum number $l = 2$.

One effect that is well explained by the meson exchange mechanism is the charge symmetry of the nuclear force. In particular, the nuclear force between two protons is the same as the one between two neutrons. (The Coulomb repulsion is obviously not the same, and must be subtracted first to isolate the nuclear force effect.) The explanation of charge symmetry is that in both cases, the force is due to the exchange of the same uncharged mesons. Charged mesons cannot be exchanged between two protons because one of the two protons would end up with two units of net charge. Similarly, charged mesons cannot be exchanged between neutrons, since one of the two neutrons would end up with a negative charge. Only protons with charge e and uncharged neutrons are acceptable as products of meson exchange in normal nuclei.

The fact that charge independence is not perfect is also well explained. Perfect charge independence would imply that the nuclear force between a neutron and a proton is the same as that between two protons or two neutrons. However, experimental data show about a percent difference. That can be understood from the fact that a proton and a neutron can exchange charged mesons in addition to uncharged ones. The proton can emit a meson with one unit of positive charge that the neutron then absorbs. Or the neutron can emit a meson with one unit of negative charge that the proton then absorbs. Either

process turns the proton into a neutron and vice versa. The charged pions are a few percent heavier than the uncharged ones, producing a slight difference in exchange force.

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 processes involving the correlated exchange of multiple pions become important. The interactions between nucleons at normal nuclear separations mainly involve two-pion exchanges. That can be understood from the fact that a two-pion composite has about twice the mass of a single pion, hence a typical range of 0.7 fm. The strong repulsion between nucleons when they get very close involves exchanges of three pions and more.

Multi-pion exchanges are much more difficult to analyze than one-pion ones. Therefore physicists often model these 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. The procedure can be justified to some extent by the fact that pions really do combine into correlated states called “resonances.” A two-pion correlated state called the ρ is experimentally detected at about 770 MeV. (Note that a true bound state of two 140 MeV pions should have a mass less than 280 MeV.) Three-pion resonances called ω and ϕ are also observed; the two are almost the same but differ somewhat in mass, 780 respectively 1020 MeV. If you think positively, there may also be a two-pion resonance at very roughly 700 MeV called σ or ε . And in addition, the interaction between nucleons also involves the exchange of a second true particle, the η meson, with a mass of 549 MeV.

All these particles are bosons, but they differ in spin and parity, and that produces Yukawa-type potentials of different forms. The pion and similarly the η are 0^- “pseudo-scalar” particles. However, the 0^+ σ is a “scalar” particle, and the 1^- ρ , ω , and ϕ are “vector” particles. These two kinds of particles add spin-orbit, $\vec{L} \cdot \vec{S}$, dependence to the potential, as well as momentum dependence. The vector particles also contribute much of the strong repulsion between nucleons when they get too close. Since the ω and ϕ resonances are very similar, often just a suitably defined ω resonance is used in OBEP potentials.

Unfortunately, for the critical range of normal nucleon separation distances, it is found that two different σ resonances are required to produce a decent potential. There is no experimental evidence for their existence. Therefore alternate approaches may be used. One approach is to crunch out the full “two-pion exchange potential,” or “TPEP.” However, that is a very nontrivial exercise. There are different ways in which two pions can be exchanged, and in addition you need to account accurately for the experimentally observed σ and ρ resonances between the two. Also, in a two pion exchange, either nucleon may be in an excited state during part of the process. Then there is the finite size of

the nucleons and mesons to worry about, especially at close nucleon spacings. A simpler approach is to formulate a suitable phenomenological potential with the right symmetries. The unknown parameters in such a potential can be fitted to experimental data.

All of that is far outside the scope of this book. Only the most elementary nuclear models will be used in the remainder of this chapter.

11.10 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.

11.10.1 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:

$$R \approx R_A \sqrt[3]{A} \quad R_A = 1.23 \text{ fm} \quad (11.11)$$

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, [21, pp. 44-59]

11.10.2 von Weizsäcker formula

The binding energy of nuclei can be approximated by the “von Weizsäcker formula,” or “Bethe-von Weizsäcker formula:”

$$E_{\text{B,vW}} = C_v A - C_s A^{2/3} - C_c \frac{Z(Z - C_z)}{A^{1/3}} - C_d \frac{(Z - N)^2}{A} - C_p \frac{o_Z + o_N - 1}{A^{C_e}} \quad (11.12)$$

where the C_i 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 [24] 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 $1.60218 \cdot 10^{-13}$ 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:”

$$m_{\text{nucleus,SE}} = Zm_p + Nm_n - \frac{E_{\text{B,vW}}}{c^2} \quad (11.13)$$

11.10.3 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.

11.10.4 Accuracy of the formula

Figure 11.7 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

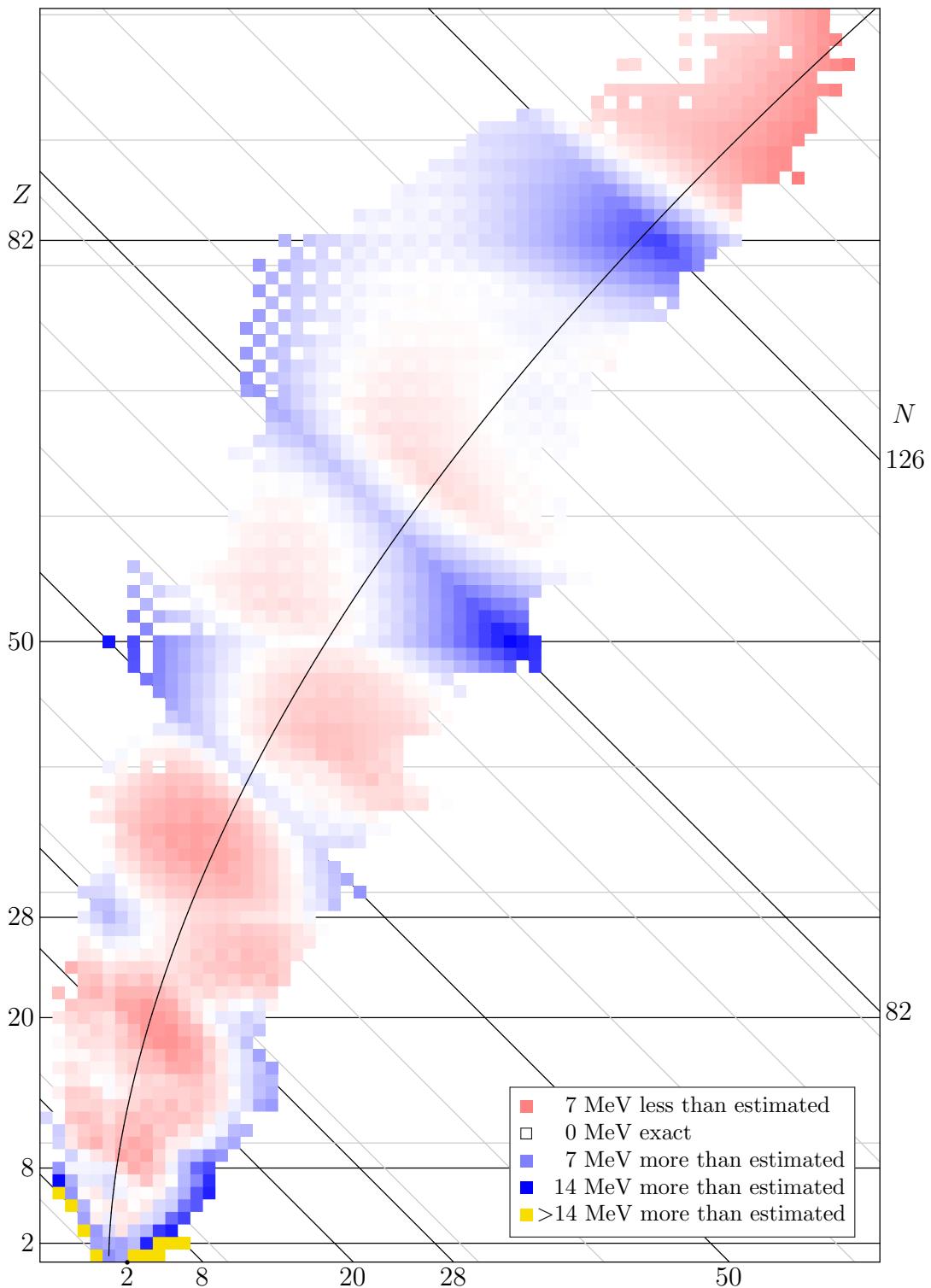


Figure 11.7: Error in the von Weizsäcker formula.

the magic number, further nucleons have a much lower energy level, and the deviation from the von Weizsäcker value decreases again.

11.11 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.

11.11.1 Decay mechanism

Figure 11.8 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.

Not shown in figure 11.8 is the unstable beryllium isotope ^8_4Be , which has a half-life of only 67 as, (i.e. $67 \cdot 10^{-18}$ 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 11.8, 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 11.8 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 6.8.2.

Figure 11.9 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,

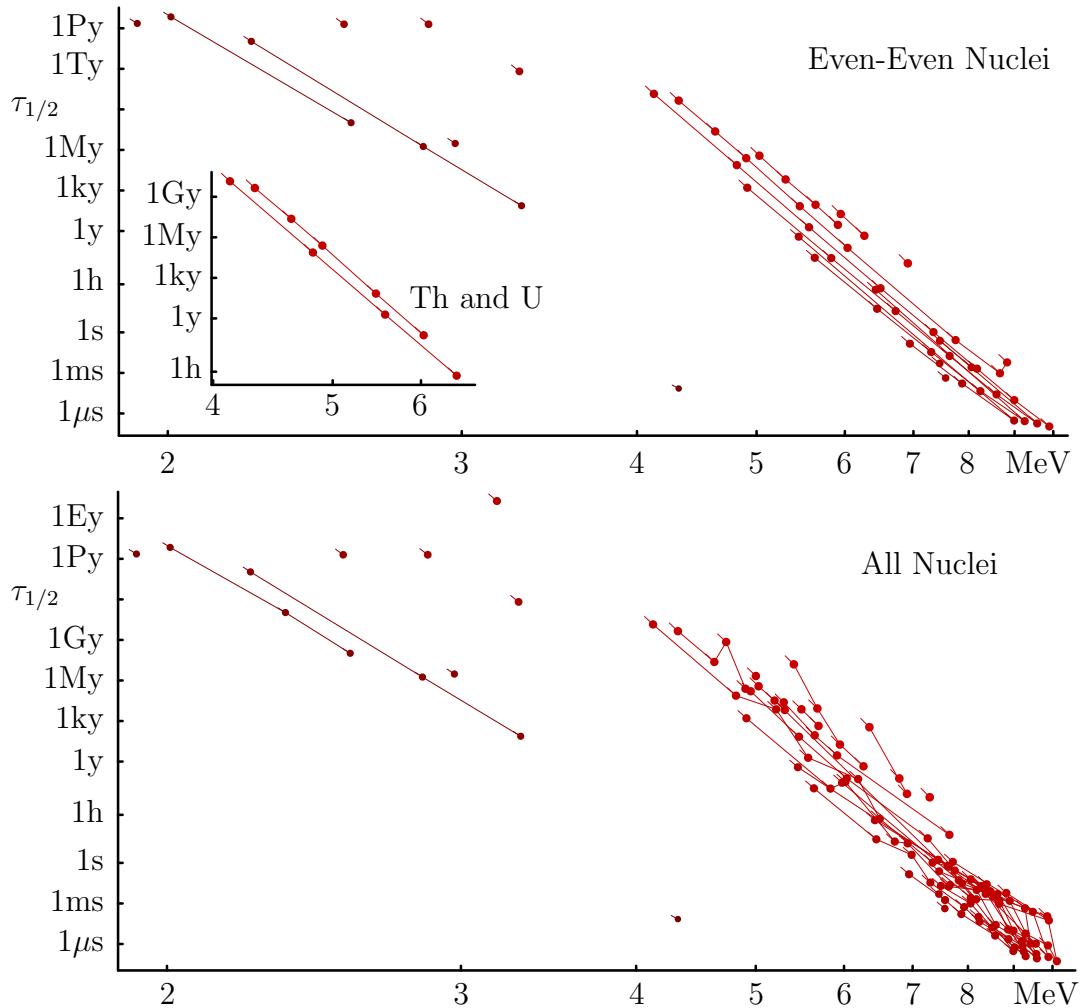


Figure 11.8: Half-life versus energy release for the atomic nuclei marked in NUBASE 2003 as showing pure alpha decay with unqualified energies. Top: only the even values of the mass and atomic numbers cherry-picked. Inset: really cherry-picking, only a few even mass numbers for thorium and uranium! Bottom: all the nuclei except one.

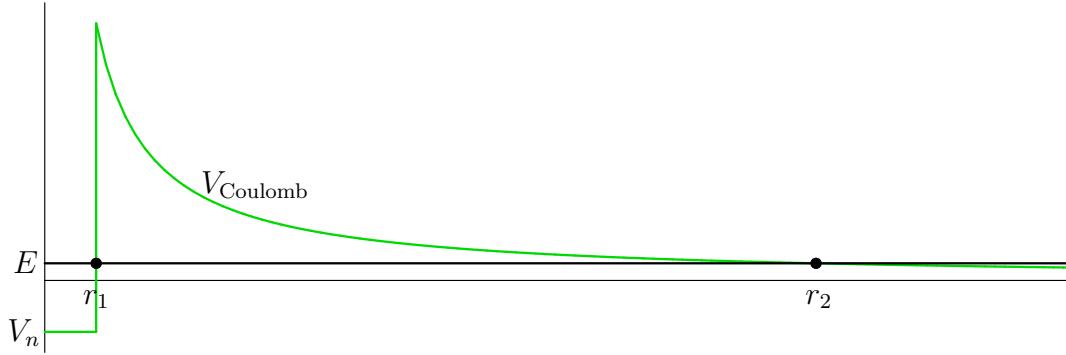


Figure 11.9: Schematic potential for an alpha particle that tunnels out of a nucleus.

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, so it is not used.

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 6.9 as

$$T \approx e^{-2\gamma_{12}} \quad \gamma_{12} = \frac{1}{\hbar} \int_{r_1}^{r_2} \sqrt{2m_\alpha(V - E)} dr \quad (11.14)$$

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.

11.11.2 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 11.6. 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. (An 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 indicated 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 11.9. It can be ballparked reasonably well from the number of nucleons A ; according to section 11.10,

$$R \approx R_A \sqrt[3]{A} \quad R_A = 1.23 \text{ fm}$$

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, [21, 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 minuscule $e^{-2\gamma_{12}}$ 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 ;

$$\frac{v_\alpha}{2r_0} e^{-2\gamma_{12}} = \frac{1}{\tau} \quad (11.15)$$

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 11.8. Forget about engineering ideas about acceptable accuracy. A 50% error in half-life would be invisible seen on the tremendous range of figure 11.8. 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 [21, p. 252].

That leaves the value of γ_{12} to be found from the integral over the barrier in (11.14). 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 ;

$$E = \frac{(Z - Z_\alpha)e Z_\alpha e}{4\pi\epsilon_0 r_2} \quad Z_\alpha = 2$$

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 11.8, 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 11.8. 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 11.10 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 $^{209}_{83}\text{Bi}$ bismuth isotope indicated by the rightmost dot in figure 11.10. Its true half-life of 19 Ey, $19 \cdot 10^{18}$ years, is grossly underestimated to be just 9 Py, $9 \cdot 10^{15}$ years. Then again, since the universe has only existed about $14 \cdot 10^9$ years, who is going to live long 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 ^8Be , which is not shown in the figure, the predicted half-life is 55 as ($55 \cdot 10^{-18}$ s), versus a true value of 67 as.

11.11.3 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 2000 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

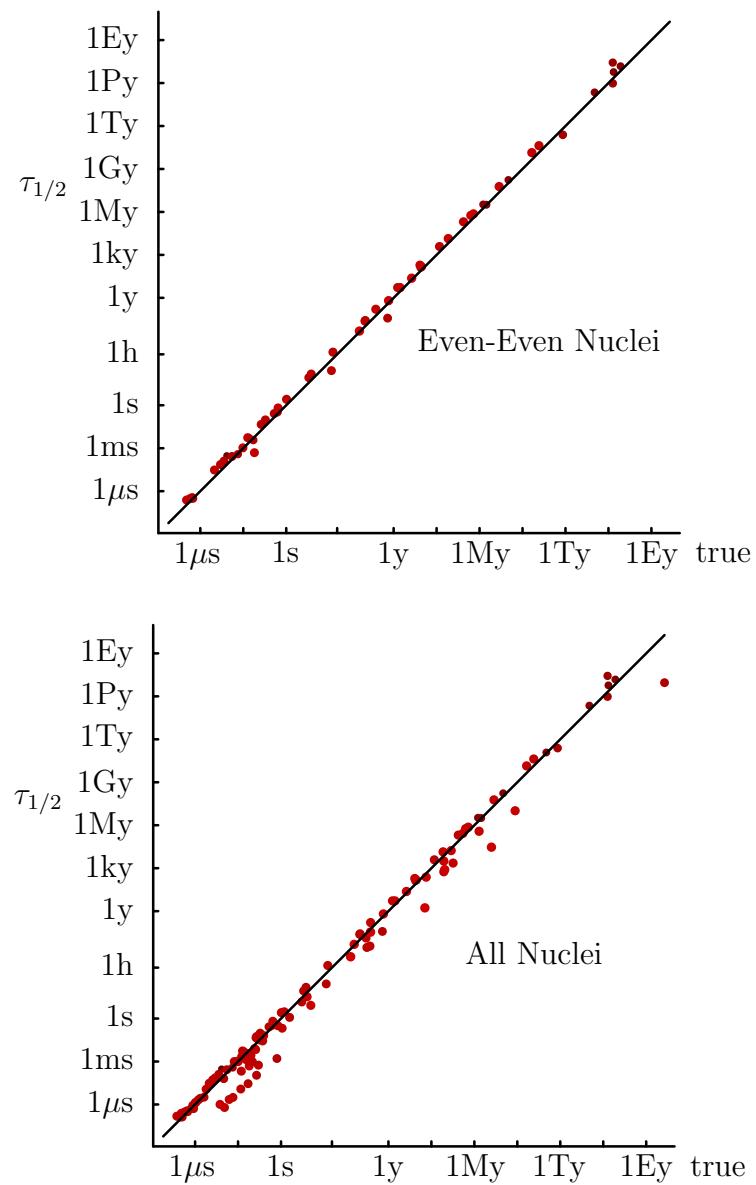


Figure 11.10: Half-life predicted by the Gamow / Gurney & Condon theory versus the true value. Top: even-even nuclei only. Bottom: all the nuclei except one.

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 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 11.10 how many nuclei that are not even-even do indeed have half-lives that are orders of magnitude larger than predicted by theory. Consider the example of bismuth-209, with a half-life 2000 times longer than predicted. Bismuth-209 has a spin, i.e. an azimuthal quantum number, of $9/2$. However, the decay product thallium-205 has spin $1/2$ in its ground state. If you check out the excited states of thallium-205, there is an excited state with spin $9/2$, 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 excited state is not possible either, because it violates conservation of parity, chapter 6.2 and 6.3.7. 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 3.1.3. 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 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 $|\frac{9}{2} - l| \leq j \leq \frac{9}{2} + l$ 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 positive parity and 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 3.2.2. 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-lives of figure 11.10. Indeed, while the predicted half-lives of many nuclei are way below the true value in the

figure, few are significantly above it.

11.11.4 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 11.2 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-lives of the events as shown in table 11.2.

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, 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.

$^{238}_{92}\text{U}$ with $\tau_{1/2} = 1.4 \cdot 10^{17}$ s							
Ejected:	^2_1H	^3_1H	^3_2He	^4_2He	^8_4Be	$^{16}_6\text{C}$	$^{20}_8\text{O}$
E , MeV:	-11.2	-10.0	-11.8	4.3	7.9	17.4	35.3
E/A_1 :	-5.6	-3.3	-3.9	1.1	1.0	1.1	1.8
r_1/r_2 :	-	-	-	0.14	0.14	0.21	0.33
γ_{12}^* :	-	-	-	85	172	237	239
γ_{12} :	-	-	-	45	92	103	74
$\tau_{1/2}$, s:	∞	∞	∞	$4 \cdot 10^{17}$	$9 \cdot 10^{58}$	$2 \cdot 10^{68}$	$8 \cdot 10^{44}$
$^{223}_{88}\text{Ra}$ with $\tau_{1/2} = 9.9 \cdot 10^5$ s							
Ejected:	^2_1H	^3_1H	^3_2He	^4_2He	^8_4Be	$^{14}_6\text{C}$	$^{18}_8\text{O}$
E , MeV:	-9.2	-9.2	-8.3	6.0	12.9	31.9	40.4
E/A_1 :	-4.6	-3.1	-2.8	1.5	1.6	2.3	2.2
r_1/r_2 :	-	-	-	0.20	0.23	0.40	0.39
γ_{12}^* :	-	-	-	69	129	156	203
γ_{12} :	-	-	-	32	53	40	53
$\tau_{1/2}$, s:	∞	∞	∞	$9 \cdot 10^4$	$4 \cdot 10^{24}$	$9 \cdot 10^{12}$	$2 \cdot 10^{24}$
$^{256}_{100}\text{Fm}$ with $\tau_{1/2} = 9.5 \cdot 10^3$ s							
Ejected:	^2_1H	^3_1H	^3_2He	^4_2He	^8_4Be	$^{14}_6\text{C}$	$^{20}_8\text{O}$
E , MeV:	-9.6	-8.5	-8.7	7.1	13.2	27.9	39.4
E/A_1 :	-4.9	-2.8	-2.9	1.8	1.6	2.0	2.0
r_1/r_2 :	-	-	-	0.22	0.22	0.31	0.35
γ_{12}^* :	-	-	-	72	145	192	249
γ_{12} :	-	-	-	31	63	63	74
$\tau_{1/2}$, s:	∞	∞	∞	$2 \cdot 10^5$	$8 \cdot 10^{32}$	$6 \cdot 10^{32}$	$6 \cdot 10^{42}$
							$2 \cdot 10^{31}$

Table 11.2: Candidates for nuclei ejected by uranium-238, radium-223, and fermium-256.

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-lives in table 11.2 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 11.2 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 11.2 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 11.2 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.

11.12 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

$$\text{magic numbers: } 2, 8, 20, 28, 50, 82, 126, \dots \quad (11.16)$$

The higher magic number values are quite clearly seen in proton pair and neutron pair removal graphs like figures 11.5 and 11.6 in section 11.8.

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_2He helium-4 nucleus, with 2 protons and 2 neutrons, is a good example. It has more than three times the binding energy of ^3_2He 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 rejected within a few centuries.

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 4.9 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.

11.12.1 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 11.11a.

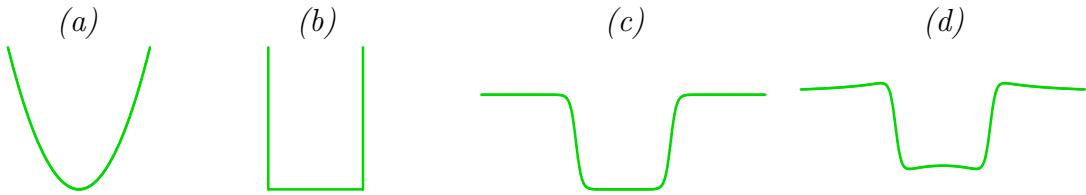


Figure 11.11: Example average nuclear potentials: (a) harmonic oscillator, (b) impenetrable surface, (c) Woods-Saxon, (d) Woods-Saxon for protons.

The energy eigenvalues of the harmonic oscillator are

$$E_n = \left(n + \frac{1}{2}\right) \hbar\omega \quad n = 1, 2, 3, \dots \quad (11.17)$$

Also, in spherical coordinates the energy eigenfunctions of the harmonic oscillator can be taken to be of the form, {A.107},

$$\begin{aligned} \psi_{nlmm_s}^{\text{ho}} &= R_{nl}^{\text{ho}}(r) Y_l^m(\theta, \phi) \uparrow & l &= n - 1, n - 3, \dots \geq 0 \\ m &= -l, -l + 1, \dots, l - 1, l & m_s &= \pm \frac{1}{2} \end{aligned} \quad (11.18)$$

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_{nl}^{ho} 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 2.6, 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 3.2, 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:

$$s, p, d, f, g, h, i, [j], k, \dots \implies l = 0, 1, 2, 3, 4, 5, 6, 7, 8, \dots$$

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 rennotated as

$$\psi_{21mm_s} \implies 2p$$

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 11.12. 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 actuality the fourth 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

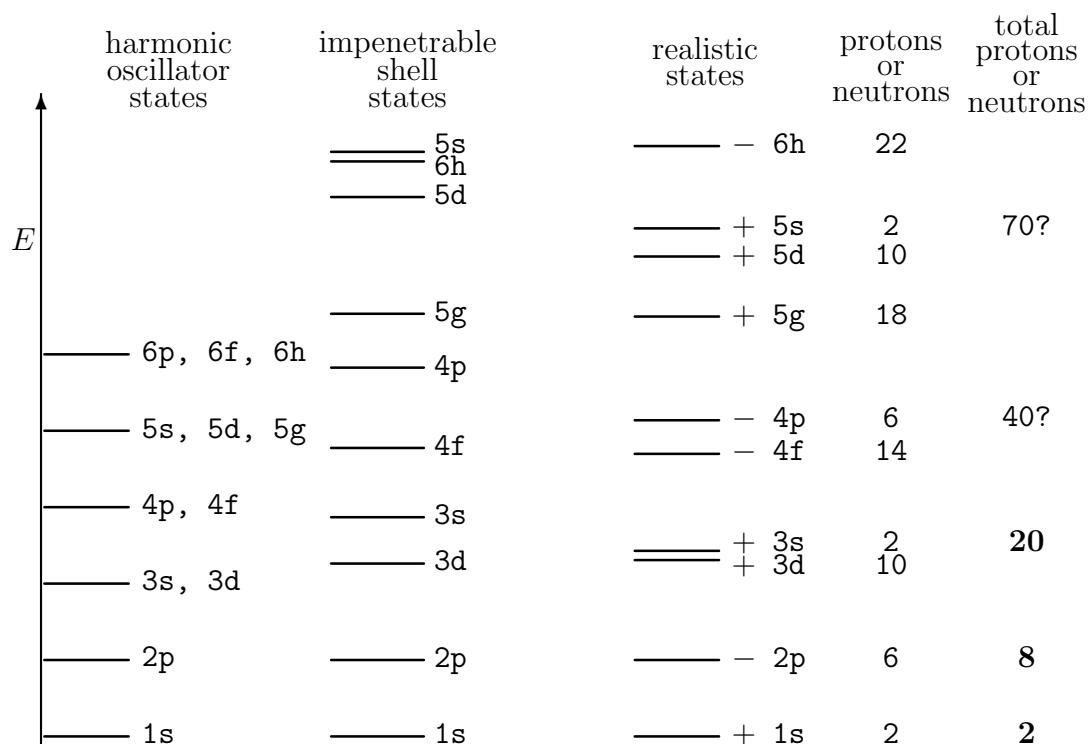


Figure 11.12: Nuclear energy levels for various assumptions about the average nuclear potential. The signs indicate the parity of the states.

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 gives the “impenetrable-shell” potential shown in figure 11.11. It too is analytically solvable, {A.108}. The energy levels are shown in figure 11.12. Unfortunately, it does not help any explaining the fourth magic number 28.

It does help understand why the shell model works at all, [[14]]. 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 11.11c. 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 11.11d. 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 11.12. 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 11.12 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 \vec{r} in it is everywhere replaced by $-\vec{r}$. The parity is “minus one,” or “negative,” or “odd,” if the wave function merely changes sign when \vec{r} is replaced by $-\vec{r}$. 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, {A.15} or {A.107}. 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 11.12. (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.

11.12.2 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 non quantum 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

$$\Delta H \propto -\hat{\vec{L}} \cdot \hat{\vec{S}}$$

Here $\hat{\vec{L}}$ is the orbital angular momentum vector and $\hat{\vec{S}}$ 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, section 12.1.6. 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 $\hat{\vec{L}} \cdot \hat{\vec{S}}$ involves \hat{L}_x and \hat{L}_y , and these do not commute with \hat{L}_z regardless of how you orient the axis system. And the same for \hat{S}_x and \hat{S}_y .

With spin-orbit interaction, energy eigenfunctions of nonzero orbital angular momentum no longer have precise orbital momentum L_z in a chosen z -direction. And neither do they have precise spin S_z in such a direction.

Therefore the energy eigenfunctions can no longer be taken to be of the form $R_{nl}(r)Y_l^m(\theta, \phi)\uparrow$. They have uncertainty in both m and m_s , so they will be combinations of states $R_{nl}(r)Y_l^m(\theta, \phi)\uparrow$ with varying values of m and m_s .

However, consider the *net* angular momentum operator

$$\hat{J} \equiv \hat{L} + \hat{S}$$

If you expand its square magnitude,

$$\hat{J}^2 = (\hat{L} + \hat{S}) \cdot (\hat{L} + \hat{S}) = \hat{L}^2 + 2\hat{L} \cdot \hat{S} + \hat{S}^2$$

you see that the spin-orbit term can be written in terms of the square magnitudes of orbital, spin, and net angular momentum operators:

$$-\hat{L} \cdot \hat{S} = -\frac{1}{2} [\hat{J}^2 - \hat{L}^2 - \hat{S}^2]$$

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 10.1.6. 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 $|L - S| \leq J \leq L + S$, 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 $|l - s| \leq j \leq l + s$ 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 precise net angular momentum j can be written as linear combinations of the fourteen $R_{nl}Y_l^m\downarrow$ states. (Figure 10.5 shows such combinations up to $l = 2$; item 2 in chapter 10.1.7 gives a general formula.) Pictorially,

$$7 \text{ } 4f\uparrow \text{ and } 7 \text{ } 4f\downarrow \text{ states} \quad \Rightarrow \quad 6 \text{ } 4f_{5/2} \text{ and } 8 \text{ } 4f_{7/2} \text{ states}$$

where the spectroscopic convention is to show the net angular momentum j as a subscript for states in which its value is precise.

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 precise square orbital and square net angular momentum,

$$-\hat{\vec{L}} \cdot \hat{\vec{S}} = -\frac{1}{2}\hbar^2[j(j+1) - l(l+1) - s(s+1)] = \begin{cases} \frac{1}{2}(l+1)\hbar^2 & \text{for } j = l - \frac{1}{2} \\ -\frac{1}{2}l\hbar^2 & \text{for } j = l + \frac{1}{2} \end{cases}$$

The eight $4p_{7/2}$ states of lowered energy form the shell that is filled at the fourth magic number 28.

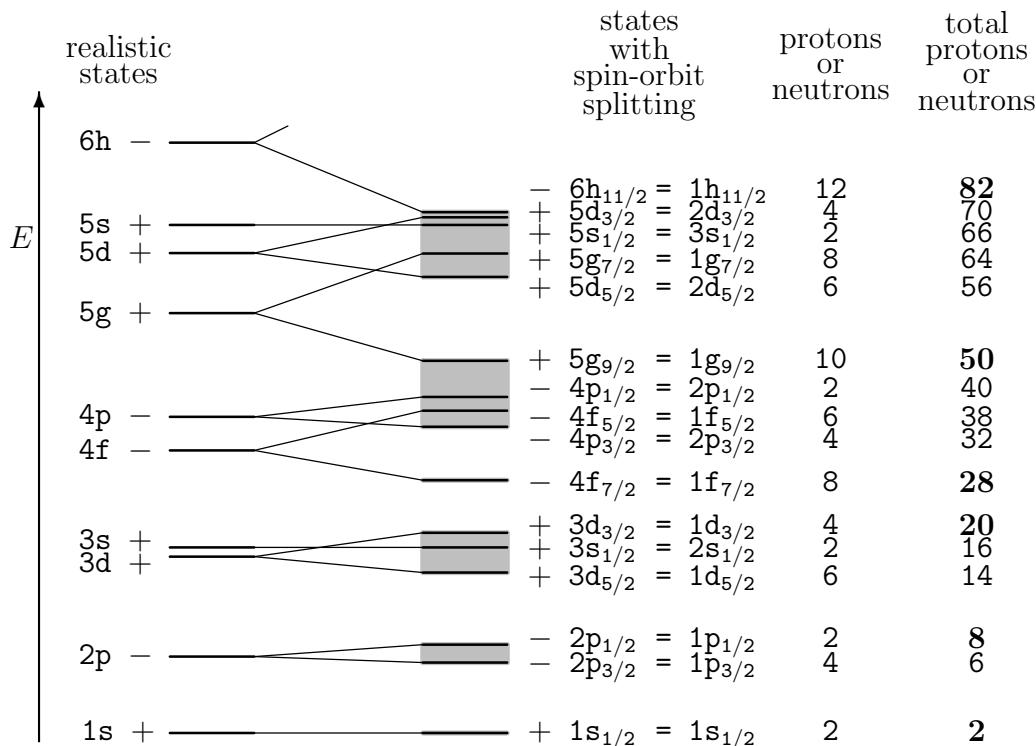


Figure 11.13: Schematic effect of spin-orbit interaction on the energy levels. The ordering within bands is realistic for neutrons. The designation behind the equals sign is the “official” one. (Assuming counting starts at 1).

Figure 11.13 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 11.13 also shows the “official” enumeration of the states. You be the judge which numbering system makes the most sense to you.

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:

$$\begin{aligned} & 1s_{1/2} \\ & 2p_{3/2} \quad 2p_{1/2} \\ & 3d_{5/2} \quad 3s_{1/2} \quad 3d_{3/2} \\ & \quad 4f_{7/2} \\ & 4p_{3/2} \quad 4f_{5/2} \quad 4p_{1/2} \quad 5g_{9/2} \\ & 5g_{7/2} \quad 5d_{5/2} \quad 6h_{11/2} \quad 5d_{3/2} \quad 5s_{1/2} \\ & 6h_{9/2} \quad 6f_{7/2} \quad 6f_{5/2} \quad 6p_{3/2} \quad 6p_{1/2} \quad 7i_{13/2} \end{aligned}$$

neutrons:

$$\begin{aligned} & 1s_{1/2} \\ & 2p_{3/2} \quad 2p_{1/2} \\ & 3d_{5/2} \quad 3s_{1/2} \quad 3d_{3/2} \\ & \quad 4f_{7/2} \\ & 4p_{3/2} \quad 4f_{5/2} \quad 4p_{1/2} \quad 5g_{9/2} \\ & 5d_{5/2} \quad 5g_{7/2} \quad 5s_{1/2} \quad 5d_{3/2} \quad 6h_{11/2} \\ & 6f_{7/2} \quad 6h_{9/2} \quad 6p_{3/2} \quad 6f_{5/2} \quad 7i_{13/2} \quad 6p_{1/2} \\ & 7g_{9/2} \quad 7d_{5/2} \quad 7i_{11/2} \quad 7g_{7/2} \quad 7s_{1/2} \quad 7d_{3/2} \quad 8j_{15/2} \end{aligned}$$

The ordering for protons follows [23, table 7-1], but not [23, p. 223], to Z=92, and then [21], whose table seems to come from Mayer and Jensen. The ordering for neutrons follows [23], with the subshells beyond 136 taken from [[8]]. However, the $7i_{13/2}$ and $6p_{1/2}$ states were swapped since the shell filling [23, table 7-1] makes a lot more sense if you do. The same swap is also found in [27, p. 255], following Klinkenberg, while [21, p. 155] puts the $7i_{13/2}$ subshell even farther down below the $6p_{3/2}$ state.

11.12.3 Example occupation levels

The purpose of this section is to explore how the shell model works out for sample nuclei.

Figure 11.14 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

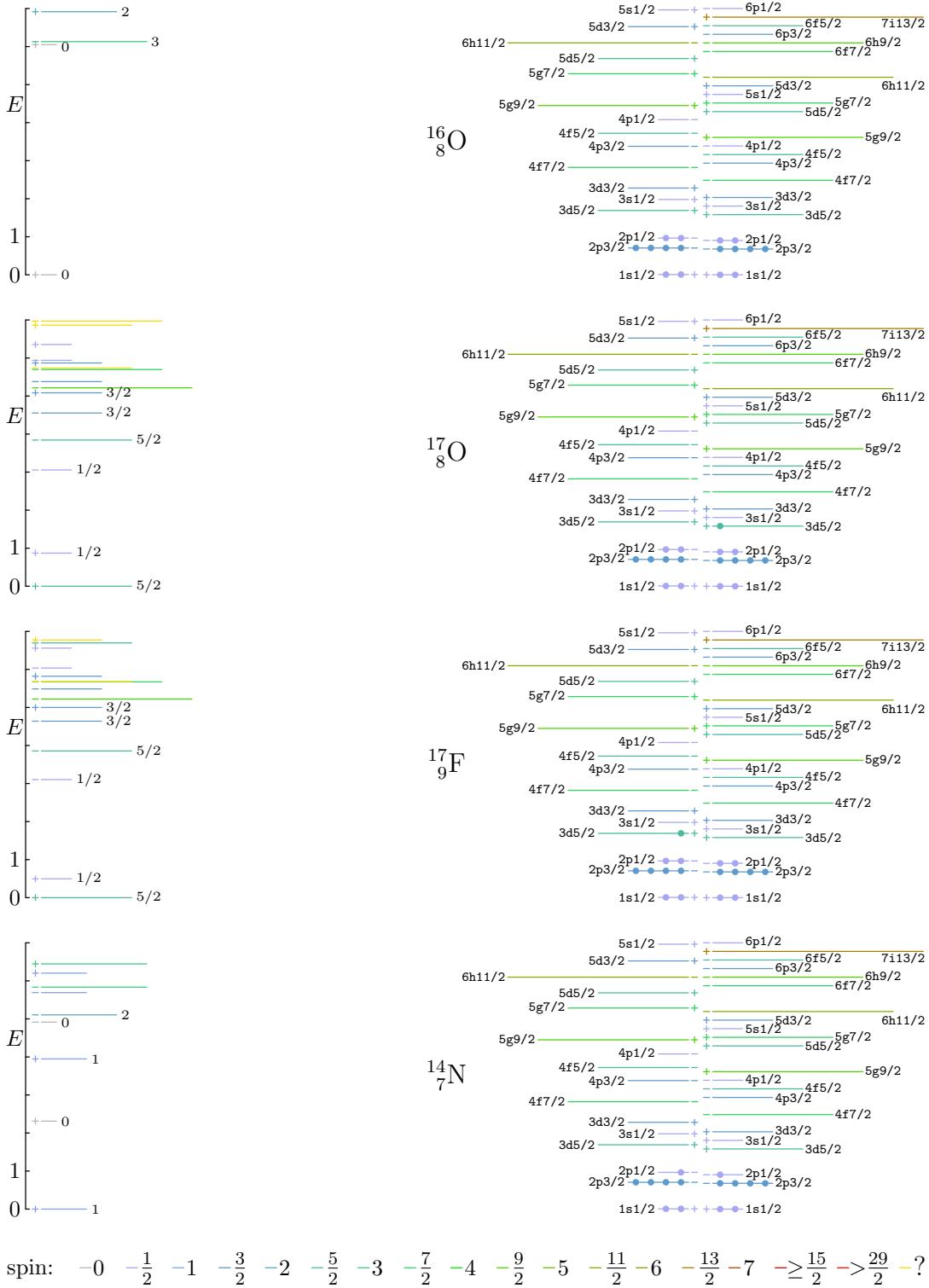


Figure 11.14: Energy levels for doubly-magic oxygen-16 and neighbors.

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 particularly stable. So it should require a great deal of energy to excite it. Indeed it does: figure 11.14 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 excited 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 11.13. 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 11.14. 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 $1/2^+$ 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 $3/2^+$ state, for kicking up the nucleon to the $3d_{3/2}$ state instead.

However, from the shell model, in particular figure 11.13, 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, {A.31}. 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 $3/2^+$ state. Some thoughts on these states will be given in the next subsection.

The fourth nucleus in figure 11.14 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 $1/2$. 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.

11.12.4 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 11.15, 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 10.1.7 item 3, or more simply from table 10.1, 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:

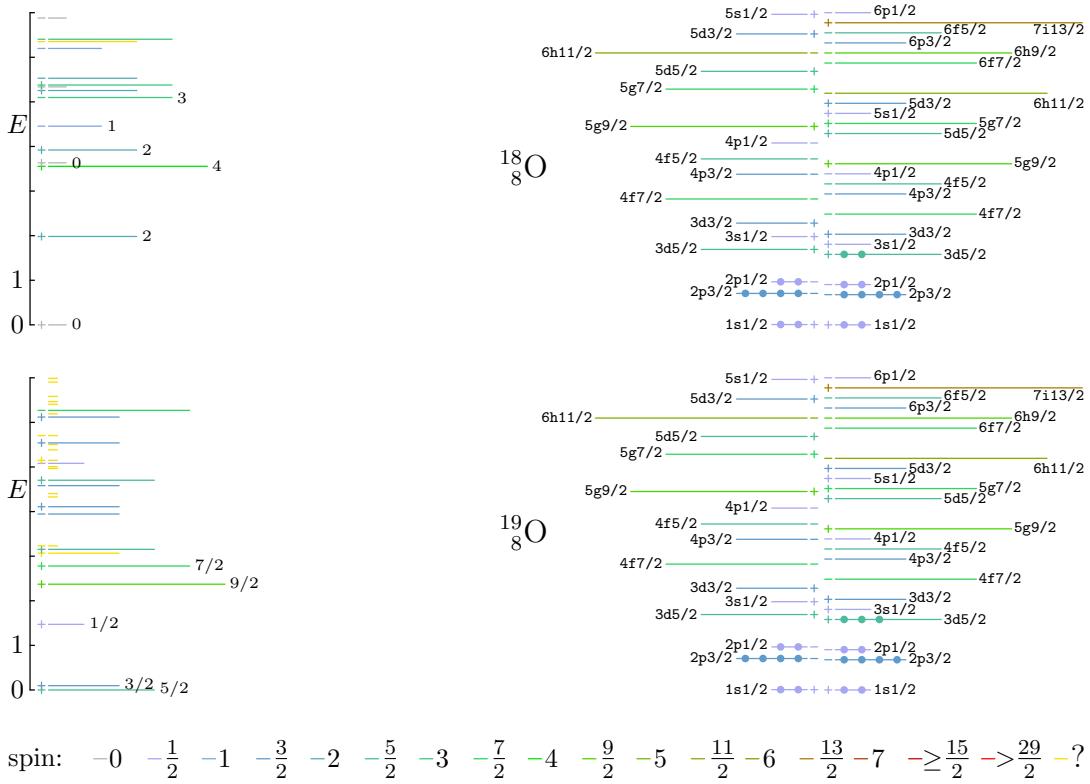


Figure 11.15: Nucleon pairing effect.

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 chapter 12.1. Here it must 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 11.15 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 11.13 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 11.14 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 $1/2^-$ 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 $5/2^-$ 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 $5/2$, [21, 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 $3/2^+$ 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 $5/2^-$ excitation is like the $3/2^-$, 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 11.15 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 10.1. And indeed, the oxygen-19 energy spectrum in figure 11.15 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 11.15 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 independent* 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 11.16. 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 10.1. 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

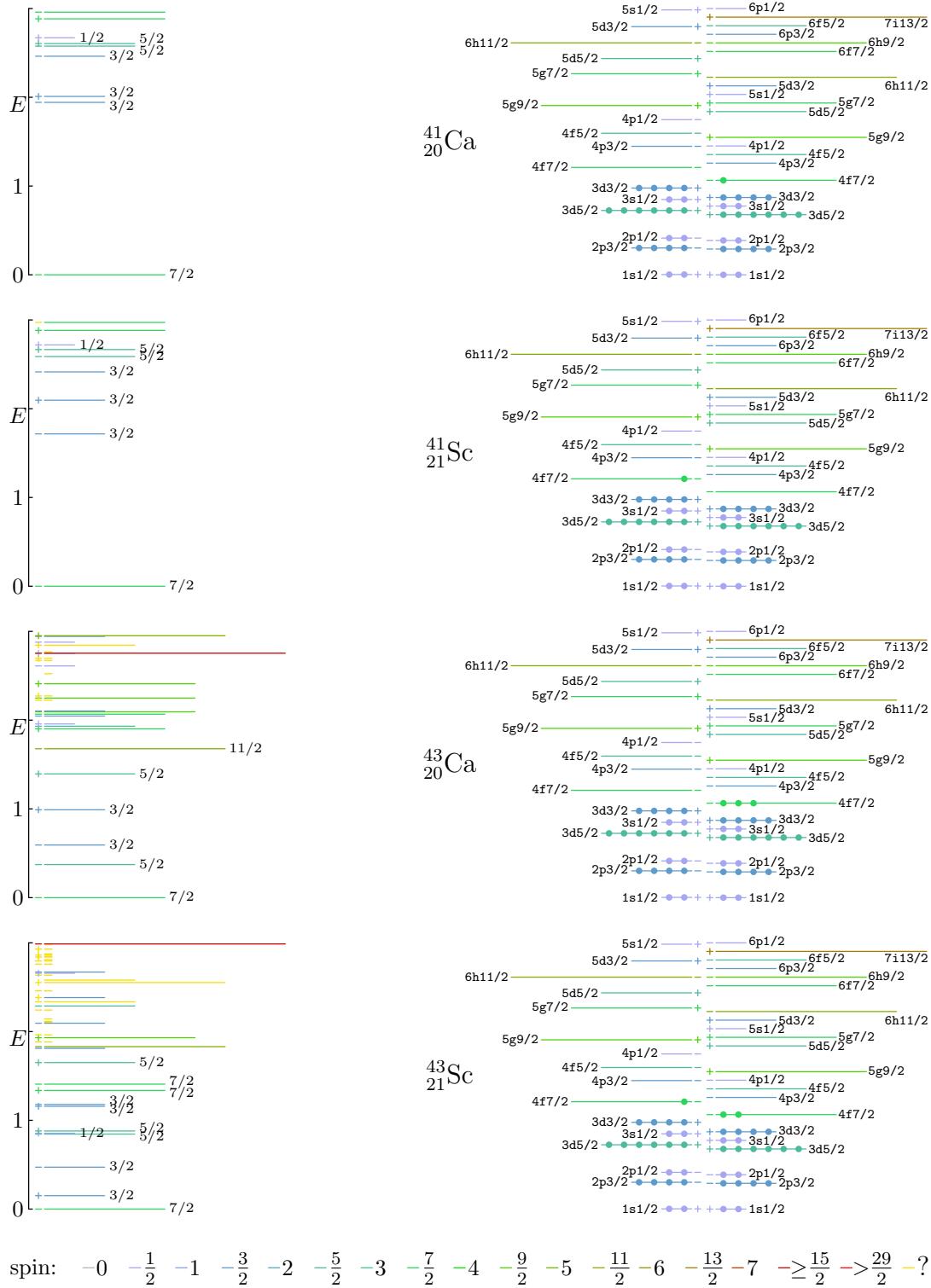


Figure 11.16: Energy levels for neighbors of doubly-magic calcium-40.

$9/2^-$ 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 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 spins 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.

11.12.5 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 11.14. 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.

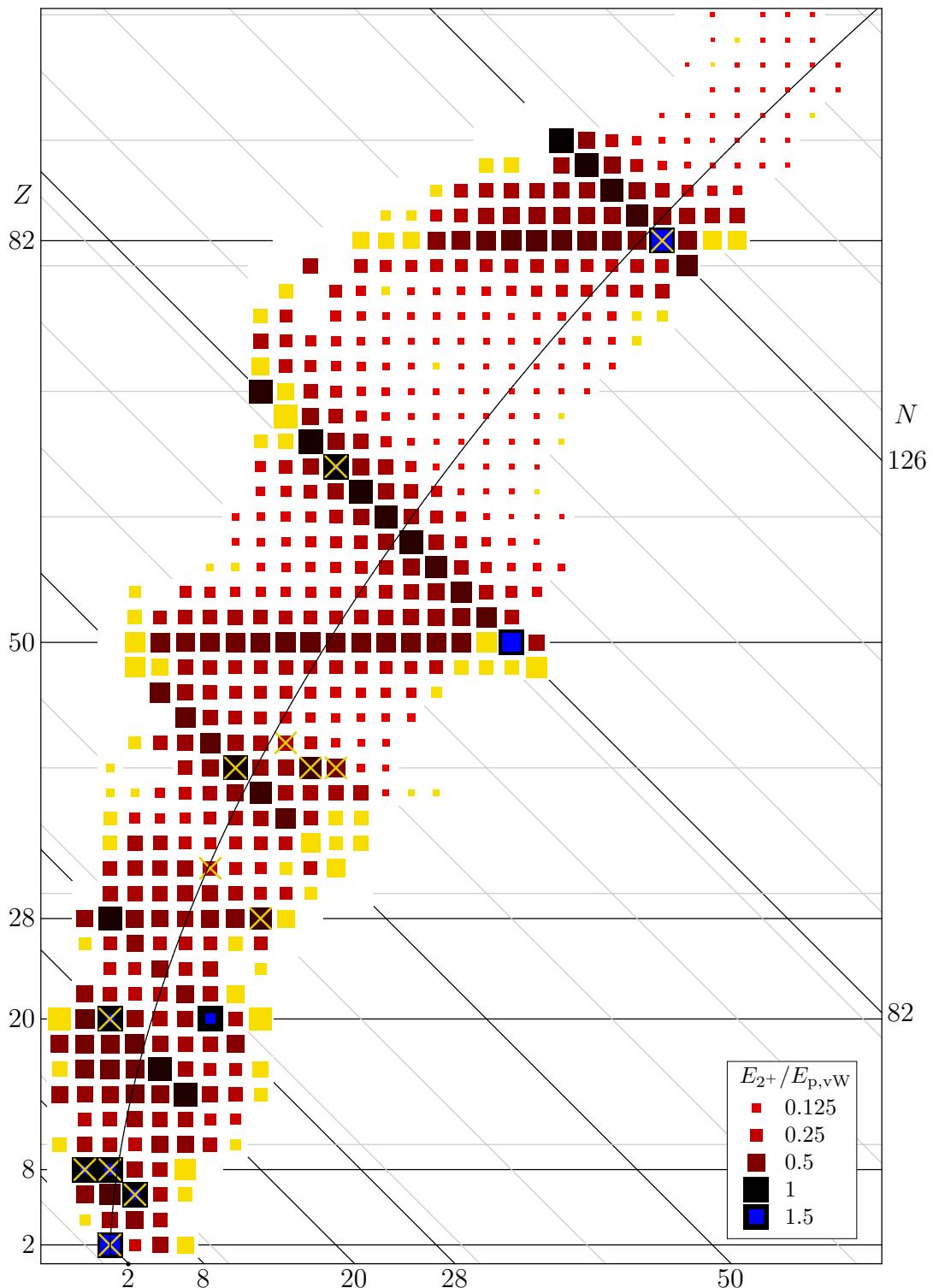
But now consider the possibility that *both* the protons and the neutrons would be in the $3d_{5/2}$ subshell. In that case, surely you would agree that these four, mutually attracting, nucleons in the same spacial orbits would significantly interact and lower their energy. Even if the unperturbed shell 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 4.3. 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 11.17. 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 already seems curious. Why would the lowest excited state

Figure 11.17: 2^+ excitation energy of even-even nuclei.

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.

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 11.17 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 pleased 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 11.18. 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 10.1. 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, [21, 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,

$$\psi_1, \psi_2, \psi_3, \dots, \psi_Q$$

Watch what happens when such states are mixed together. The energy follows

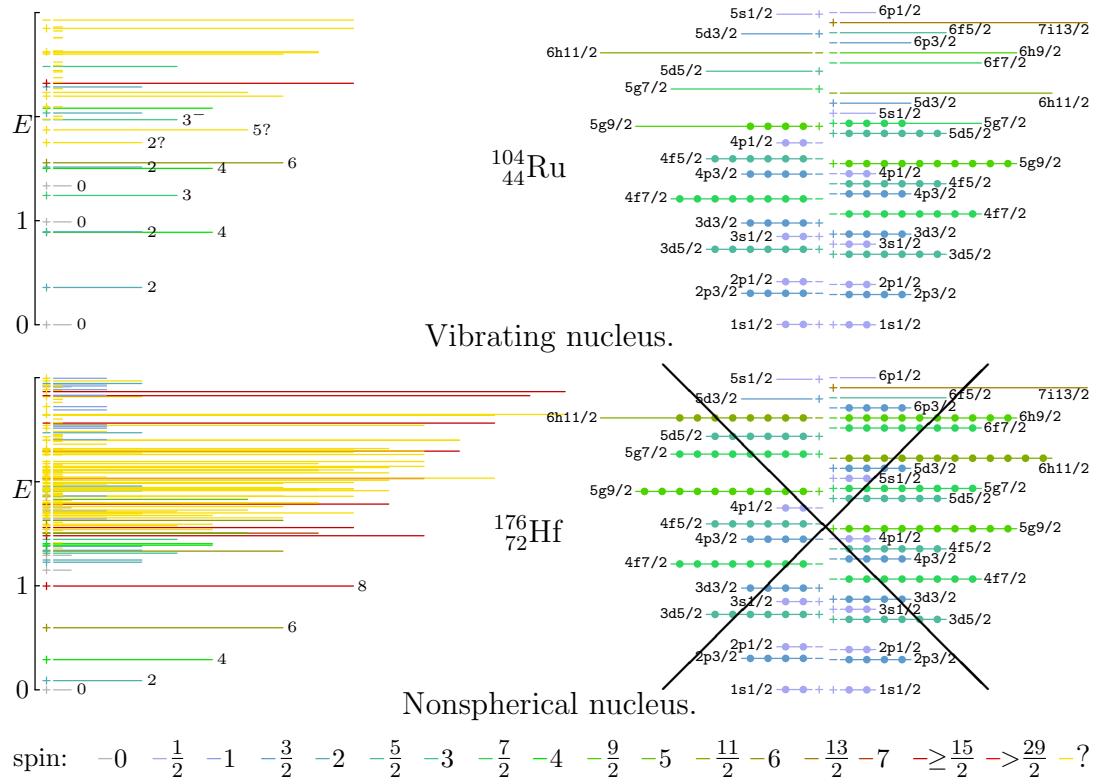


Figure 11.18: Collective motion effects.

from the Hamiltonian coefficients

$$\begin{aligned}
 E_1 &\equiv \langle \psi_1 | H \psi_1 \rangle & \varepsilon_{12} &\equiv \langle \psi_1 | H \psi_2 \rangle & \varepsilon_{13} &\equiv \langle \psi_1 | H \psi_3 \rangle & \dots & \varepsilon_{1Q} &\equiv \langle \psi_1 | H \psi_Q \rangle \\
 \varepsilon_{12}^* &\equiv \langle \psi_2 | H \psi_1 \rangle & E_2 &\equiv \langle \psi_2 | H \psi_1 \rangle & \varepsilon_{23} &\equiv \langle \psi_2 | H \psi_3 \rangle & \dots & \varepsilon_{2Q} &\equiv \langle \psi_2 | H \psi_Q \rangle \\
 \varepsilon_{13}^* &\equiv \langle \psi_3 | H \psi_1 \rangle & \varepsilon_{23}^* &\equiv \langle \psi_3 | H \psi_2 \rangle & E_3 &\equiv \langle \psi_3 | H \psi_3 \rangle & \dots & \varepsilon_{3Q} &\equiv \langle \psi_3 | H \psi_Q \rangle \\
 &\vdots &&\vdots &&\vdots &&\ddots &&\vdots \\
 \varepsilon_{1Q}^* &\equiv \langle \psi_Q | H \psi_1 \rangle & \varepsilon_{2Q}^* &\equiv \langle \psi_Q | H \psi_2 \rangle & \varepsilon_{3Q}^* &\equiv \langle \psi_Q | H \psi_3 \rangle & \dots & E_Q &\equiv \langle \psi_Q | H \psi_Q \rangle
 \end{aligned}$$

By assumption, the energy levels E_1, E_2, \dots of the states are all about the same, and if the unperturbed shell model was exact, the perturbations $\varepsilon_{..}$ would all be 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 4.3, 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 11.13.

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, chapter 12.1. 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 11.17 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 11.18.

11.12.6 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 11.19 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 [23, 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 11.15.

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 10.1 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 11.14. 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 11.13 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

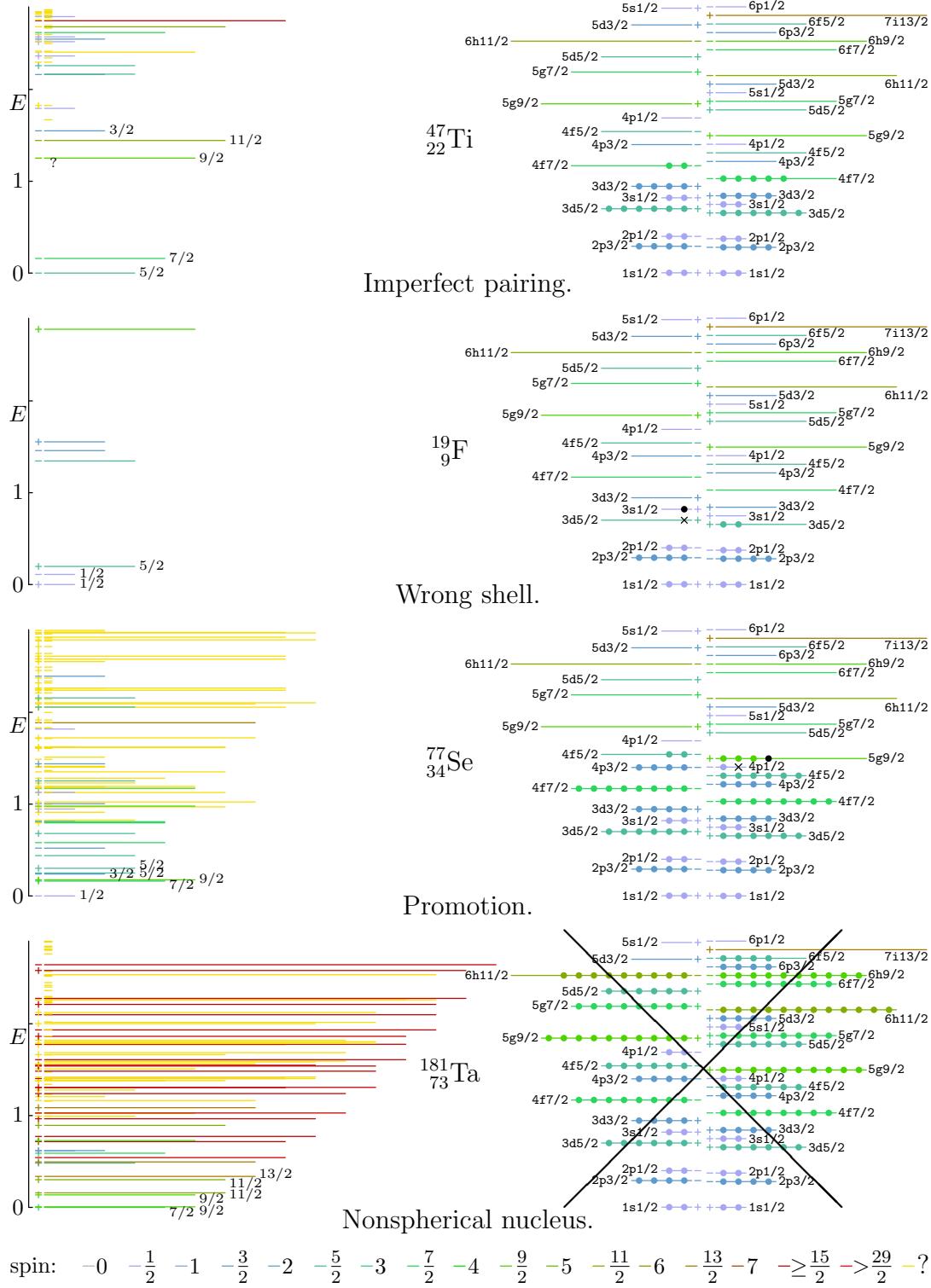


Figure 11.19: Failures of the shell model.

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 $1/2^-$ 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 (11.11). 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 $1/2$ 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 [23, 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.

11.13 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 11.18, and many other even-even nuclei with such energies may be found in figure 11.17. 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 11.12.5.

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 11.17, they are the very small squares. Examples are hafnium-176 in figure 11.18 and tantalum-181 in figure 11.19.

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 11.10.

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.

11.13.1 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 [27, 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 (11.11) 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 (11.12) and (11.11), the nuclear equivalent of the surface tension is

$$\sigma = \frac{C_s}{4\pi R_A^2} \quad (11.19)$$

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.109}. 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

$$\delta = \varepsilon l \sin(\omega t - \varphi) \bar{Y}_l^m(\theta, \phi) \quad (11.20)$$

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, 3.1. The \bar{Y}_l^m are essentially the spherical harmonics of orbital angular momentum fame, chapter 3.1.3. 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 $l \geq 2$, and the frequency of vibration is then,

{A.109}:

$$\omega = \sqrt{\frac{E_{s,l}^2}{\hbar^2} \frac{1}{A} - \frac{E_{c,l}^2}{\hbar^2} \frac{Z^2}{A^2}} \quad (11.21)$$

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,

$$E_{s,l} = \frac{\hbar c}{R_A} \sqrt{\frac{(l-1)l(l+2)}{3} \frac{C_s}{m_p c^2}} \quad E_{c,l} = \frac{\hbar c}{R_A} \sqrt{\frac{2(l-1)l}{2l+1} \frac{e^2}{4\pi\epsilon_0 R_A m_p c^2}} \quad (11.22)$$

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

$$E_{s,2} \approx 35 \text{ MeV} \quad E_{c,2} \approx 5.1 \text{ MeV} \quad (11.23)$$

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, [23, 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.

11.13.2 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

$$E_{\text{vibration}} = \sqrt{E_{s,2}^2 \frac{1}{A} - E_{c,2}^2 \frac{Z^2}{A^2}} \quad E_{s,2} \approx 35 \text{ MeV} \quad E_{c,2} \approx 5.1 \text{ MeV} \quad (11.24)$$

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 11.17. 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, [23, p. 342, 354] or small, [21, 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 10.2. 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 11.15 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 11.20. 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. Nuclei that have a 2^+ lowest excited state, immediately followed by a $0^+, 2^+, 4^+$ triplet, in any order, are marked with a “V” for vibrational. Green squares marked with a V have in addition the rough energy ratio of 2 between the second and first excited energies predicted by the vibrating liquid drop model. These requirements are clearly nontrivial, so it is encouraging that so many nuclei except the very heavy ones satisfy them. Still, many even-even nuclei do not. Obviously liquid-drop vibrations are only a part of the complete story.

Much heavier vibrating nuclei than oxygen-18 are tellurium-120 in figure 11.21 as well as the earlier example of ruthenium-104 in figure 11.18. 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.

As subsection 11.13.1 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.

11.13.3 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 11.20. 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 (11.21), predicts that the nuclear shape

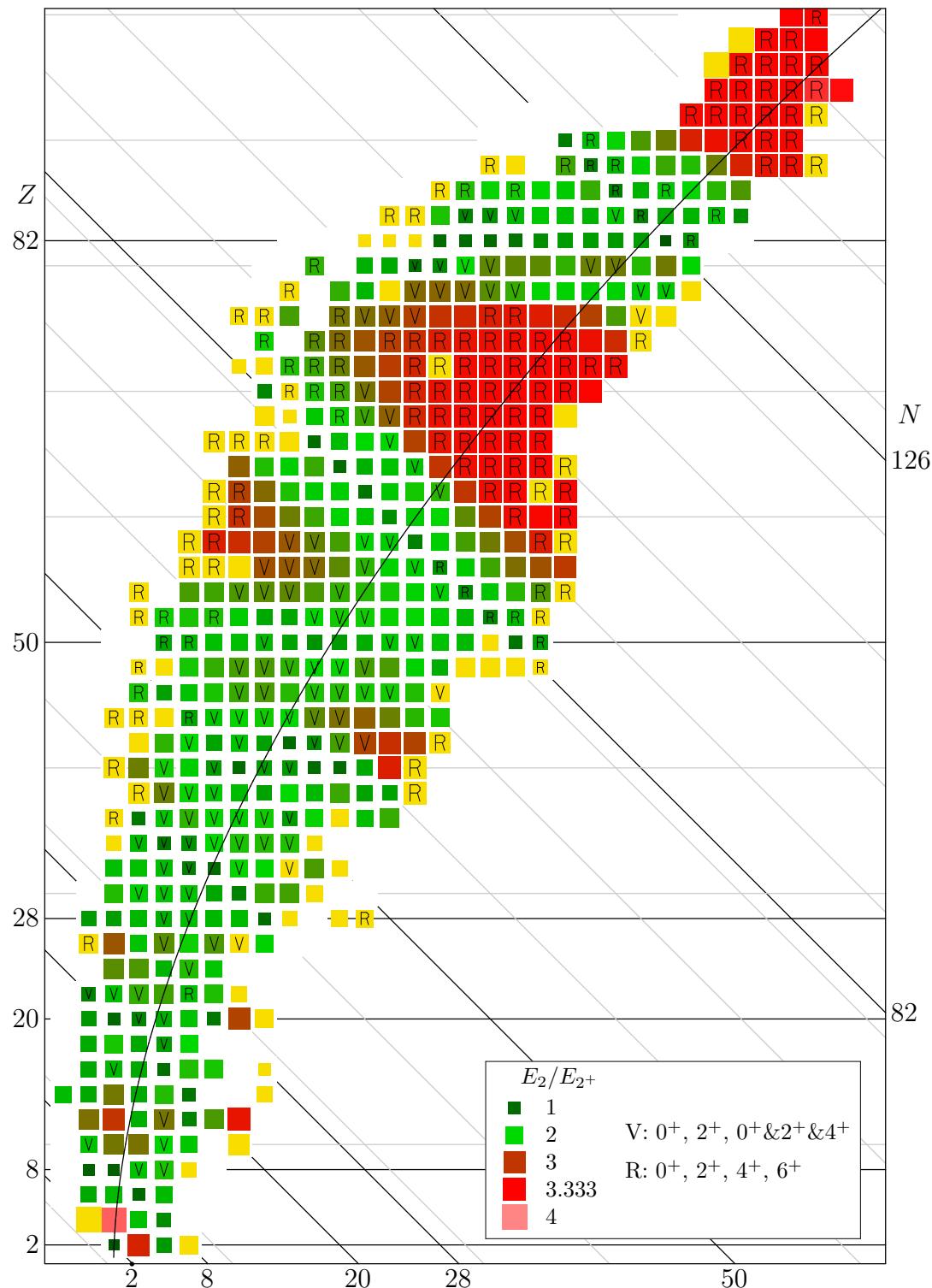


Figure 11.20: An excitation energy ratio for even-even nuclei.

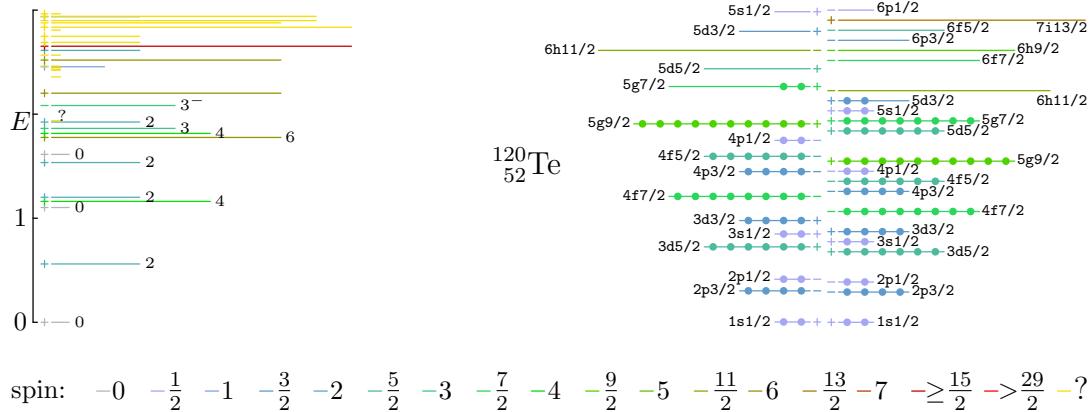


Figure 11.21: Textbook vibrating nucleus tellurium-120.

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 11.13.1, 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.

11.13.4 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.

Basic notions in nuclear rotation

Classically speaking, the kinetic energy of a solid body due to rotation around an axis is $T_R = \frac{1}{2}\mathcal{I}_R\omega^2$, where \mathcal{I}_R is the moment of inertia around the axis and ω the angular velocity. Quantum mechanics does not use angular velocity but angular momentum $L = \mathcal{I}_R\omega$, and in these terms the kinetic energy is $T_R = L^2/2\mathcal{I}_R$. Also, the square angular momentum L^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:

$$T_R = \frac{\hbar^2}{2\mathcal{I}_R} j(j+1) - \frac{\hbar^2}{2\mathcal{I}_R} j_{\min}(j_{\min}+1) \quad (j_{\min} \neq \frac{1}{2}) \quad (11.25)$$

Here j_{\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

$$E_j = E_{\min} + T_R$$

where E_{\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 $\hbar^2/2\mathcal{I}_R$. The classical value for the moment of inertia \mathcal{I}_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 (11.11). Plugging in the numbers, the ballpark for rotational energies becomes

$$\frac{35}{A^{5/3}} [j(j+1) - j_{\min}(j_{\min}+1)] \text{ MeV}$$

For a typical nonspherical nucleus like hafnium-177 in figure 11.22, taking the intrinsic state to be the ground state with j_{\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

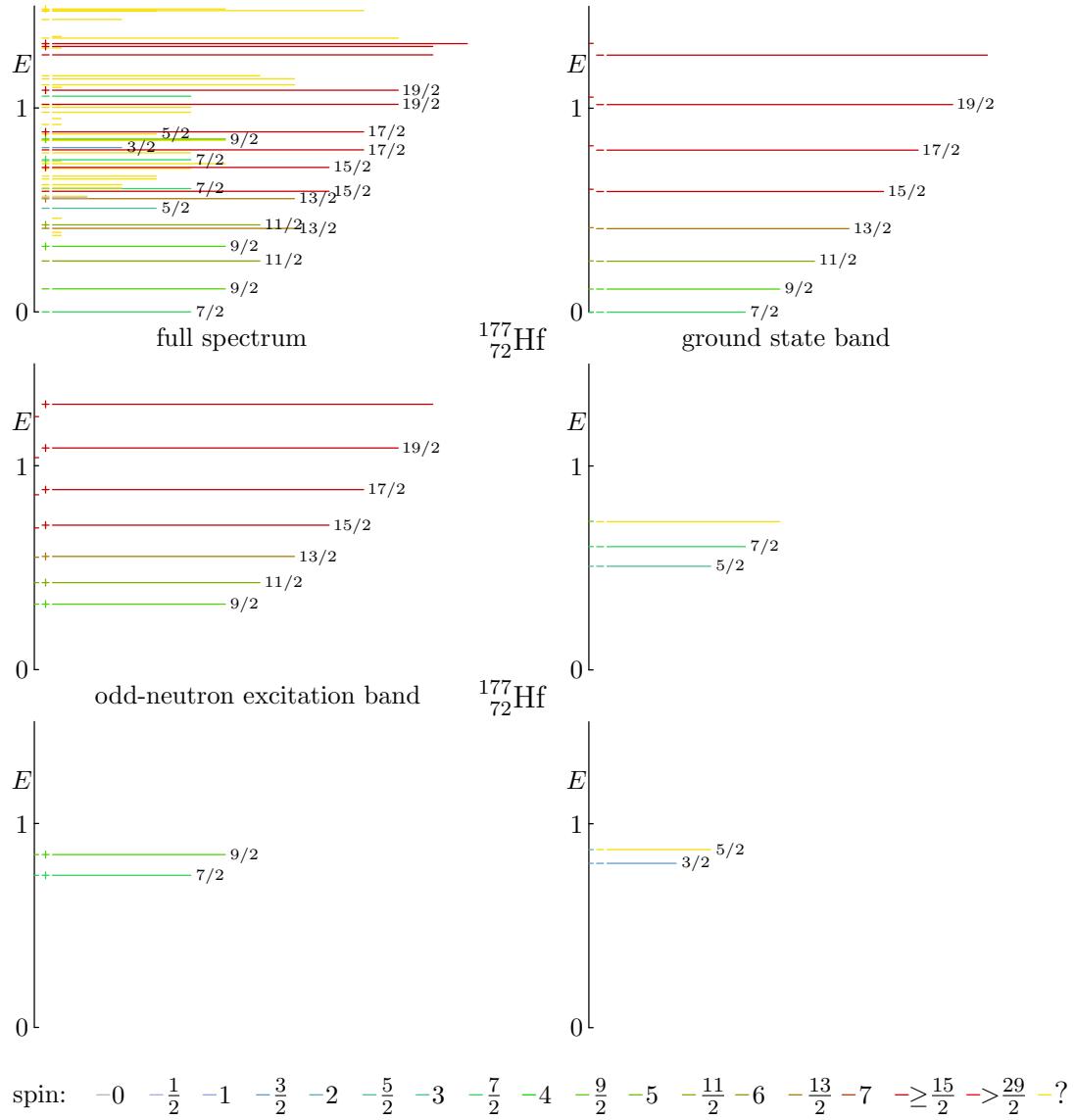


Figure 11.22: Rotational bands of hafnium-177.

as a spheroid shape rotating around an axis normal to the axis of symmetry. As 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 6.1.9. 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, [21, 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 $T_R = \frac{1}{2}I_R\omega^2$, with T_R the kinetic energy of the liquid. If typical numbers are plugged into this picture, [21, 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, [27, pp. 493ff]

Basic rotational bands

Consider the spectrum of the deformed nucleus hafnium-177 in figure 11.22. At first the spectrum seems a mess. However, take the ground state to be an “intrinsic state” with a spin j_{\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 $9/2, 11/2, 13/2, \dots$. And indeed, a sequence of such excited states can be identified in the spectrum, as shown in the top right of figure 11.22. 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 (11.25)? 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 (11.25). 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 (11.25). The energies obtained in this way are indicated by the spin-colored tick marks on the axis in the top-right graph of figure 11.22. 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, (11.25).

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 be 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 $7/2\pm$ around the nuclear axis and odd parity. (For a non-spherical 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 built 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 11.22. (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 11.22. 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, [21, p. 156].

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:

$$\vec{L} = \vec{L}_{\text{rot}} + \vec{L}_{\text{min}}$$

Now in the expression for rotational energy, (11.25) 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, [27, 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 11.13.4, 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_{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 $|j_{\text{min}}\rangle$ with magnetic quantum number j_{min} along the axis and a state $| - j_{\text{min}}\rangle$ with magnetic quantum $-j_{\text{min}}$.

Next, the kinetic energy of rotation is, since $\vec{L}_{\text{rot}} = \vec{L} - \vec{L}_{\text{min}}$,

$$\frac{1}{2\mathcal{I}_R} \vec{L}_{\text{rot}}^2 = \frac{1}{2\mathcal{I}_R} \vec{L}^2 - \frac{1}{\mathcal{I}_R} \vec{L} \cdot \vec{L}_{\text{min}} + \frac{1}{2\mathcal{I}_R} \vec{L}_{\text{min}}^2$$

As long as the middle term in the right hand side averages away, the normal formula (11.25) for the energy of the rotational states is fine. This happens if there is no correlation between the angular momentum vectors \vec{L} and \vec{L}_{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

$$\frac{1}{\mathcal{I}_R} \vec{L} \cdot \vec{L}_{\text{min}} = \frac{1}{\mathcal{I}_R} [L_x L_{x,\text{min}} + L_y L_{y,\text{min}} + L_z L_{z,\text{min}}]$$

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 $|\pm j_{\text{min}}\rangle$ into some combination of states $|\pm j_{\text{min}} \pm 1\rangle$, chapter 10.1.10. Since there is no rotational momentum in the axial direction, \vec{L} and \vec{L}_{min} have quite similar effects on the wave function, but it is not the same, for one because \vec{L} “sees” the complete nuclear momentum. If j_{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 4.3. But if $j_{\text{min}} = \frac{1}{2}$ then $|j_{\text{min}} - 1\rangle$ and $| - j_{\text{min}} + 1\rangle$ are part of the state itself. In that case, the x and y components of the $\vec{L} \cdot \vec{L}_{\text{min}}$ term produces a contribution to the energy that does not average out, and the larger \vec{L} is, the larger the contribution.

The expression for the kinetic energy of nuclear rotation then becomes

$$T_R = \frac{\hbar^2}{2\mathcal{I}_R} \left\{ \left[j(j+1) + a(-1)^{j+\frac{1}{2}} (j + \frac{1}{2}) \right] - [j_{\text{min}}(j_{\text{min}}+1) - a] \right\} \quad j_{\text{min}} = \frac{1}{2}$$

(11.26)

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 $1/2^-$ ground state rotational band of tungsten-183, figure 11.23. To compute the rotational kinetic energies in the band using (11.26) requires the values of both \mathcal{I}_R and a . The measured energies of the $3/2^-$ and $5/2^-$ states above the ground state can be used to do so. That produces a

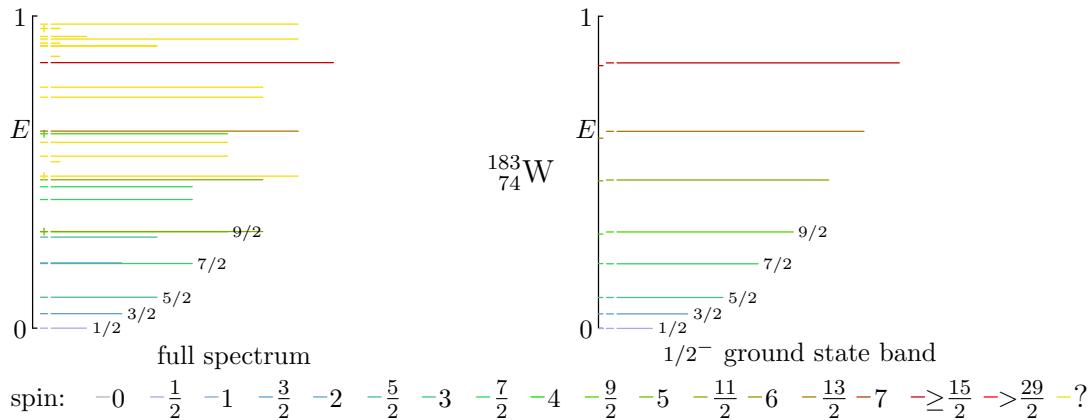


Figure 11.23: Ground state rotational band of tungsten-183.

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 11.23 show, the prediction is quite good. Note in particular that the energy interval between the $9/2^-$ and $7/2^-$ states is *less* than that between 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 11.24. 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 11.24. 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.

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: $0^+, 2^+, 4^+, \dots$

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 cancelling each other.

As an example, consider erbium-164. The ground state band in the top right of figure 11.25 consists of the states $0^+, 2^+, 4^+, \dots$ as expected. The energies initially agree well with the theoretical prediction (11.25) shown as tick marks.

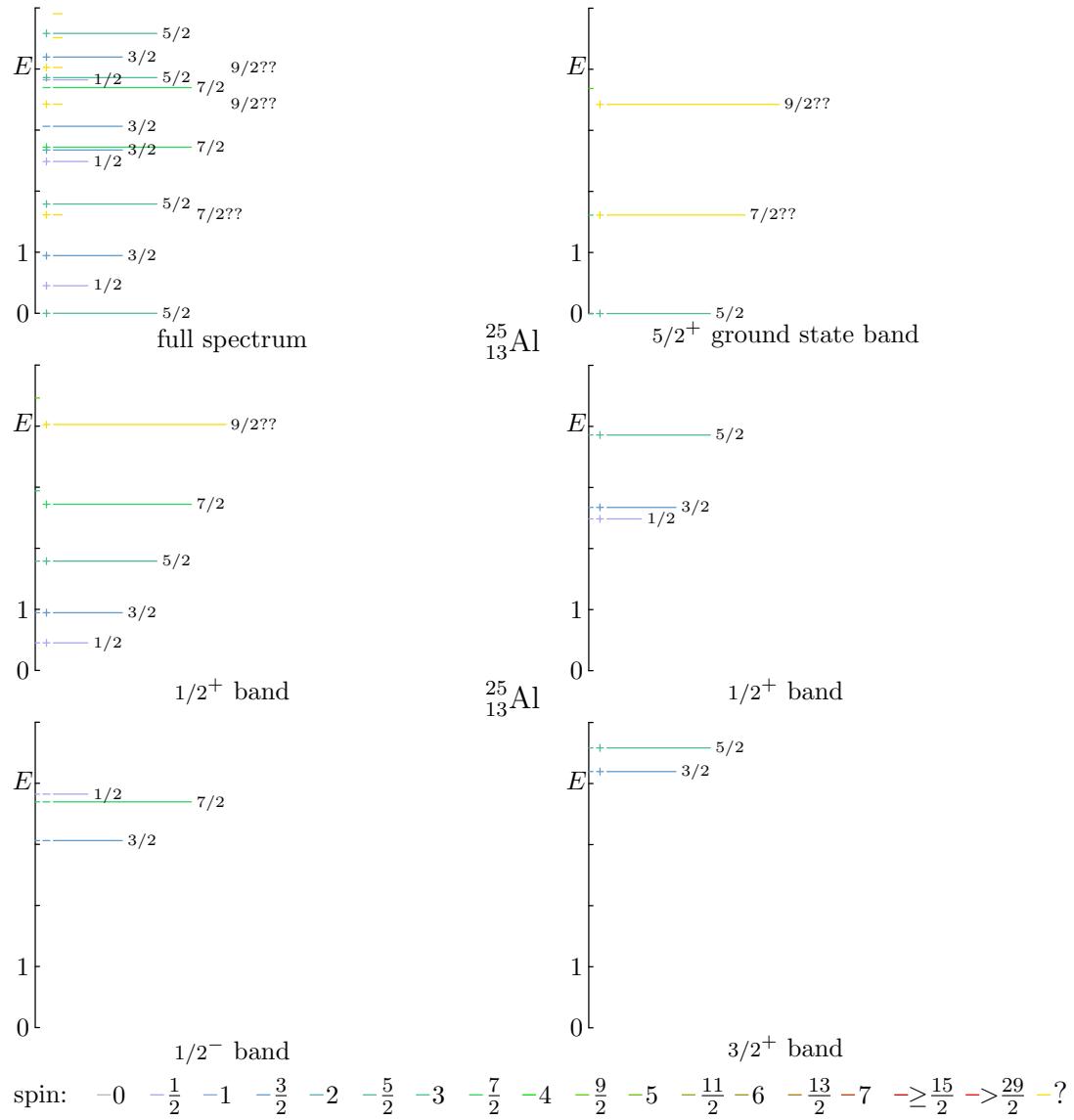


Figure 11.24: Rotational bands of aluminum-25.

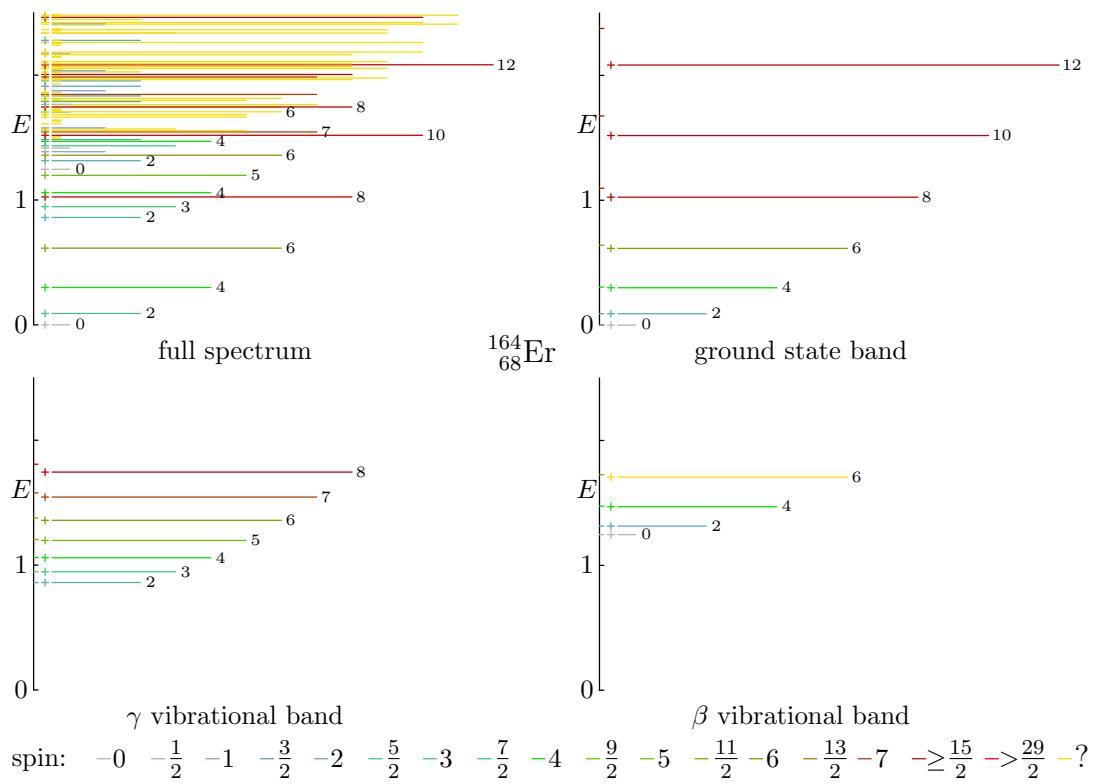


Figure 11.25: Rotational bands of erbium-164.

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 (11.25) again well. The moments of inertia are 46%, 49% and 61% of the solid sphere value.

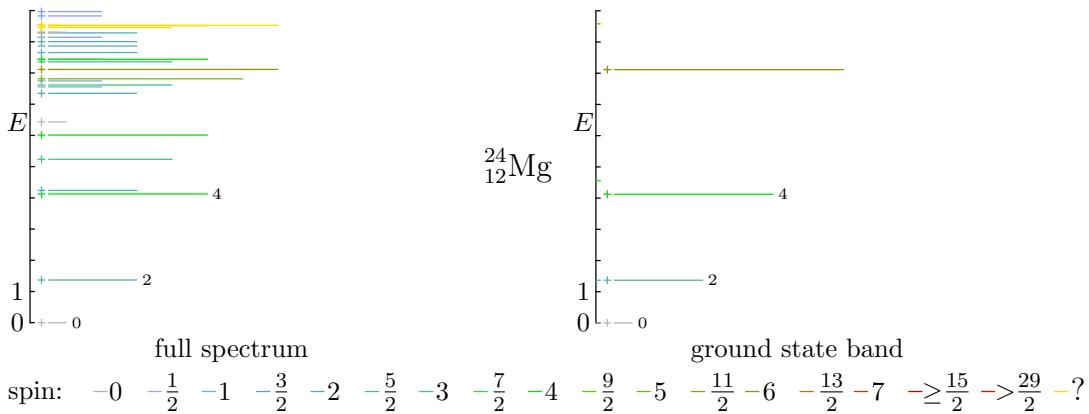


Figure 11.26: Ground state rotational band of magnesium-24.

Light even-even nuclei can also be deformed and show rotational bands. As an example, figure 11.26 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^- , ..., [27, 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.

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 (11.25)

$$\frac{\hbar^2}{2\mathcal{I}_R} 4(4+1) \Big/ \frac{\hbar^2}{2\mathcal{I}_R} 2(2+1) = \frac{10}{3}$$

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 11.20, 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.

Non-axial 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 11.27 shows an example. The primary ground state band in the top right quickly develops big deviations from the axially symmetric theory (11.25) 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.)

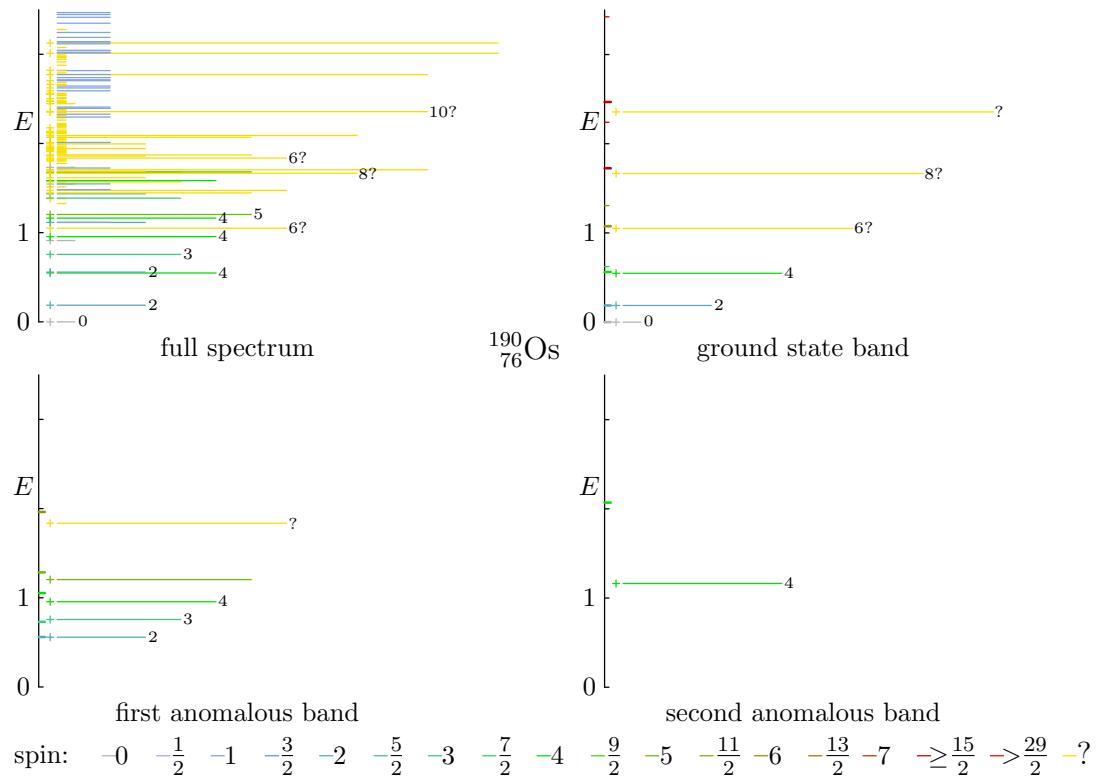


Figure 11.27: Rotational bands of osmium-190.

The predicted energy levels of the first anomalous band also agree well with the predicted values. The identification of the bands was taken from [27, p. 416], but since they do not list the energies of the second anomalous band, that value was taken from [23, 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 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.

11.14 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.

11.14.1 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, 11.2. 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 disproportionately 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. 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 $^{293}_{118}\text{Ei}$ 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 11.11.2 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 11.13.1.

11.14.2 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 [23]. 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 11.28 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 11.7. Indeed, nuclei with 50 protons are seen to have the highest fission energy release in figure 11.28. 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 11.28.

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

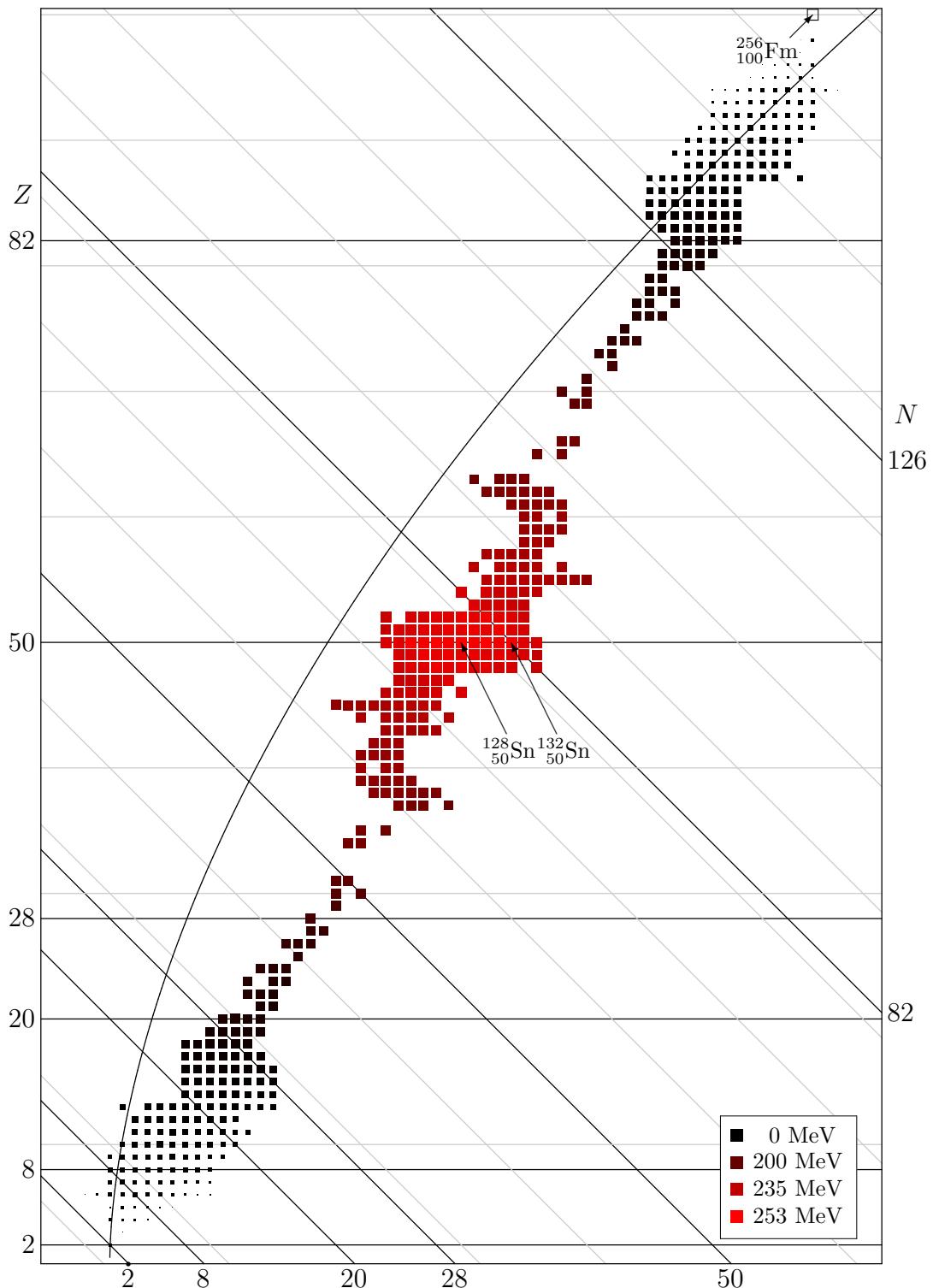


Figure 11.28: Simplified energetics for fission of fermium-256.

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 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.

11.15 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 11.11.3.

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.

11.15.1 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 11.29. There are no known exceptions to this rule.

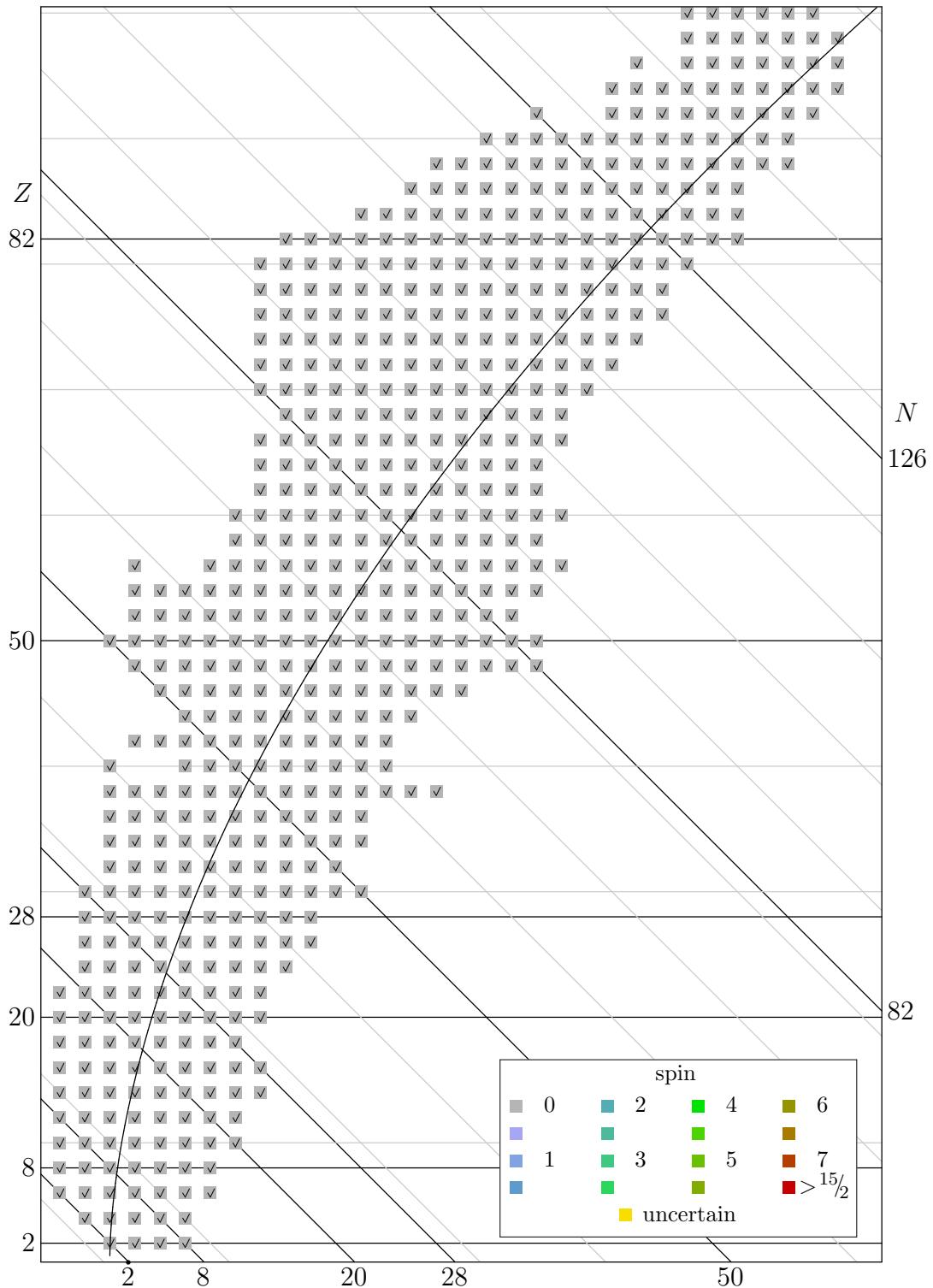


Figure 11.29: Spin of even-even nuclei.

11.15.2 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 11.12.2. This ordering is indicated by the colored lines in figures 11.30 and 11.31. Nuclei for which the resulting nuclear spin prediction is correct are indicated with a check mark.

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. They are the rotational nuclei; roughly the red squares in figure 11.20, the smallest squares in figure 11.17. Their main regions are for atomic number above $Z = 82$ and in the interior of the $Z < 82$, $N > 82$ wedge.

For almost all remaining nuclei near the stable line, the spin can be explained in terms of the shell model using various reasonable excuses, [23, p. 224ff]. However, it seems more interesting to see how many spins can be correctly predicted, rather than justified after the fact. It may be noted that even if the wrong value is predicted, the true value is usually either another value in the same major shell, or one less than such a value.

One modification of the odd-particle shell model has been allowed for in the figures. It is that if the subshell being filled is above one of lower spin, a particle from the lower subshell may be promoted to the higher one; it can then pair up at higher spin. Since the odd nucleon is now in the lower shell, the spin of the nucleus is predicted to be the one of that shell. The spin is lowered. In a sense of course, this gives the theory a second shot at the right answer. However, 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. Also, no promotion was allowed if the nucleon number was below 32. Nuclei for which spin lowering due to promotion can explain the observed spin are indicated with an “L” or “I” in figures 11.30 and 11.31. 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 greatly increases the number of nuclei whose spin can be correctly predicted. 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. For some other nuclei, the

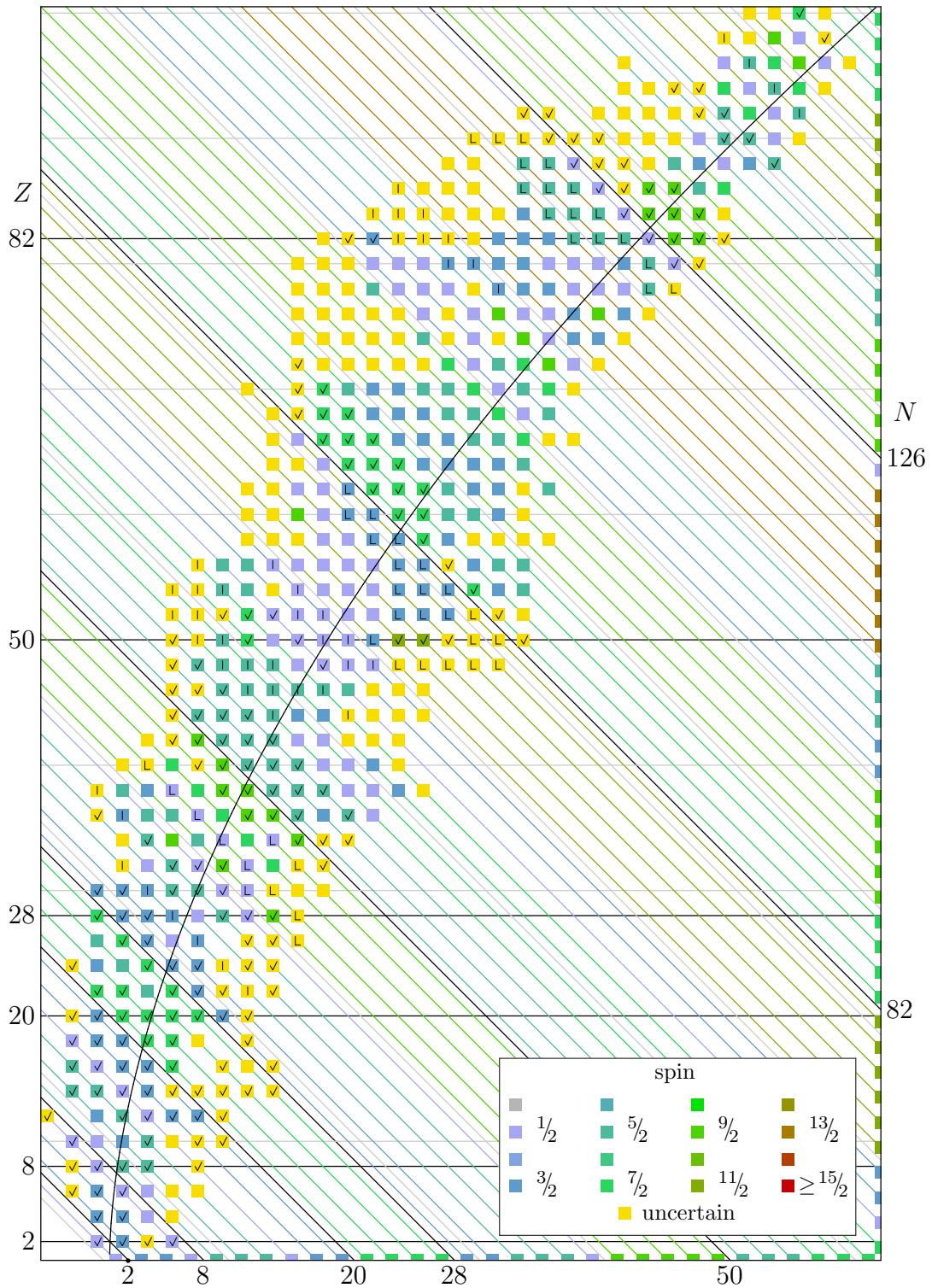


Figure 11.30: Spin of even-odd nuclei.

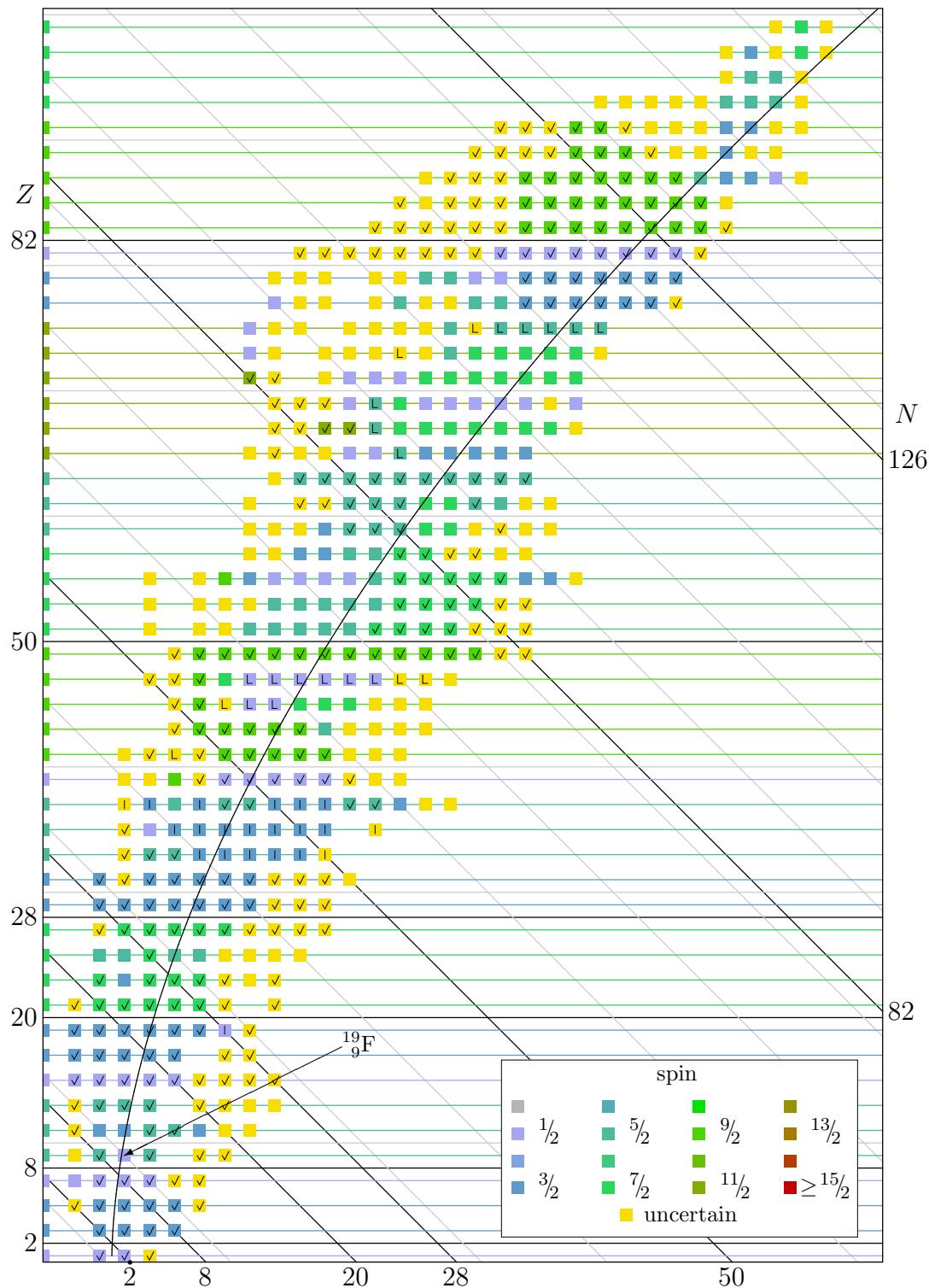


Figure 11.31: Spin of odd-even nuclei.

basic shell model is valid but the odd-particle assumption fails. In particular, for subshells with at least three particles and three holes (empty spots for additional nucleons) the spin is often one unit less than that of the odd nucleon. This 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 discussed in section 11.12.6.

In a few exceptional cases, like the unstable nitrogen-11 and beryllium-11 mirror 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.

11.15.3 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 comes from combining the net angular momenta j^p of the odd proton and j^n of the odd neutron. In particular, quantum mechanics allows any integer value for j in the range

$$|j^p - j^n| \leq j \leq j^p + j^n \quad (11.27)$$

That gives a total of $2 \min(j^p, j^n) + 1$ different possibilities, two at the least. 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 two odd nucleons, just like the deuterium nucleus does. To describe the rules, the net angular momentum j of an odd nucleon and its spin $s = \frac{1}{2}$ will be called “parallel” if $j = l + s$, with l the orbital angular momentum. The idea is that then the spin acts to increase j , so it must be in the same direction as j . (Of course, this is a simplistic one-dimensional picture; angular momentum and spin are really both three-dimensional vectors with uncertainty.) Similarly, the total angular momentum and spin are called “opposite” if $j = l - s$. Following this picture, now make the assumption that the spins of proton and neutron are parallel. Then:

1. The angular momenta j^p and j^n of the odd proton and neutron will be opposite to each other if one is parallel to its spin, and the other is opposite to its spin. In that case the prediction is that the total spin of the nucleus is $j = |j^p - j^n|$.
2. Otherwise the angular momenta will be parallel to each other and the total spin of the nucleus will be $j = j^p + j^n$. New and improved version: if that fails, assume that the angular momenta are opposite anyway and the spin is $j = |j^p - j^n|$.

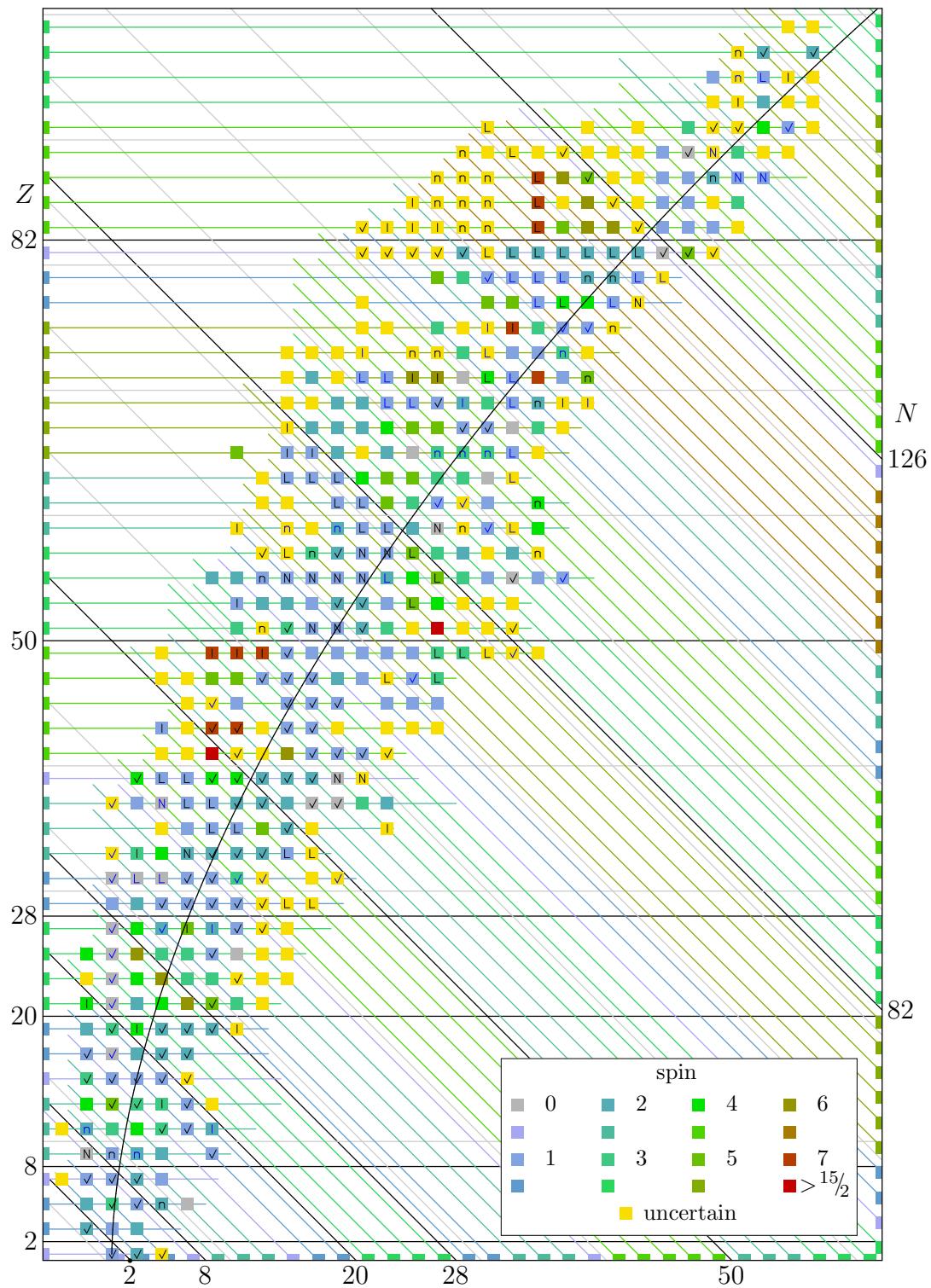


Figure 11.32: Spin of odd-odd nuclei.

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 11.12.2. The nuclei that satisfy the Nordheim rules under that assumption are indicated with a check mark in figure 11.32. A blue check mark means that the new and improved version has been used. 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 [23, p. 239] suggest that the proton and neutron angular momenta be taken from the neighboring pairs of nuclei of odd mass number. Figure 11.33 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 11.33, 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 [23, 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 11.34, 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

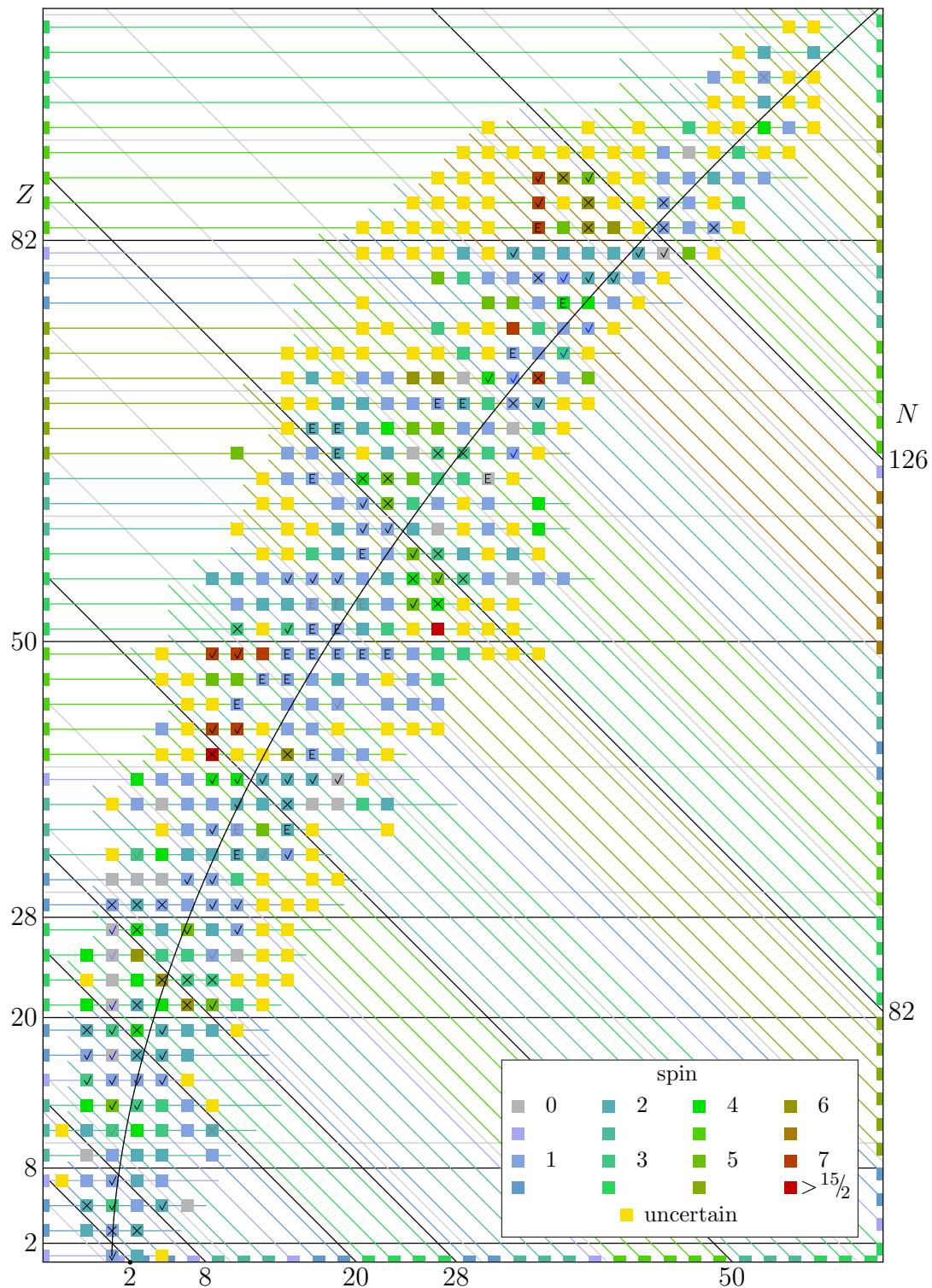


Figure 11.33: Selected odd-odd spins predicted using the neighbors.

helpful if there is supporting evidence.

The nuclei marked with “E” in figure 11.33 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 11.32 from the nuclei with spin zero (grey) and blue check marks.

11.16 Parity Data

The parity of a nucleus is even, or one, if its wave function stays the same if \vec{r} in it is everywhere replaced by $-\vec{r}$. The parity is odd, or minus one, if the wave function gets multiplied by -1 under the same conditions. 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.

11.16.1 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 11.35. There are no known exceptions to this rule.

11.16.2 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 11.12.2. 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 11.36 and 11.37. Failures are in red. Small grey signs are shell model values if the nucleons fill the shells in the normal order.

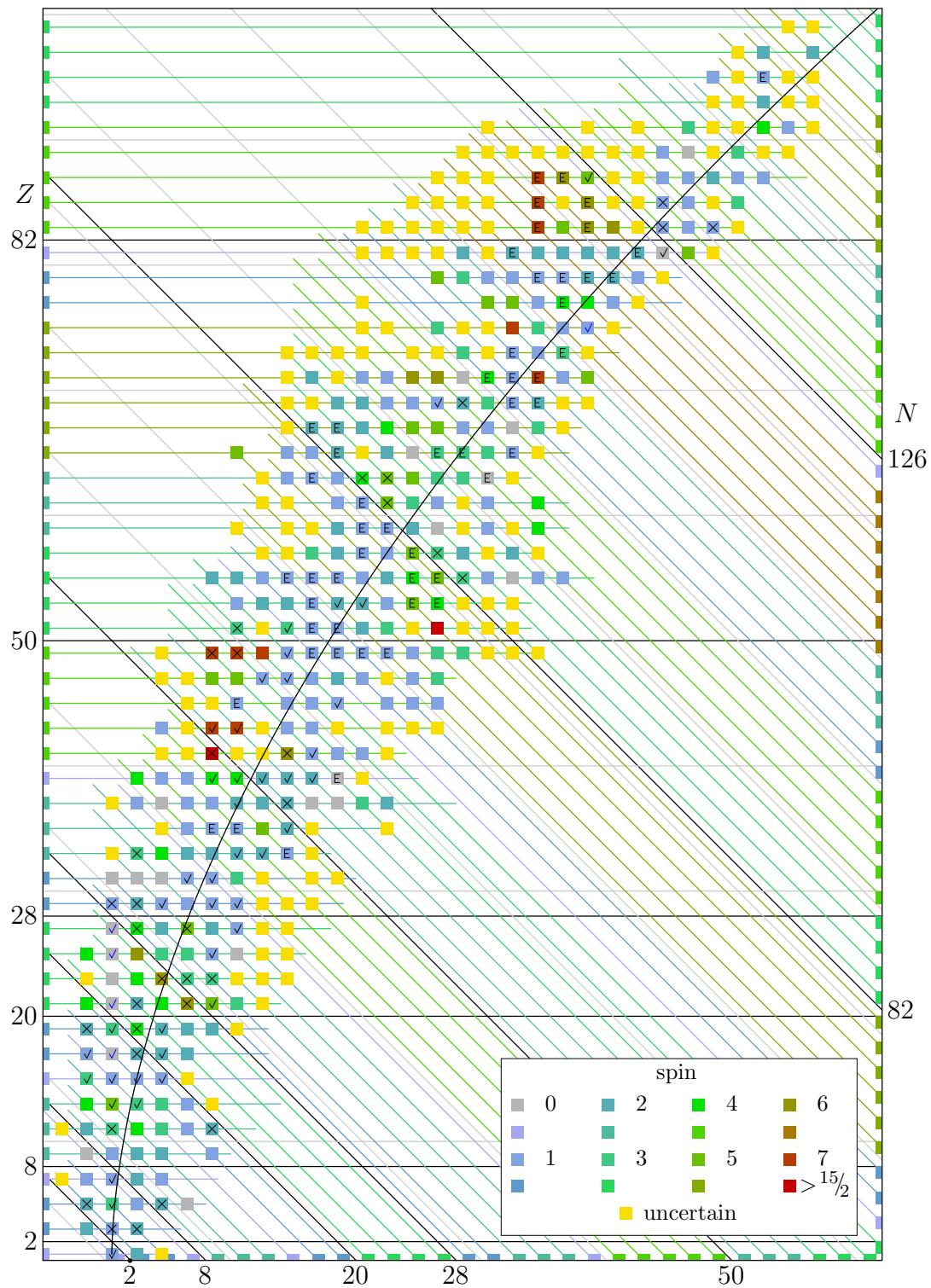


Figure 11.34: Selected odd-odd spins predicted from theory.

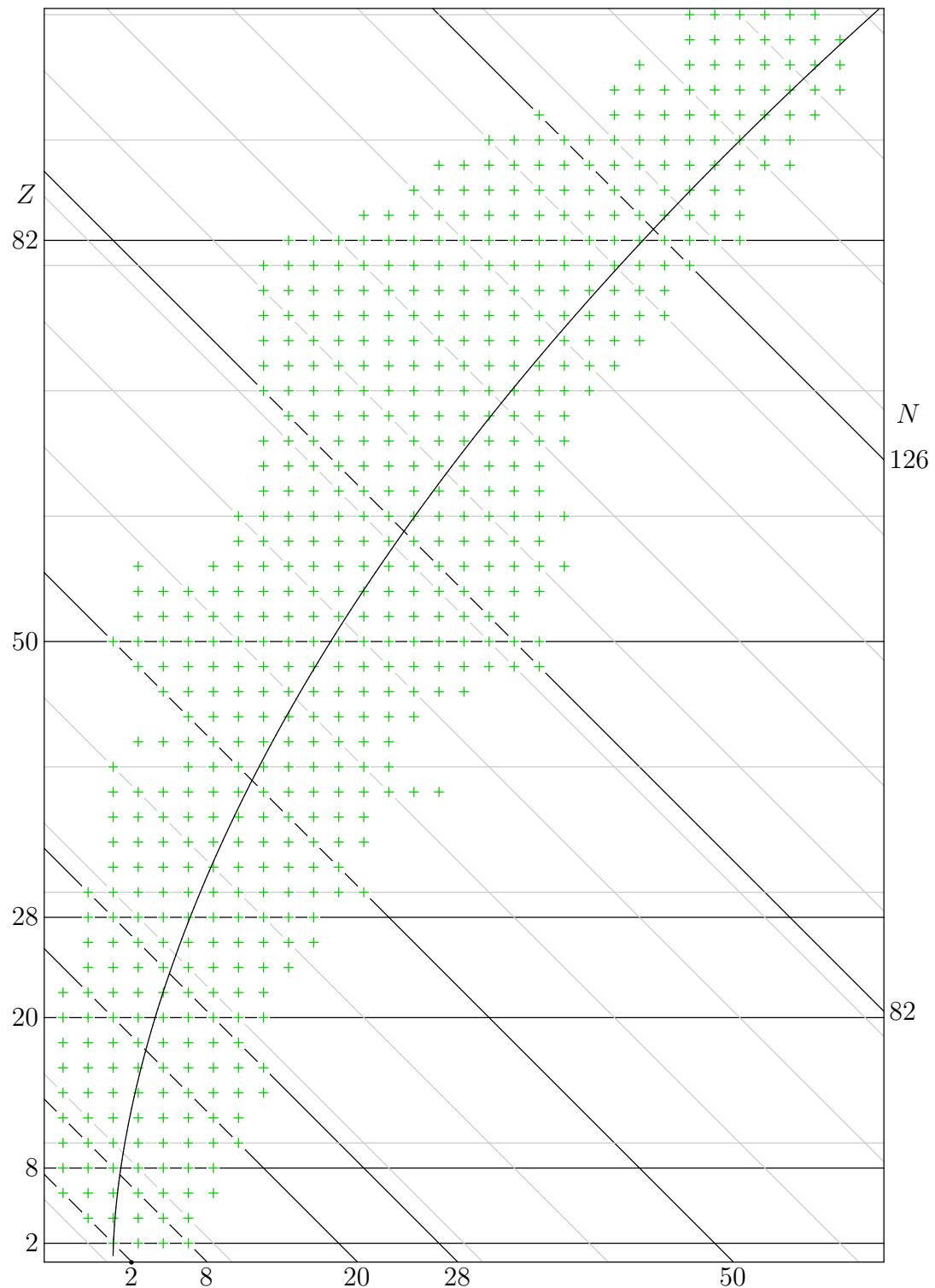


Figure 11.35: Parity of even-even nuclei.

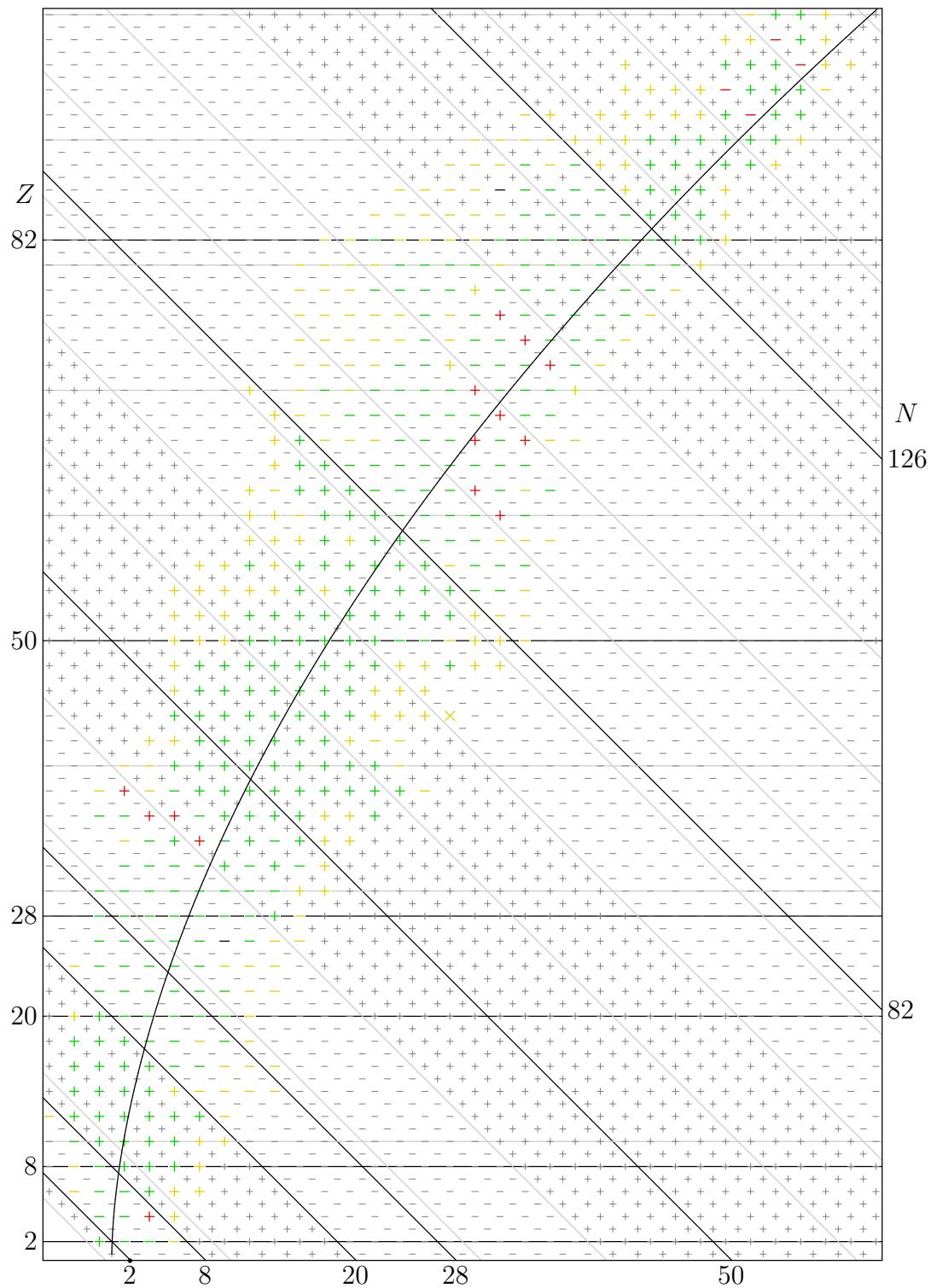


Figure 11.36: Parity of even-odd nuclei.

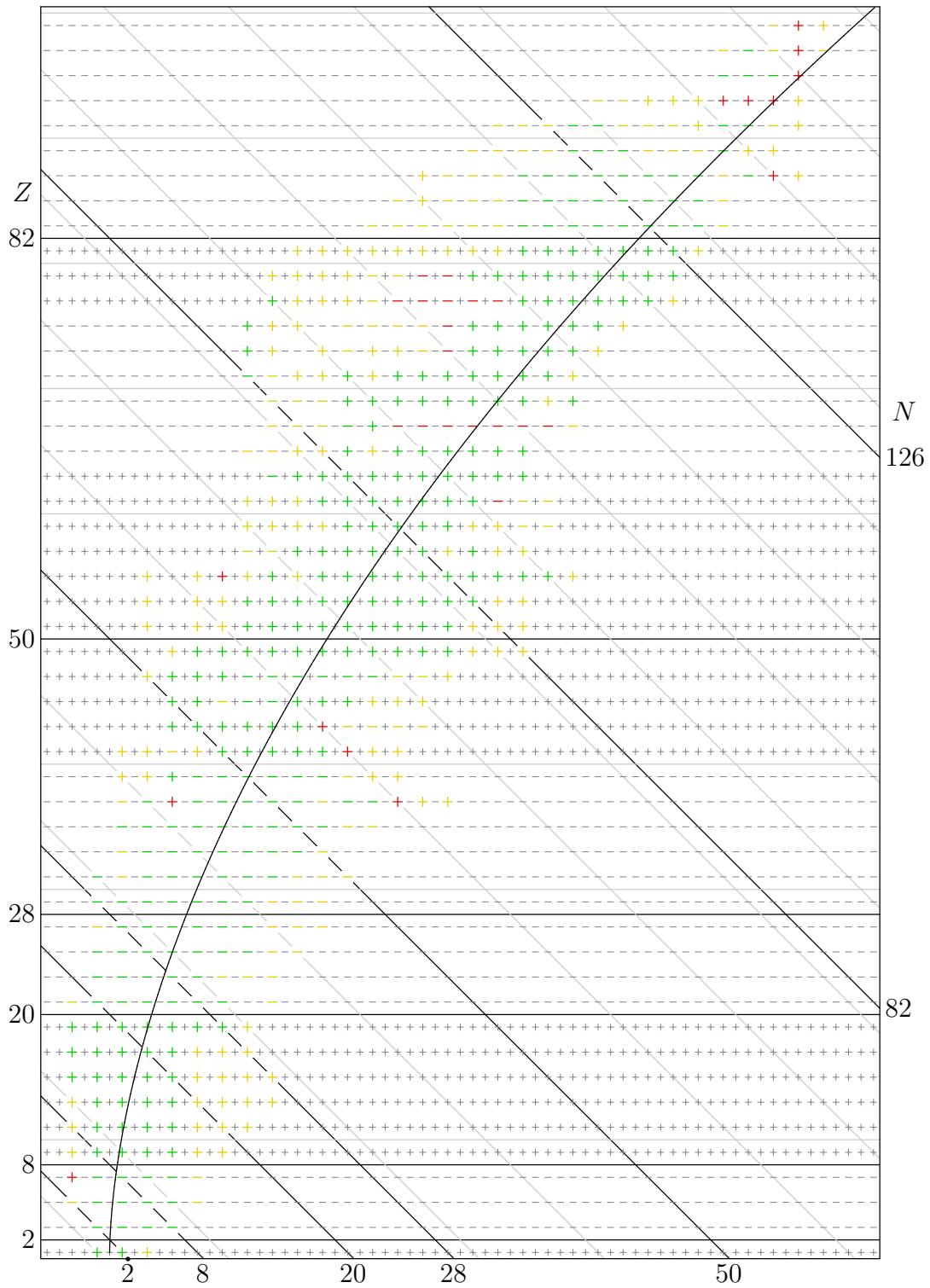


Figure 11.37: Parity of odd-even nuclei.

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 11.12.6.

11.16.3 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 11.38. Hits are green, failures red, and unable-to-tell black. Small grey signs are shell model values.

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.

11.16.4 Parity Summary

Figure 11.39 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 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 11.12.2 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.

11.17 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

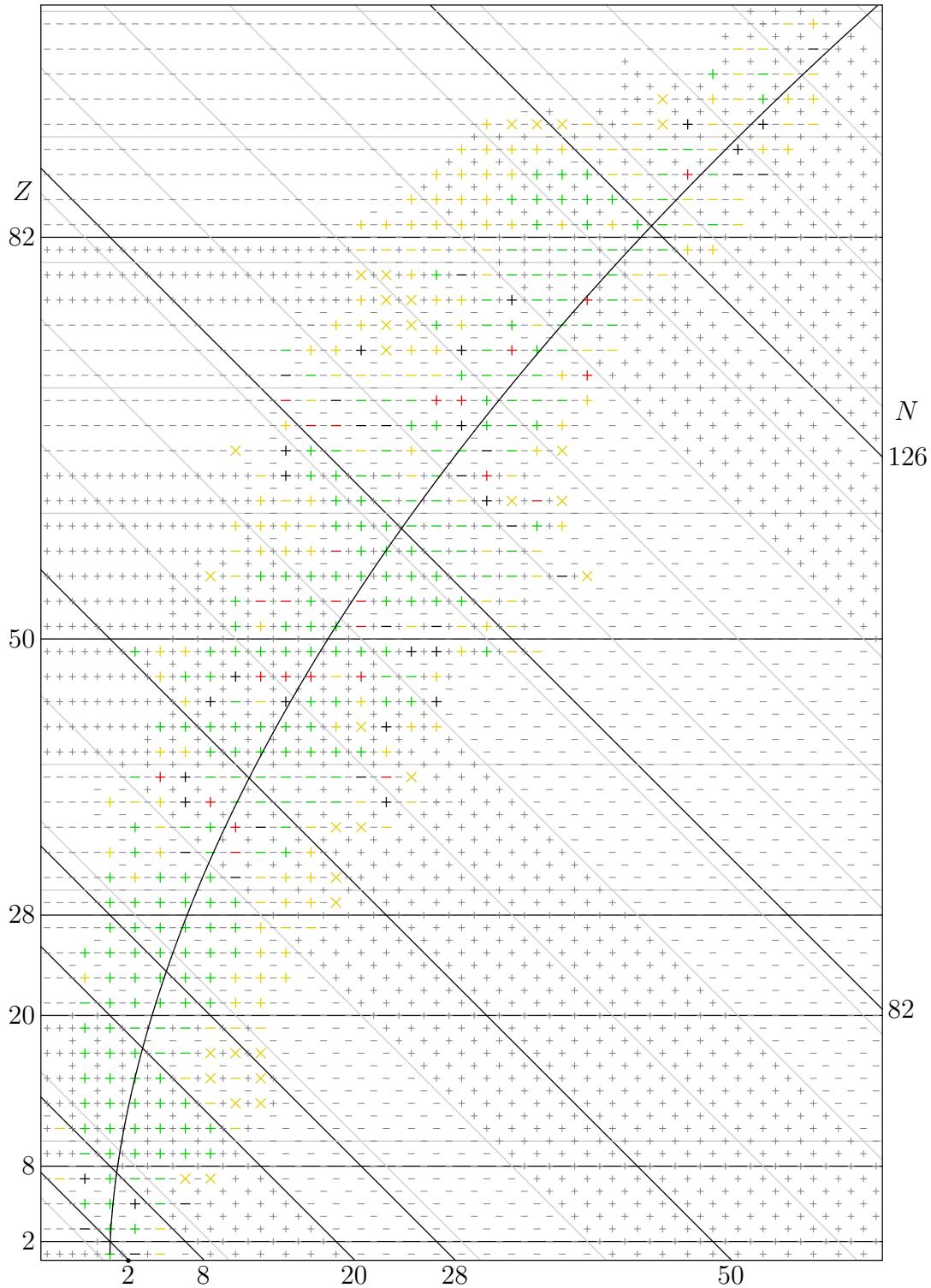


Figure 11.38: Parity of odd-odd nuclei.

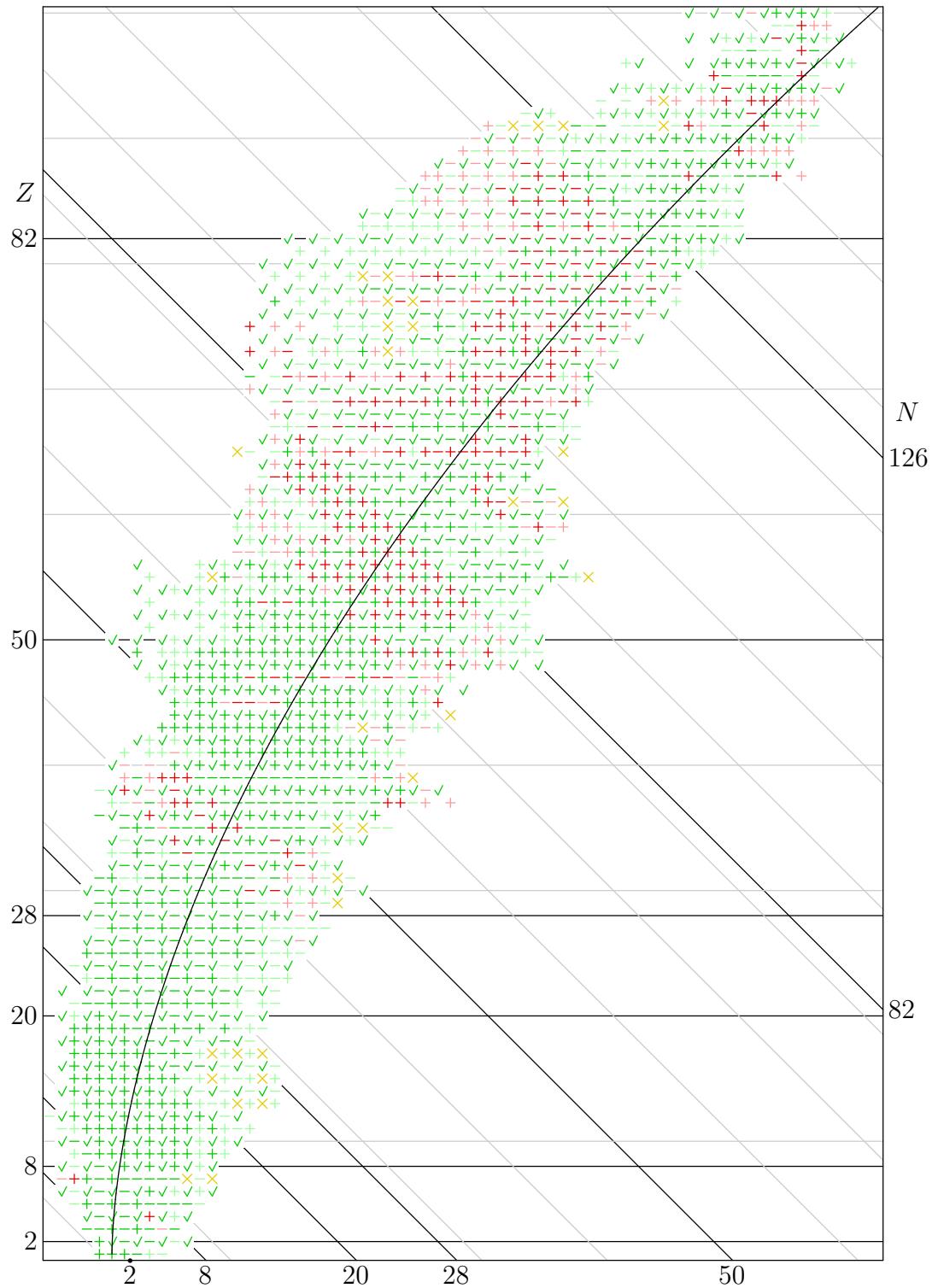


Figure 11.39: Parity versus the shell model.

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.

11.17.1 Classical description

This subsection explains the magnetic dipole and electric quadrupole moments from a classical point of view.

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 electrodynamics 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 10.4 or 10.6.

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{B}_{ext} . The energy involved in this alignment is

$$-\vec{\mu} \cdot \vec{B}_{\text{ext}}$$

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

$$V = \int \varphi \rho_c d^3\vec{r}$$

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

$$V = \varphi_0 \int \rho_c d^3\vec{r} + \sum_{i=1}^3 \left(\frac{\partial \varphi}{\partial r_i} \right)_0 \int r_i \rho d^3\vec{r} + \sum_{i=1}^3 \sum_{j=1}^3 \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial r_i \partial r_j} \right)_0 \int r_i r_j \rho d^3\vec{r}$$

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 cancelled 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 $\int r_i r_j \rho d^3\vec{r}$ for $i \neq j$, so what is left is

$$V = V_{pc} + \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial x^2} \right)_0 \int x^2 \rho d^3\vec{r} + \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial y^2} \right)_0 \int y^2 \rho d^3\vec{r} + \frac{1}{2} \left(\frac{\partial^2 \varphi}{\partial z^2} \right)_0 \int z^2 \rho d^3\vec{r}$$

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 10.4. All that then leads to a definition of an electric quadrupole moment for a single axis, taken to be the z -axis, as

$$Q \equiv \frac{1}{e} \int (3z^2 - r^2) \rho d^3\vec{r}$$

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 confusing 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 10.5.3.

11.17.2 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.

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_p . Therefore, a good unit to describe magnetic strengths in terms of is

the so-called “nuclear magneton”

$$\mu_N \equiv \frac{e\hbar}{2m_p} \quad (11.28)$$

In those terms, the magnetic magnetic dipole moment operator of a single proton is

$$\frac{1}{\hbar} \hat{\vec{L}}_p \mu_N$$

But quantum mechanics brings in a complication, chapter 10.6. 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

$$\hat{\vec{\mu}}_p = \frac{1}{\hbar} \left(\hat{\vec{L}} + g_p \hat{\vec{S}} \right) \mu_N \quad g_p \approx 5.59 \quad (11.29)$$

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:

$$\hat{\vec{\mu}}_n = \frac{1}{\hbar} g_n \hat{\vec{S}} \mu_N \quad g_n \approx -3.83 \quad (11.30)$$

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

$$\hat{\vec{\mu}} = \frac{1}{\hbar} \left[\sum_{i=1}^Z \left(\hat{\vec{L}}_i + g_p \hat{\vec{S}}_i \right) + \sum_{i=Z+1}^A g_n \hat{\vec{S}}_i \right] \mu_N \quad (11.31)$$

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 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:

$$\Delta E = \langle \Psi | -\hat{\mu}_z B | \Psi \rangle$$

The fact that that is possible is a consequence of small perturbation theory, as covered in chapter 12.1.

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 \hat{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 \hat{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 $g\mu_N$ 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 $m_j = j, j-1, \dots, -j$. The numerical value of the magnetic dipole moment μ is therefore *defined* to be the expectation value of $\hat{\mu}_z$ in the nuclear state in which m_j has its largest value j , call it the $|jj\rangle$ state:

$$\mu \equiv \langle jj | \hat{\mu}_z | jj \rangle \quad (11.32)$$

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 6.2. See also the discussion of the Zeeman effect, and in particular the weak Zeeman effect, in chapter 12.1.

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 3.1.3. A magnetic dipole strength requires directionality,

there must be a north pole and a south pole. That cannot occur for nuclei of spin zero.

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 :

$$Q \equiv \langle jj | \sum_{i=1}^Z 3z_i^2 - r_i^2 | jj \rangle \quad (11.33)$$

Note that there is a close relation with the spherical harmonics;

$$3z^2 - r^2 = \sqrt{\frac{16\pi}{5}} r^2 Y_2^0 \quad (11.34)$$

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 $j \leq \frac{1}{2}$.

It makes nuclei with spin $\frac{1}{2}$ 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.

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 subsection, what is needed is the expectation value of the magnetic moment

operator $\hat{\mu}_z$ 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,

$$\mu = \frac{1}{\hbar} \langle \psi_{nljj} | L_z + g_p \hat{S}_z | \psi_{nljj} \rangle \mu_N$$

while for an even-odd nucleus

$$\mu = \frac{1}{\hbar} \langle \psi_{nljj} | g_n \hat{S}_z | \psi_{nljj} \rangle \mu_N$$

The unit μ_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 \hat{L}_z and spin \hat{S}_z following chapter 10.1.7, 2.

It is then found that for an odd proton, the magnetic moment is

$j = l - \frac{1}{2} :$	$\mu_{p1} = \frac{1}{2} \frac{j}{j+1} (2j+3-g_p) \mu_N$	
$j = l + \frac{1}{2} :$	$\mu_{p2} = \frac{1}{2} (2j-1+g_p) \mu_N$	

(11.35)

while for an odd neutron

$j = l - \frac{1}{2} :$	$\mu_{n1} = -\frac{1}{2} \frac{j}{j+1} g_n \mu_N$	
$j = l + \frac{1}{2} :$	$\mu_{n2} = \frac{1}{2} g_n \mu_N$	

(11.36)

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:

$$Q = Q_p = \langle \psi_{nljj} | 3z^2 - r^2 | \psi_{nljj} \rangle$$

Evaluation, {A.110}, gives

$Q_p = -\frac{2j-1}{2j+2} \langle r^2 \rangle$	
--	--

(11.37)

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

$$\langle r^2 \rangle \approx \frac{3}{5} R^2$$

Next consider the case that there are not one but $I \geq 3$ 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 11.12.4, 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:

$$Q_{2j_p} = -Q_p \quad (11.38)$$

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, [21, 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.

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

$$\boxed{\text{even-even, ground state band: } \mu = g_R j \mu_N \quad g_R \approx \frac{Z}{A}} \quad (11.39)$$

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, [27, p. 392],

$$\boxed{\mu = \left[g_R j + \frac{j_{\min}^2}{j+1} (g_{\text{int}} - g_R) \right] \mu_N \quad g_R \approx \frac{Z}{A} \quad j_{\min} \neq \frac{1}{2}} \quad (11.40)$$

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.16}.

A model of a spheroidal nucleus produces the following relationship between the intrinsic quadrupole moment and the one that is measured:

$$Q = \frac{3j_{\min}^2 - j(j+1)}{(j+1)(2j+3)} Q_{\text{int}} \quad (11.41)$$

where j_{\min} is the angular momentum of the nucleus when it is not rotating. Derivations may be found in [27] or [23]. 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.

11.17.3 Magnetic moment data

Figure 11.40 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 $0.1 \mu_N$, 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 $1/2^-$ 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 qualitative 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 (11.40). 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, [27, 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 11.40. Interestingly enough, the average is like a collective rotation, (11.39).

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 11.40 that that just does not happen.

Even-even nuclei do not have magnetic moments in their ground state. Figure 11.41 shows the magnetic moments of the first excited 2^+ state of these nuclei. The values are in fairly good agreement with the prediction (11.39) 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

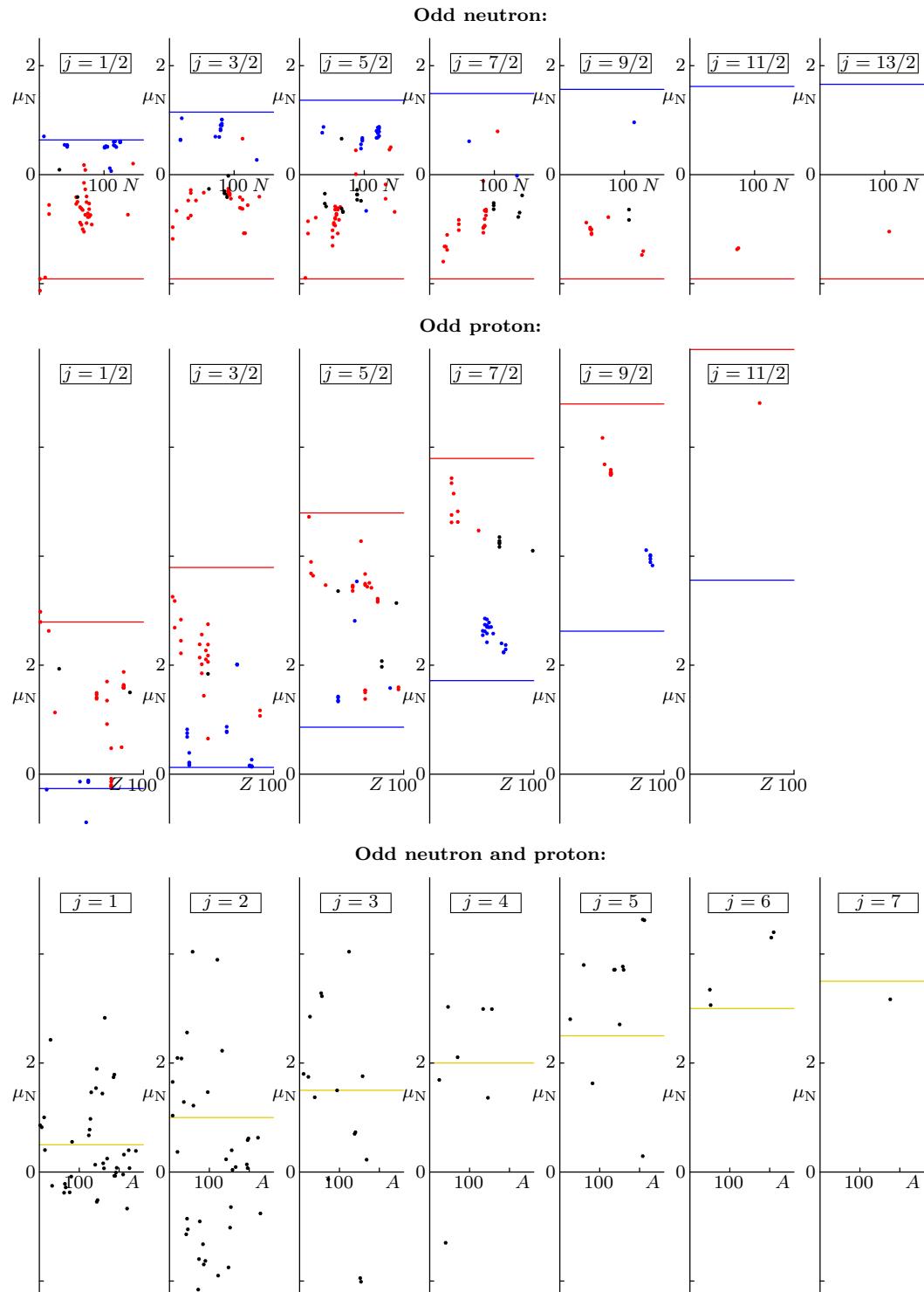
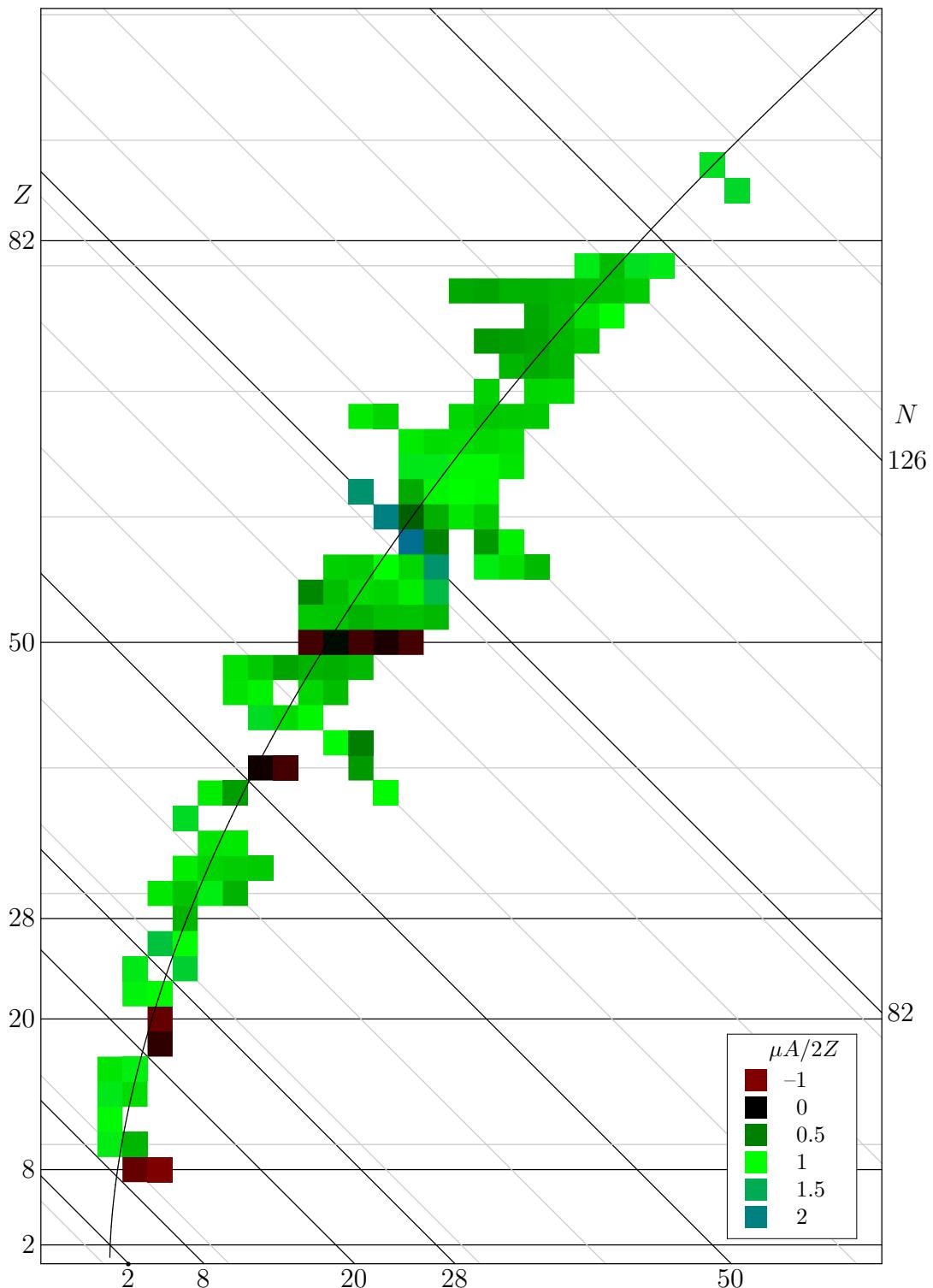


Figure 11.40: Magnetic dipole moments of the ground-state nuclei.

Figure 11.41: 2^+ magnetic moment of even-even nuclei.

maximum error in the shown data is about a quarter of a nuclear magneton, subjectively evaluated.

11.17.4 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

$$Q_{p \text{ ballpark}} \sim \frac{3}{5} R^2$$

where R is the nuclear radius computed from (11.11). This value corresponds to the area of the square marked “a proton’s” in the legend of figure 11.42. As discussed in subsection 11.17.2, 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 11.42 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

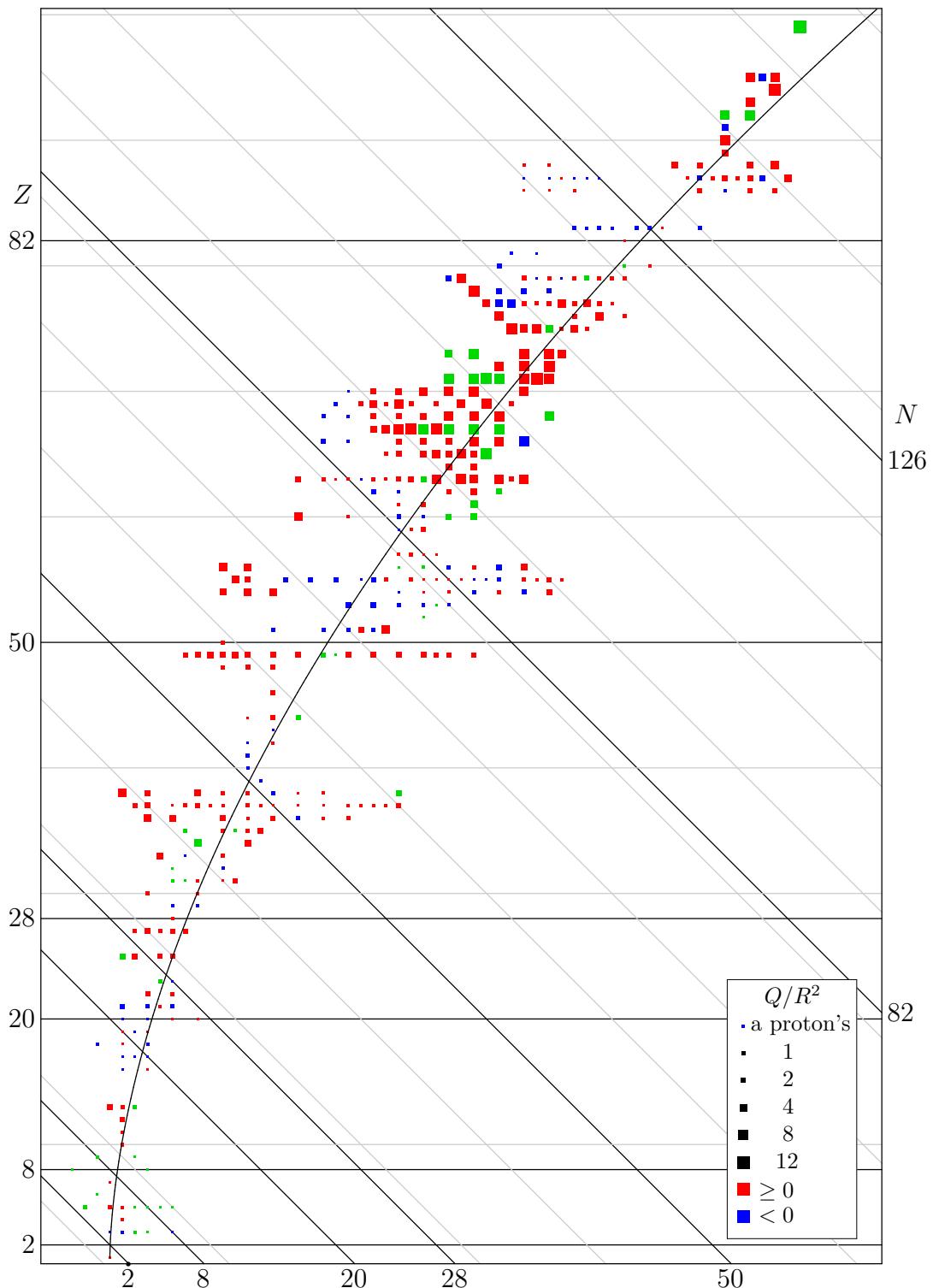


Figure 11.42: Electric quadrupole moment.

for these nuclei. They are the deformed nuclei that also showed up in figures 11.17 and 11.20. 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.

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 11.43 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 (11.41), 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 11.43 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 11.42 leaves out magnesium-23, whose reported quadrupole moment of 1.25 barn is far larger than 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.

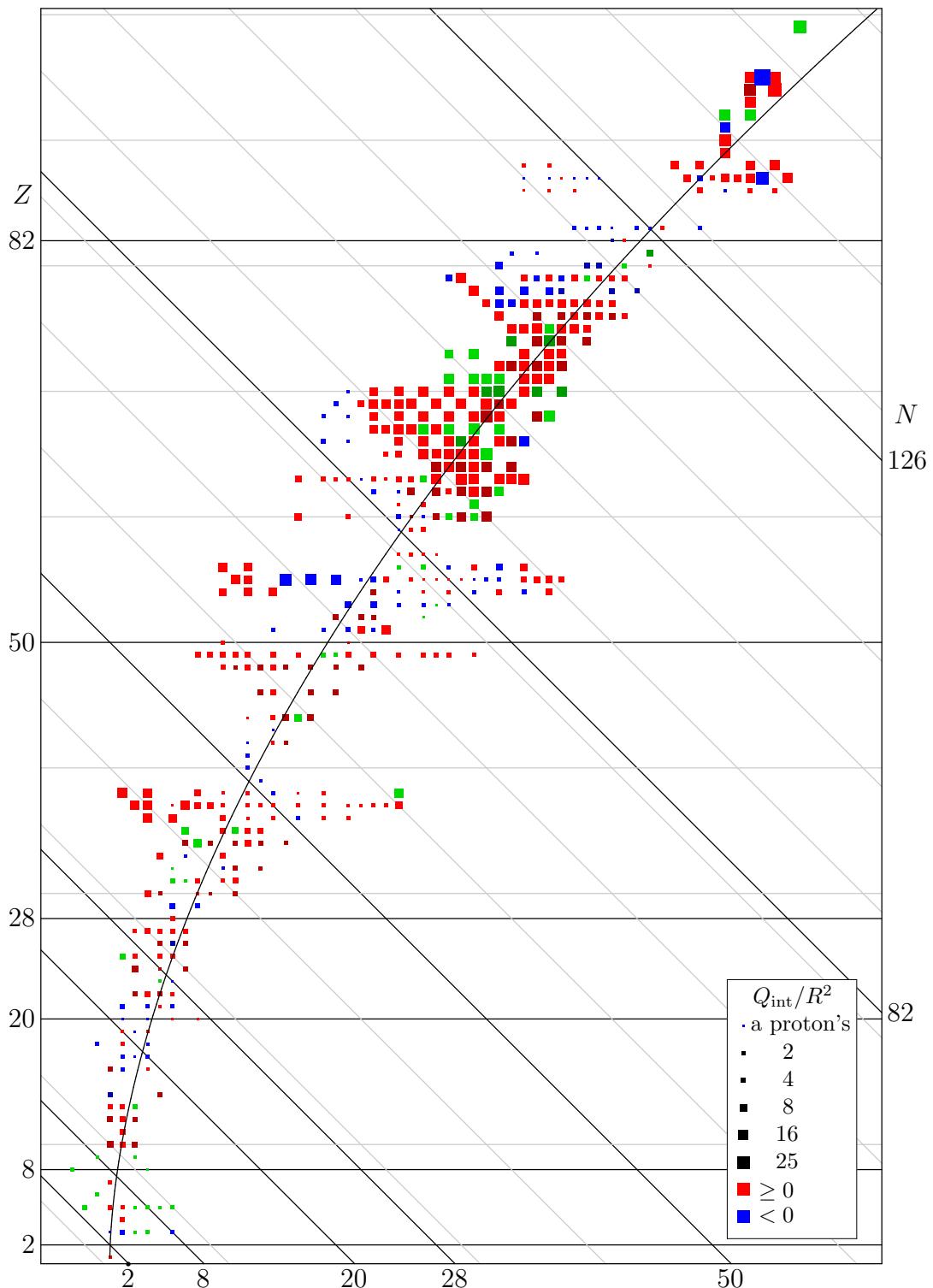


Figure 11.43: Electric quadrupole moment corrected for spin.

11.18 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.

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 type, called nucleon, and give that particle a property called “subtype.” If the subtype is $\frac{1}{2}$, it is a proton, and if the subtype is $-\frac{1}{2}$ it is a neutron. That makes subtype a property that is mathematically much like the spin S_z in a given z -direction. But of course, there is no physical “subtype axis;” nucleons do not change type if you rotate the coordinate system. Therefore subtype is conventionally indicated by the symbol T_3 , not T_z , with no physical meaning attached to the 3-axis.

As far as the exclusion principle is concerned, there is no problem. Earlier, a single spatial state could hold two protons, one with spin up and the other with spin down, and two neutrons, also one with spin up and the other with spin down. Now the same state can hold four nucleons, because with two spin states times two subtype states, there are four different nonspatial states for the nucleons. Still, there is a change in thinking.

Consider for example the possible states of a two-nucleon system like the deuteron. In the deuteron, the two nucleons are in the same spatial state and in a triplet spin state, say both spin up. As far as subtype is concerned, generally speaking the possible subtype combinations for a two-nucleon system are:

$$\text{pp} \quad \text{pn} \quad \text{np} \quad \text{nn}$$

where p stands for proton type and n for neutron type. But for the deuteron, one nucleon must be a proton and the other a neutron. That is so if the combined subtype state is either pn or np.

But in the new way of thinking these states by themselves are no longer acceptable. A system of two identical nucleons should be antisymmetric under exchange of the nucleons, and neither the pn nor the np state is consistent with that requirement. (Note that in the new way of thinking the nucleons are considered identical particles whether or not one is a proton and the other a neutron; that is just like they are considered identical particles whether or not one is spin up and the other spin down.) To satisfy the antisymmetrization requirement, it is useful to rewrite the subtype states as antisymmetric singlet and symmetric triplet states. That is just like was done for spin states in chapter 4.5.6; it gives:

$$\text{singlet: } \frac{\text{pn} - \text{np}}{\sqrt{2}} \quad \text{triplet: } \text{pp} \quad \frac{\text{pn} + \text{np}}{\sqrt{2}} \quad \text{nn}$$

Under the given deuteron conditions of symmetric spatial and spin states, the subtype state must be the antisymmetric singlet one. Otherwise the antisymmetrization requirement is not satisfied. So there is a 50% chance that nucleon 1 is a proton and nucleon 2 a neutron, and a 50% chance that it is the other way around. Protons and neutrons have lost their identity.

That is a change in thinking. It still makes no physical difference as long as the potential is formulated such that the pn and np parts are each energy eigenfunctions by themselves. Suppose you formulate an energy eigenstate using nucleon 1 as proton and nucleon 2 as neutron, as you would normally want to do. That would be the energy eigenfunction of the pn state. Obviously, if you simply renumber the neutron as nucleon 1 and the proton as nucleon 2, it is physically still the same energy eigenstate. It is just renamed the np state. And a linear combination of two eigenstates of the same energy is still an eigenstate of that energy. So the singlet subtype state will have the same energy as the pn and np states separately.

But chapter 4.3 showed that the energy of a system that can be in either one of two equivalent states, like the pn and np states here, might be lowered or raised by taking linear combinations of the states. That only happens if the two states are not really energy eigenstates, but only approximations to it. In particular, chapter 6.1.5 showed that there must be a possibility for one state to change in the other. In the present context, there must be a possibility for the proton and neutron to swap identity. Such a swap is in fact observed in proton-neutron scattering experiments, [21, p. 110]. To that extent, the new way of thinking affects the nuclear models that you would formulate for the deuteron.

For two nucleon systems in general, if either the spatial or the spin state would be antisymmetric, then any of the triplet subtype states would be allowed. There are three of them. The pp state has net $T_3 = 1$ and would correspond to ^2He , the diproton. The $(\text{pn} + \text{np})/\sqrt{2}$ state has $T_3 = 0$ and would correspond to an excited state of ^2H , the deuteron. The nn state has net $T_3 = -1$ and would correspond to the dineutron. As far as the Coulomb force can be ignored and the nuclear force is truly charge independent, all of these three very different nuclei would have the same energy.

However, excited states of the deuteron do not exist, and neither do the diproton or dineutron, (except conceivably for extremely short times.) Still, there is a lesson here: the value of T_3 does not make a difference if electromagnetic effects, and the weak force, can be ignored. The three triplet states are all the same as far as the nuclear force is concerned. Even if they are physically different nuclei.

What does make a difference is whether it is a singlet or a triplet state. If this was nucleon spin instead of nucleon subtype, you would say that what makes a difference is the net square spin \vec{S}^2 or its quantum number s : a singlet

spin state has $s = 0$ while every triplet one has $s = 1$.

Apparently, it is worthwhile to extend the idea of square spin to square subtype \vec{T}^2 . But of course, spin is a three dimensional vector, (with at most one nonzero component that can be certain, that is.) The 3-axis of subtype is completely artificial, so what to make of T_1 and T_2 ? The solution is to define corresponding operators. By definition, the operator \hat{T}_3 turns a p state into $\frac{1}{2}\mathbf{p}$ and an n state into $-\frac{1}{2}\mathbf{n}$. A “charge creation” operator T^+ is defined to turn an n state into a suitable multiple of a p state. So in effect it creates a unit charge e . Because there are no nucleons with charge $2e$, T^+ is defined to turn an p state into zero. Similarly, the T^- charge annihilation operator is defined to turn a p state into a multiple of an n state, and an n state into zero. Next the operators \hat{T}_1 and \hat{T}_2 are defined as

$$\hat{T}_1 = \frac{1}{2}T^+ + \frac{1}{2}T^- \quad \hat{T}_2 = -i\frac{1}{2}T^+ + i\frac{1}{2}T^-$$

This makes the mathematics of subtype just like that of spin, chapter 10.1 or 12.2. The quantum number of square subtype, the equivalent of the azimuthal quantum number s for spin, will be defined to be t_t . Since \hat{T}_3 has already been defined to be free of any factor \hbar , there is no need for a separate quantum number for it. However, various sources still define, rather superfluously, the equivalent of the Pauli spin matrices to be $\tau_i = 2\hat{T}_i$ for $i = 1, 2, 3$.

With all three components defined in operator form, subtype becomes a very useful tool to understand isobars, nuclei with the same mass number A . Since each proton contributes to the net T_3 an amount $\frac{1}{2}$ and each neutron an amount $-\frac{1}{2}$, the total T_3 is

$$T_3 = \frac{1}{2}(Z - N) \tag{11.42}$$

In other words, the value of T_3 is fixed for a given nucleus. In particular $-2T_3$ gives the neutron excess of the nucleus. Also, the maximum value that t_t can have is $\frac{1}{2}A$. And since t_t must be at least $|T_3|$, the minimum that t_t can be is $\frac{1}{2}|Z - N|$.

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 . For example, among the $A = 2$ isobars, deuterium has $T_3 = 0$. Then the lowest possible value of t_t is zero too. And that is indeed the ground state value, as seen above. A state with $t_t = 1$, if it existed, would be an excited state. And the versions of such a $t_t = 1$ state with $T_3 = 1$, respectively -1 , instead of zero would describe bound states of the diproton, respectively dineutron.

But there is no stable $t_t = 1$ state for two nucleons. A better example is provided by the $A = 14$ isobars. Figure 11.44 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. The von Weizsäcker value

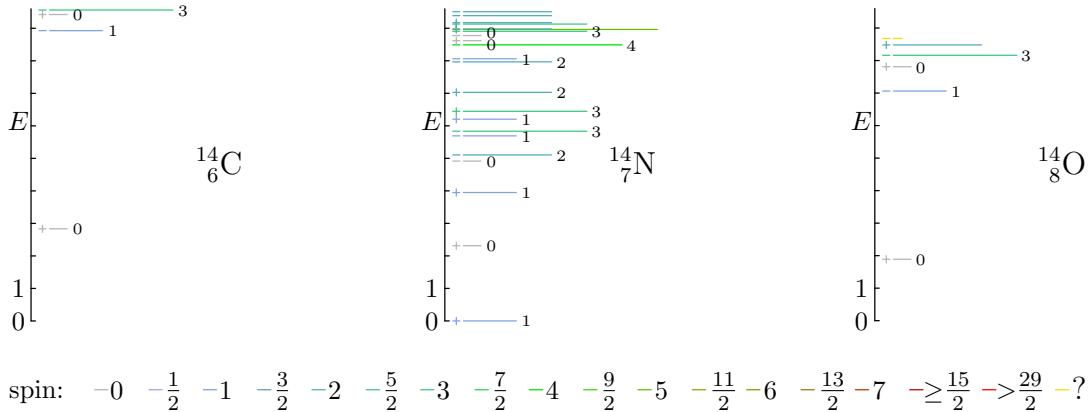


Figure 11.44: Isobaric analog states.

for the Coulomb energy has been subtracted to isolate the nuclear force effects. (The net effect is simply to shift the normal spectra of carbon-14 and oxygen-14 up by 2.83, respectively 1.89 MeV.)

Carbon-14 and oxygen-14 are mirror nuclei, so you would expect them to have pretty much the same sort of energy levels. Indeed their lowest four energy states have identical spin and parity and similar energies. Also, both even-even nuclei have a 0^+ ground state, as they should. But note the surprising result that the odd-odd nitrogen-14 nucleus actually has less energy in its 1^+ ground state than these even-even nuclei. Normally even-even nuclei have less energy than odd-odd ones.

Subtype arguments can explain that quite nicely. Nitrogen-14 assumes a $t_t = 0$ state in its ground state. However, carbon-14 and oxygen-14 cannot do so, because they have $T_3 = \pm 1$, and that makes a $t_t = 1$ state the best they can manage. Such a state has more energy, raising their ground state energies.

Traces of the lower energy of light nuclei with $T_3 = 0$ can also be detected in figures like 11.2, 11.3, and 11.4, T_3 being zero straight above the deuteron.

Next, every $t_t = 1$ state should have three versions: $T_3 = -1$, a carbon-14 state, $T_3 = 1$, an oxygen-14 state, and $T_3 = 0$, which means nitrogen-14. Such sets of states, different only in the value of T_3 , are called “isobaric analog states.”

For example, the ground states of carbon-14 and oxygen-14 should have $t_t = 1$ and about the same energy, and they do. More interestingly, there should be a similar 0^+ state with $t_t = 1$ in the nitrogen-14 spectrum. Indeed, if you inspect the energy levels for nitrogen-14, exactly halfway in between the carbon-14 and oxygen-14 ground state energies, there it is!

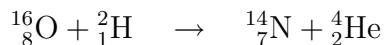
The first three excited levels of carbon-14 and oxygen-14 show the same features. In each case there is a corresponding nitrogen-14 state with exactly the

same spin and parity and $t_t = 1$ right in between the 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.)

Figure 11.44 also shows that nitrogen-14 has a lot more low energy states than carbon-14 or oxygen-14. Subtype explains 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. 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 at all.

Despite such imperfections, consideration of subtype is quite helpful for understanding the energy levels of light nuclei. Now all that is needed is a good name. “Subtype” or “subclass” 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 subtype and spin, since both follow the exact same mathematics. To maximize that potential for confusion, physicists decided to call subtype “spin.” Of course, physicists themselves still have to know whether they are talking about spin or subtype. Therefore some physicists talked about “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 subtype 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:



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. 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 11.44. Indeed, experiments show that the lowest excited state is only produced in negligible amounts compared to the surrounding states.

Selection rules for which reactions and 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 conserves neither T_3 nor \vec{T} , as beta decay shows.

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 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.

11.19 Beta decay

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.

11.19.1 Energetics Data

In beta decay, or more specifically, beta-minus decay, a nucleus converts a neutron into a proton by emitting an electron. An electron antineutrino is also emitted. The number of neutrons N decreases by one unit, and the number of protons Z increases by one.

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 neutrino and electron and the (much smaller) one that the nucleus itself picks up during the decay.

Still, the bottom line is that the nuclear mass reduction must be at least the rest mass of the electron. In energy units, it must be at least 0.511 MeV, the

rest mass energy of the electron.

Figures 11.45 through 11.48 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 excesses.

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 nuclei are right on top of the band.

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 hand half square is red. The energy release is now positive downward, so it is now below the center bands that the mass reduction is sufficient to produce the rest mass of a positron that can carry the proton's positive charge away. The positron, the anti-particle of the electron, has the same mass but opposite charge as the electron.

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.

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.

Electron capture is also called inverse beta decay. However, it is not really the inverse process. The reason is that these processes also emit neutrinos. In particular, beta decay emits an electron antineutrino in addition to the electron. Inverse beta decay does not capture an electron antineutrino, just an electron; instead it emits an electron neutrino. In fact, during the formation of neutron stars, inverse beta decay on steroids, tremendous amounts of these neutrinos are emitted, taking along a large amount of the available energy of the star.

More generally, in nuclear reactions absorption of a particle works out much the same as emission of the corresponding anti-particle, at least as far as conserved quantities other than energy are concerned. 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 beta decay, a neutron with spin $\frac{1}{2}$ produces a proton and an electron,

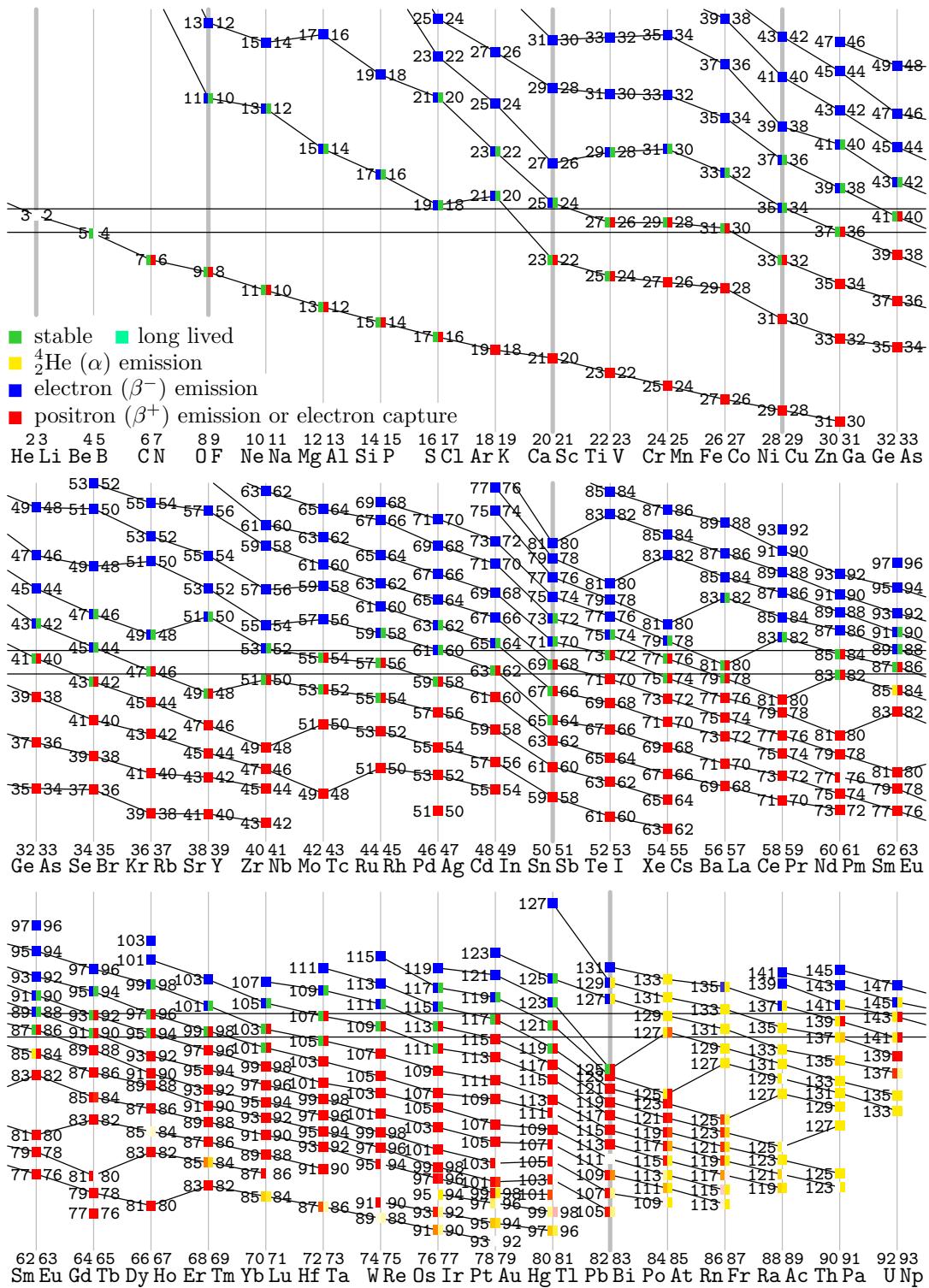


Figure 11.45: Energy release in beta decay of even-odd nuclei.

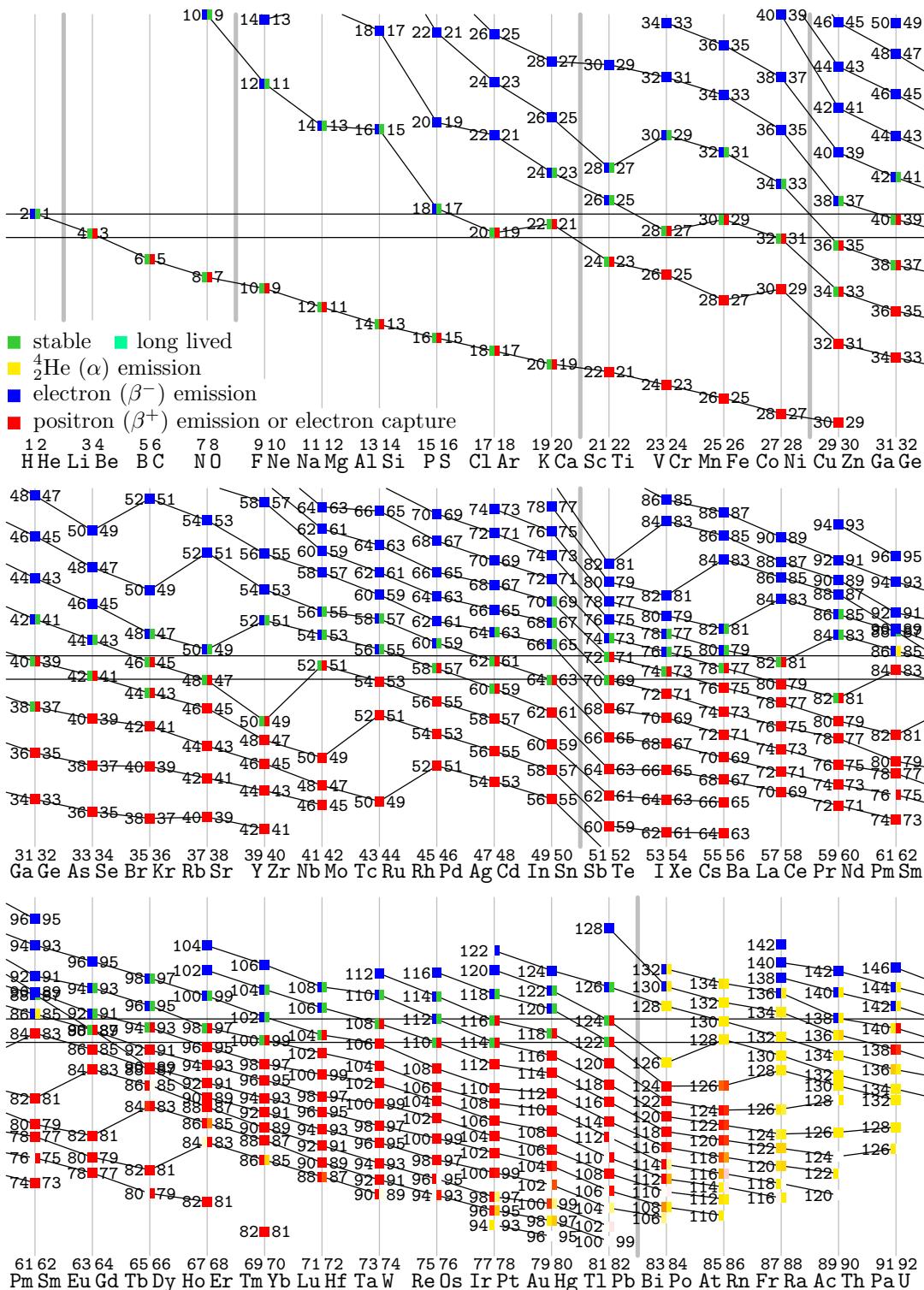


Figure 11.46: Energy release in beta decay of odd-even nuclei.

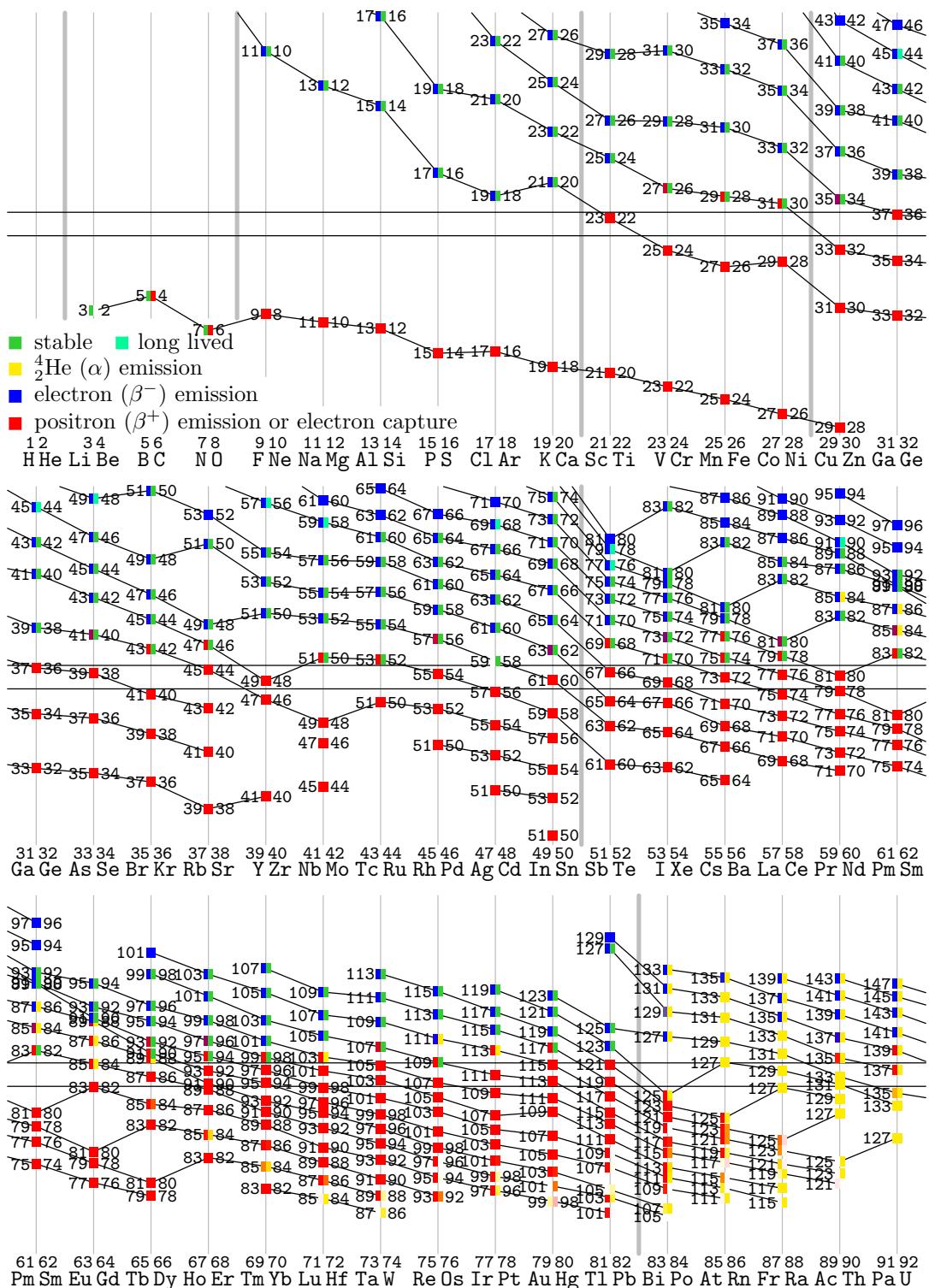


Figure 11.47: Energy release in beta decay of odd-odd nuclei.

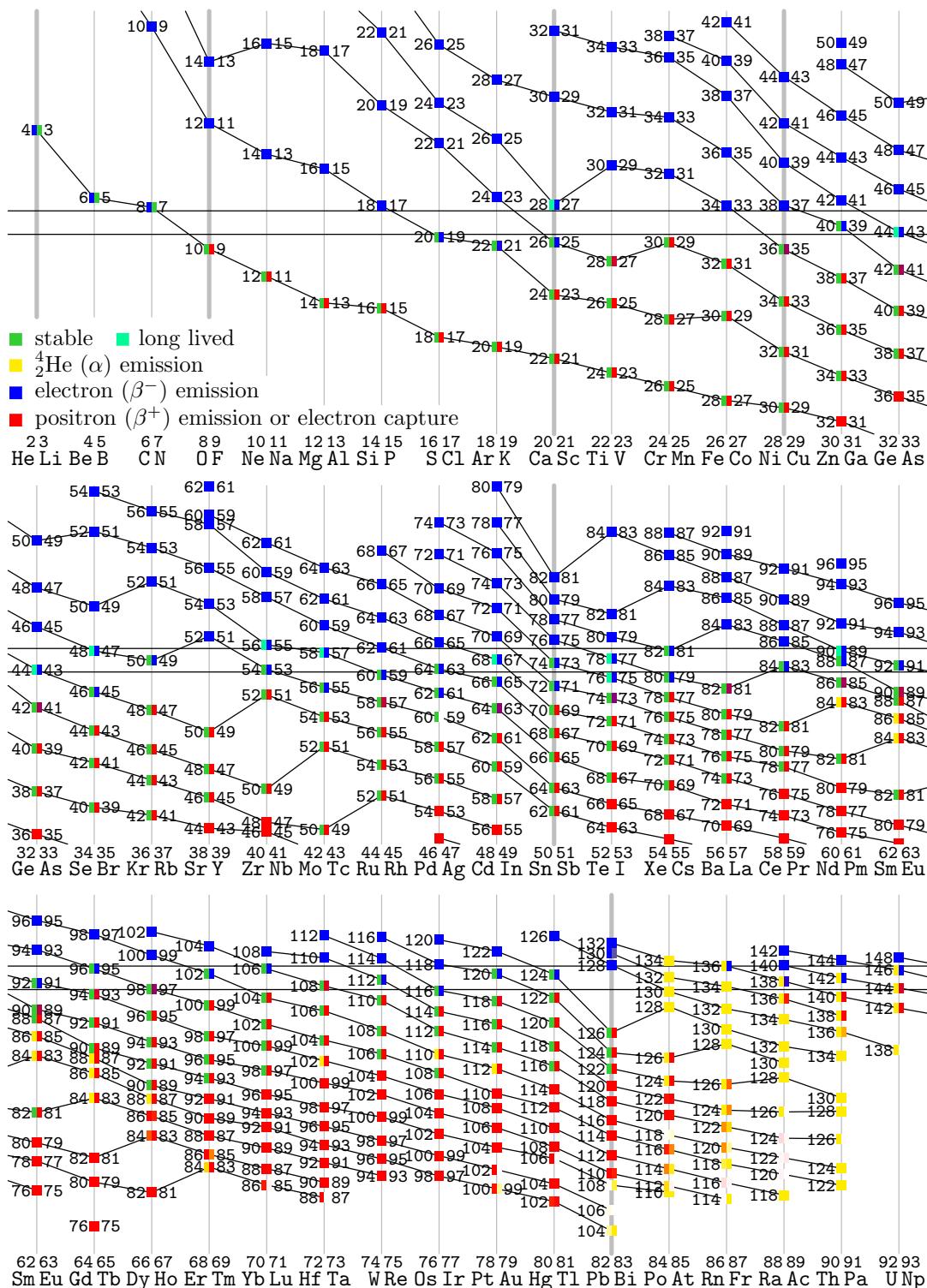


Figure 11.48: Energy release in beta decay of even-even nuclei.

each with spin $\frac{1}{2}$, and a neutrino is needed to ensure that angular momentum is conserved. However, it can do that either by being emitted or being absorbed.

Below the center bands in figures 11.46 through 11.48, both electron capture and positron emission are possible. Electron capture has still an energy reduction advantage of two electron masses over positron emission. On the other hand, the wave functions of atomic electrons are so big compared to the size of the nucleus that the electron is very unlikely to be found inside the nucleus. A high-energy positron created by the nucleus itself can have a much shorter wave length.

Note that $^{40}_{19}\text{K}$ potassium-40, with 21 neutrons, appears above the band in figure 11.47, indicating that it suffers beta decay. But it also appears below the band in figure 11.48, so that it also suffers electron capture and positron emission.

The magic neutron numbers are quite visible in the figures. For example, diagonal bands at neutron numbers 50, 82, and 126 are prominent in all four figures. Consider for example figure 11.46. 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 11.46. 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 Nobel prize acceptance lecture [[8]] 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 11.45, 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 9-20 shell into a proton in the same shell. For the later pairs, beta decay converts a neutron in the 21-28 shell to a proton in the same shell. Only for argon-39, beta decay converts a neutron in the 21-28 shell into a proton in the lower energy 9-20 shell. The lowering of the major shell releases additional energy, and the decay has enough energy to proceed.

In figures 11.45 and 11.46, 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 them it is energetically favorable to convert protons into neutrons, rather than the other way around.

11.19.2 Von Weizsäcker approximation

Since the von Weizsäcker formula of section 11.10.2 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

$$Z_{bd} = A \frac{4C_a + m_n - m_p - m_e + C_c C_z A^{-1/3}}{8C_a + 2C_c A^{2/3}} \quad (11.43)$$

where the constants C are as given in section 11.10.2. 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 11.49. 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

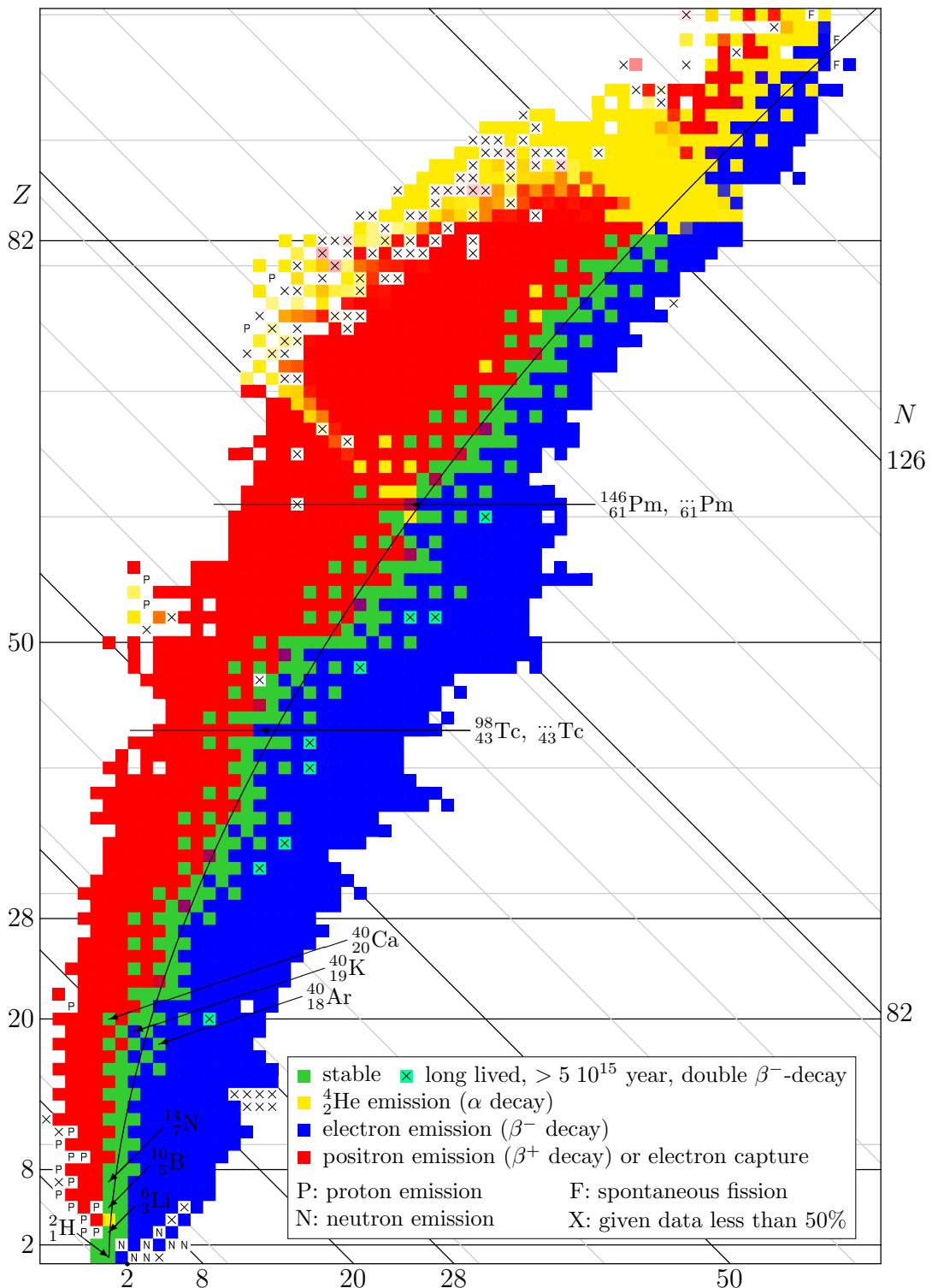


Figure 11.49: Examples of beta decay.

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 $^{180m}_{73}\text{Ta}$ 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 $^{180}_{73}\text{Ta}$ has a half life of only 8 hours, as a cooperative heavy odd-odd nucleus should.)

As an example of the instability of odd-odd nuclei, consider the curious case of potassium-40, $^{40}_{19}\text{K}$. 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 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 ($5.9 \cdot 10^{21}$ 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 11.49. 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 11.49, 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.

11.19.3 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:

$$Q = m_{N1}c^2 - m_{N2}c^2 - m_\alpha c^2 \quad (11.44)$$

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:

$$Q = \frac{1}{2}m_{N2}v_{N2}^2 + \frac{1}{2}m_\alpha v_\alpha^2$$

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:

$$m_{N1}\vec{v}_{N1} = m_{N2}\vec{v}_{N2} + m_\alpha\vec{v}_\alpha$$

but since the velocity of the initial nucleus is zero,

$$m_{N2}\vec{v}_{N2} = -m_\alpha\vec{v}_\alpha$$

Square both sides and divide by $2m_{N2}$ to get:

$$\frac{1}{2}m_{N2}v_{N2}^2 = \frac{m_\alpha}{m_{N2}}\frac{1}{2}m_\alpha v_\alpha^2$$

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

$$Q = m_{N1}c^2 - m_{N2}c^2 - m_e c^2 - m_{\bar{\nu}} c^2 \quad (11.45)$$

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 non-negligible: 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, [21, 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, {A.4} (A.7).

Consider first the extreme case that the electron comes out at rest. The relativistic energy expression gives for the kinetic energy of the neutrino:

$$T_{\bar{\nu}} = \sqrt{(m_{\bar{\nu}}c^2)^2 + (pc)^2} - m_{\bar{\nu}}c^2 \quad (11.46)$$

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:

$$T_{N2} = \frac{p^2}{2m_{N2}} = pc \frac{pc}{2m_{N2}c^2} \quad (11.47)$$

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 $T_{\bar{\nu}} = Q - T_{N2}$ 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:

$$pc = \sqrt{\frac{[E^2 - (E_{N2} + E_e)^2][E^2 - (E_{N2} - E_e)^2]}{4E^2}} \quad E = E_{N2} + E_e + Q \quad (11.48)$$

where $E_{N2} = m_{N2}c^2$ and $E_e = m_e c^2$ 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_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

$$Q = m_{N1}c^2 + m_e c^2 - m_{N2}c^2 - m_{\bar{\nu}}c^2 - E_{B,ce} \quad (11.49)$$

where $E_{B,ce}$ 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 3.2, by simply replacing e^2 in it by e^2Z . That gives:

$$E_{B,ce} \sim 13.6 Z^2 \text{ eV} \quad (11.50)$$

The ballparks for electron capture in figure 11.52 use

$$E_{B,ce} \sim \frac{1}{2}(\alpha Z)^2 m_e c^2 \left(1 + \frac{1}{4}(\alpha Z)^2\right) \quad (11.51)$$

in an attempt to partially correct for relativistic effects, which are significant for heavier nuclei. Here $\alpha \sim 1/137$ 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 (11.50). See chapter 12.1.6 for a justification.

11.19.4 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 $53 \cdot 10^{18}$ years, more than a billion times the entire lifetime of the universe. And when decay does happen, physicists are not quite sure anyway how much of it is beta decay rather than double beta decay.

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 electron 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.

Allowed decays

To see what sort of beta 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:

$$|j_{N1} - j_{e\bar{\nu}}| \leq j_{N2} \leq j_{N1} + j_{e\bar{\nu}} \quad (11.52)$$

where j_{N1} indicates the spin of the nucleus before the decay, j_{N2} the one after it, and $j_{e\bar{\nu}}$ is the combined angular momentum of electron and antineutrino. Since $j_{e\bar{\nu}} = 1$ 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

$$\boxed{\text{allowed: } |\Delta j_N| \leq 1 \quad \Delta \pi_N = \text{no}} \quad (11.53)$$

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 $1/2^+$ to $1/2^+$ 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, [21, p. 290].

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

$$\frac{pR}{\hbar} = \frac{pcR}{\hbar c} = \frac{pcR}{197 \text{ MeV fm}}$$

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.111}, 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 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:

$$\boxed{\text{first-forbidden: } |\Delta j_N| \leq 2 \quad \Delta \pi_N = \text{yes}} \quad (11.54)$$

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 are those that are not allowed ones:

$$\boxed{\text{second-forbidden: } |\Delta j_N| = 2 \text{ or } 3 \quad \Delta \pi_N = \text{no}} \quad (11.55)$$

In third forbidden decays, $l = 3$. The transitions that become possible that were not in first forbidden ones are:

$$\boxed{\text{third-forbidden: } |\Delta j_N| = 3 \text{ or } 4 \quad \Delta \pi_N = \text{yes}} \quad (11.56)$$

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:

fourth-forbidden:	$ \Delta j_N = 4 \text{ or } 5$	$\Delta\pi_N = \text{no}$	(11.57)
-------------------	----------------------------------	---------------------------	---------

This allows decay to either the 5^+ and 4^+ excited states of scandium-48. However, fourth forbidden decays are generally impossibly slow.

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 or 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.

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 11.50. 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 11.50 used a ballpark value of the mass number for each Z value, {A.111}.

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 $53 \cdot 10^{18}$ years.

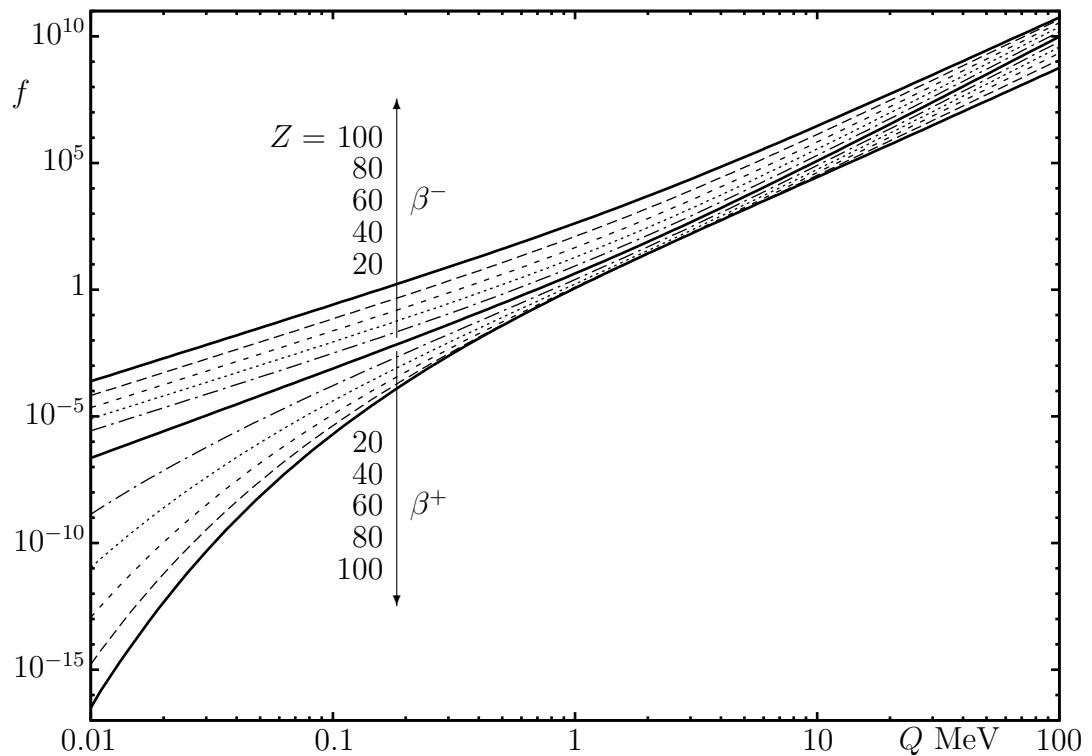


Figure 11.50: The Fermi integral. It shows the effects of energy release and nuclear charge on the beta decay rate of allowed transitions. Other effects exists.

11.19.5 Data and Fermi theory

Figure 11.51 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. 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 11.51; many would be invisible shown on the true scale.

The decay rates in figure 11.51 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. (Yellow respectively green squares indicate nuclei for which the degree to which the decay is forbidden could not be found by the automated procedures used.)

Figure 11.52 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. These are often so slow that they must be shown as minimum-size dots in figure 11.51 to be visible. However, in figure 11.52 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 $150 \cdot 10^{15}$ year, followed by cadmium-113 with $8 \cdot 10^{15}$ year, followed by indium-115 with $441 \cdot 10^{12}$ 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 11.51 become full squares in figure 11.52.

Another eye-catching success story is ${}^3_1\text{H}$, the triton, which suffers beta decay into ${}^3_2\text{He}$, the stable helion. The decay is allowed, but because of its minuscule 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 note {A.111}. 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 chapters 4.3 and 12.2.4.) The strength of the interaction is given by empirical constants.

Note that for many nuclei no ballparks were found. One major reason is that

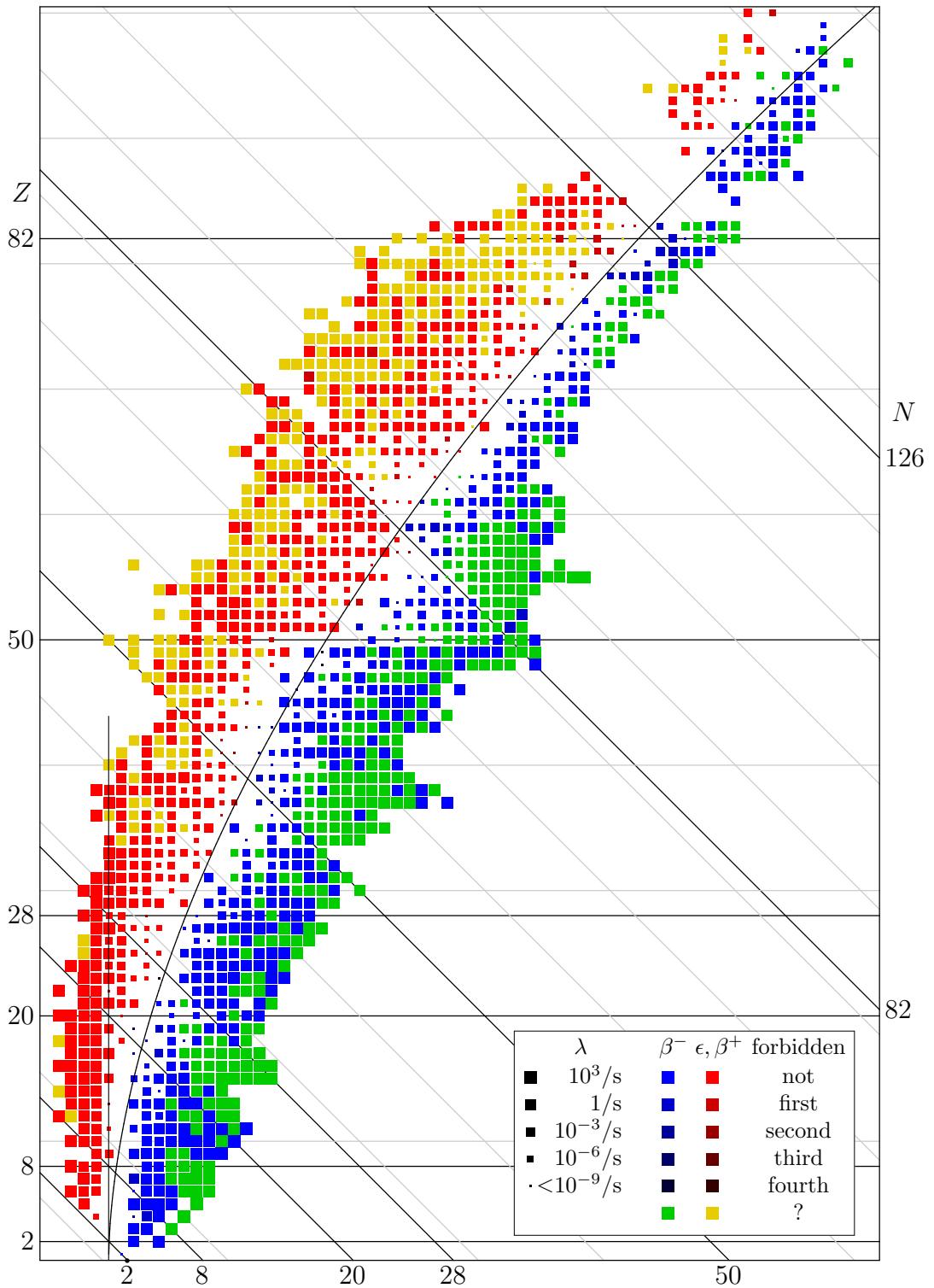


Figure 11.51: Beta decay rates.

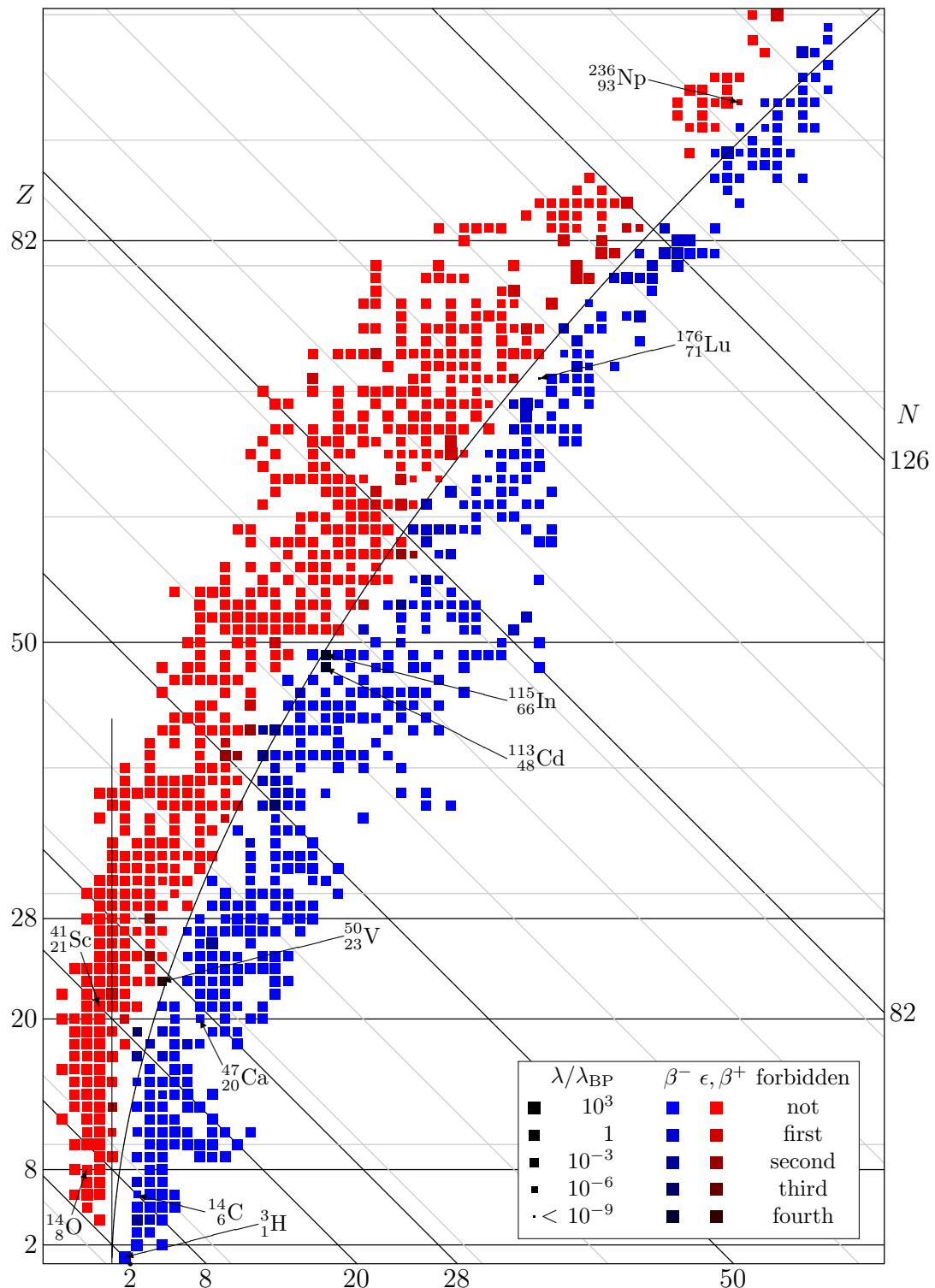


Figure 11.52: Beta decay rates as fraction of a ballparked value.

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 $^{163}_{67}\text{Ho}$, $^{193}_{78}\text{Pt}$, $^{194}_{80}\text{Hg}$, $^{202}_{82}\text{Pb}$, and $^{205}_{82}\text{Pb}$, and for the beta decay of $^{187}_{75}\text{Re}$ and $^{241}_{94}\text{Pu}$. According to the normal Fermi theory, the decay cannot occur if the Q -value is negative. To be sure, there is some energy slop, and decays with slightly negative Q -values could theoretically still happen, {A.111}. But that is irrelevant here because the Q -values in question are much smaller than the estimated electronic binding energy (11.8). 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 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 11.51 and 11.52 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 $^{253}_{100}\text{Fm}$. 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 11.51 and 11.52 are between mirror nuclei, and all are superallowed. They are full-size squares in figure 11.52. 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 4.7. 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 11.44. 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 $40 \cdot 10^{12}$ 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, [21, 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.

11.19.6 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, 6.2. 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.

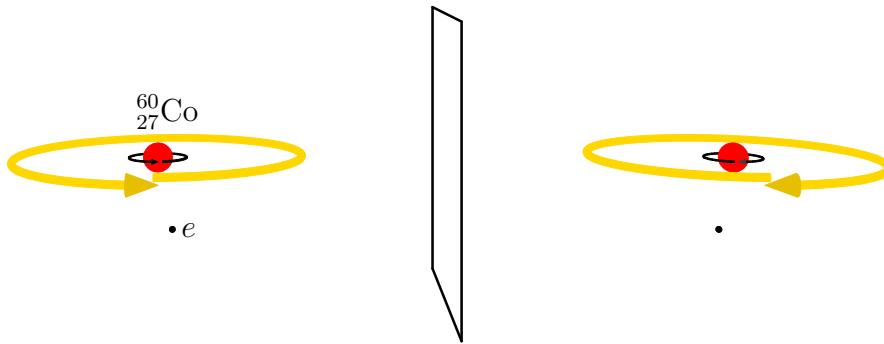


Figure 11.53: Parity violation. In the beta decay of cobalt-60, left, the electron preferentially comes out in the direction that a left-handed screw rotating with the nuclear spin would move. Seen in the mirror, right, that becomes the direction of a right-handed screw.

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 11.53. It was then observed 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} , [21, 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 11.53 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 11.53 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.

11.20 Gamma Decay

Excited nuclei can lower their energy by emitting a photon of electromagnetic radiation. That is called gamma decay.

Gamma decay of excited nuclei is the direct equivalent of the decay of excited electron states in atoms. Atomic decays were discussed in considerable detail for hydrogen in chapters 6.3 and 12.2.4. But the energy of the photons emitted by nuclei is typically even higher than the x-ray photons emitted by inner electrons. Therefor the radiation emitted by nuclei is generally referred to as “gamma rays.”

The analysis of gamma decay is also in many respect similar to that for alpha and beta decay discussed earlier in this chapter. However, the type of nucleus does not change in gamma decay, just its energy.

The existing data on gamma decay is enormous: there is a large number of stable and unstable nuclei, many have large numbers of excited energy levels, and a single excited level can often decay to multiple lower energy levels. The coverage given in this section will therefore be mostly anecdotal rather than comprehensive.

A second way that nuclei can get rid of excess energy is by kicking an atomic electron out of the surrounding atom. This process is called “internal conversion” because the electron is outside the nucleus. If the excitation energy is high, it is also possible for the nucleus to create an electron and positron pair from scratch. Since the 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.” Both “internal” processes are particularly important for transitions between nuclear states of zero spin, because such transitions cannot occur through gamma decay. They also play an important part in many transitions between nuclear states with very different spin.

11.20.1 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:

$$Q = E_{N1} - E_{N2} = pc = \hbar\omega \quad (11.58)$$

Typical tabulations list nuclear excitation energies directly, rather than as nuclear masses.

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:

$$Q = E_{N1} - E_{N2} = T_e + E_{B,e} \quad (11.59)$$

11.20.2 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, the photon normally comes out without orbital angular momentum. Since the photon has one unit of spin, the total angular momentum it carries off is one unit. That means that the nuclear spin cannot change by more than one unit.

The nuclear spin can stay the same, because in classical terms the one unit of spin can go into changing the direction of the nuclear spin instead of its magnitude. However, that only works if the nuclear spin is nonzero. Gamma decay from spin zero to spin zero is not possible.

Parity must also be preserved. Parity is even, or 1, if the wave function stays the same if \vec{r} is everywhere replaced by $-\vec{r}$. Parity is odd, or -1 , if the wave function changes sign. Unlike for angular momentum, where combined angular momentum is found by adding individual angular momenta, parities are multiplied together. In particular, in allowed decays it may be assumed that the photon has negative parity. Therefore, for the combined parity of photon and final nucleus to be equal to the parity of the initial nucleus, the parity of the nucleus must reverse during the decay. (The weak force does not preserve parity and creates a very small uncertainty in nuclear parity that will be ignored.)

Allowed transitions are called electric transitions because the nucleus interacts with the electric field of the photon. More specifically, they are called “electric dipole transitions” for reasons originating in classical electromagnetics. For practical purposes, a dipole transition is one in which the photon comes out with no orbital angular momentum.

Transitions in which the spin change is greater than one unit, or in which the parity does not change, or in which the spin stays zero, are called “forbidden.” Despite the name, such decays will occur given enough time, but they are generally much slower.

One reason that forbidden transitions can occur is that the nucleus can interact with the magnetic field of the photon instead of its electric field. This produces what are called magnetic transitions. Magnetic transitions tend to be noticeably slower than electric ones. In magnetic dipole transitions, it may be assumed that the photon has even parity, so the nuclear parity does not change. The spin can again change by up to one unit, but it cannot stay at zero.

Transitions in which the nuclear spin changes by more than one unit are possible through emission of a photon with nonzero orbital angular momentum. Classically, a photon with momentum p that comes out of a nucleus of radius R would have a ballpark momentum pR . However in quantum mechanics, an-

angular momentum is discretized in units of order \hbar . Since pR is typically only a small fraction of \hbar , the ballparked classical value of the angular momentum is impossible.

To understand what happens instead, consider transitions in which the photon comes out with one unit of orbital angular momentum. In that case the orbital azimuthal quantum number $l = 1$. Such transitions are called “quadrupole” ones. The ratio of the classical angular momentum pR to the quantum unit of angular momentum \hbar may be thought of as the quantum amplitude of the $l = 1$ state. The probability of being found in the $l = 1$ state is given by the square of the amplitude. Therefore, the probability that the photon comes out with orbital angular momentum $l = 1$ instead of zero can be ballparked as $(pR/\hbar)^2$. The decay rate for transitions with $l = 1$ is slowed down by that factor.

The factor can be expressed more conveniently by rewriting the photon momentum in terms of the nuclear energy release $Q = p/c$. In these terms, $l = 1$ decays are slowed down by a factor $(QR/\hbar c)^2$. Since the energy release has a ballpark value of a MeV, the nuclear radius a ballpark of a few fm, and $\hbar c$ is about 200 MeV fm, you see that transitions with $l = 1$ are normally orders of magnitude slower than those with $l = 0$. However, transitions between states that differ in angular momentum by two units cannot occur through dipole transitions. Quadrupole transitions will therefore dominate such decays.

Every unit increase in the value of l slows down the decays by another factor $(QR/\hbar c)^2$. (A justification for these claims can be found in {A.112}.) Therefore, transitions in which the spin changes by a large amount are normally extremely slow.

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 less than 10^{10} years, and so tantalum-180m occurs naturally.

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 11.54 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 six units of orbital angular momentum, and that just does not happen in a thousand years. Nor in a petayear.

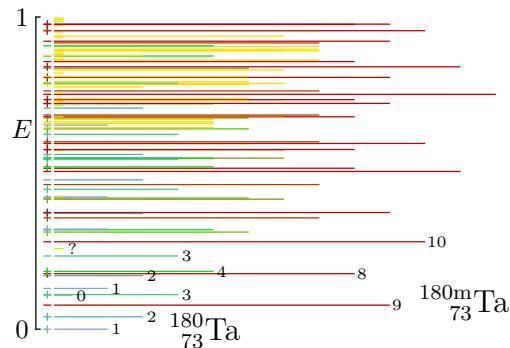


Figure 11.54: Energy levels of tantalum-180.

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 death that waits below.

Electric transitions are often generically indicated as $E\ell$ and magnetic ones as $M\ell$. Here ℓ indicates the nuclear spin change that the transition can achieve. So electric dipole transitions are $E1$, and magnetic dipole transitions are $M1$. Electric quadrupole transitions are $E2$ and magnetic ones are $M2$. And so on. For electric transitions, the nuclear parity changes when ℓ is odd. For magnetic transitions, it changes when ℓ is even.

In summary, the transition rules are:

	maximum spin change	nuclear parity change	slow down	
$E\ell :$	ℓ	if ℓ is odd	$(QR/\hbar c)^{(\ell-1)}$	(11.60)
$M\ell :$	ℓ	if ℓ is even	$(QR/\hbar c)^{(\ell-1)}$	
No spin 0 to spin 0 transitions.				

That leaves transitions from nuclear spin 0 to nuclear spin 0. Such transitions cannot occur through emission of a photon, period. A photon cannot have zero angular momentum. You might think that the intrinsic spin of the photon could be cancelled through one unit of orbital angular momentum. However, it turns out that for a photon the spin in the direction of propagation cannot be zero. Therefore it does not have a full set of three independent spin states. And that prevents it from fully cancelling its spin using a corresponding single unit of orbital angular momentum.

Another way to understand this peculiar behavior is from classical electro-

magnetics. Angular momentum is related to angular variation, and a nucleus that has zero angular momentum is spherically symmetric. Classical electrostatics says that a spherically symmetric charge distribution has the same electric field outside the nucleus regardless of the internal radial charge distribution. And if the electromagnetic field outside the nucleus does not change with time, it cannot radiate energy away.

Decay from an excited state with spin zero to another state that also has spin zero is possible through internal conversion and 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. It can be shown that in gamma decay, the quantum number of square isospin t_t should only change by up to one unit. Additional rules apply for nuclei with $T_3 = 0$, which are nuclei with the same number of protons and neutrons. For these, in E1 decays t_t cannot stay the same. M1 transitions that leave t_t the same are also weak in such nuclei.

That is illustrated by the decay of the 1^- isobaric analog state common to carbon-14, nitrogen-14, and oxygen-14 in figure 11.44. This state has $t_t = 1$. For oxygen-14, it is the lowest excited state, and its decay to the 0^+ , $t_t = 1$, ground state is an E1 transition that is allowed by the spin, parity, and isospin selection rules. And indeed, the 1^- excited state rapidly decays to the ground state; the half-life is about 0.000 12 fs (femtoseconds). That is even faster than the Weisskopf ballpark for a fully allowed decay, subsection 11.20.4, which gives about 0.007 fs here. Theoretically, it is expected that the decay rate of the 1^- state in carbon-14 is the same as for its mirror twin oxygen-14, although experimentally it has only been established that the half-life is less than 7 fs. But the $1^-, t_t = 1$, to $0^+, t_t = 1$, decay is not allowed for nitrogen-14, because nitrogen-14 has the same number of protons as neutrons, and then t_t must change. Consistent with this, the nitrogen-14 1^- state has a much larger half-life of 27 fs. And in addition, only 2% of the decays that produce that half-life are to the isobaric $0^+, t_t = 1$, state; decay is mostly to the $1^+ t_t = 0$ ground and second excited states.

11.20.3 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 [21, p. 174] ballparks the minimum lifetime to be considered a true isomeric state at roughly 10^{-9} s. However, this book will not take isomers serious unless they have a lifetime comparable to 10^{-3} 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 11.13. 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 11.55 shows the half-lives of the longest-lasting excited 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 11.56 shows the islands for odd Z , even N nuclei.

For odd-odd nuclei, figure 11.57, 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.

For even-even nuclei, figure 11.58, there is very little isomeric activity.

11.20.4 Weisskopf estimates

Gamma decay rates can be ballparked using the so-called “Weisskopf estimates,” {A.112}:

$$\lambda_{E\ell} = C_{E\ell} A^{2\ell/3} Q^{2\ell+1} \quad \lambda_{M\ell} = C_{M\ell} A^{(2\ell-2)/3} Q^{2\ell+1} \quad (11.61)$$

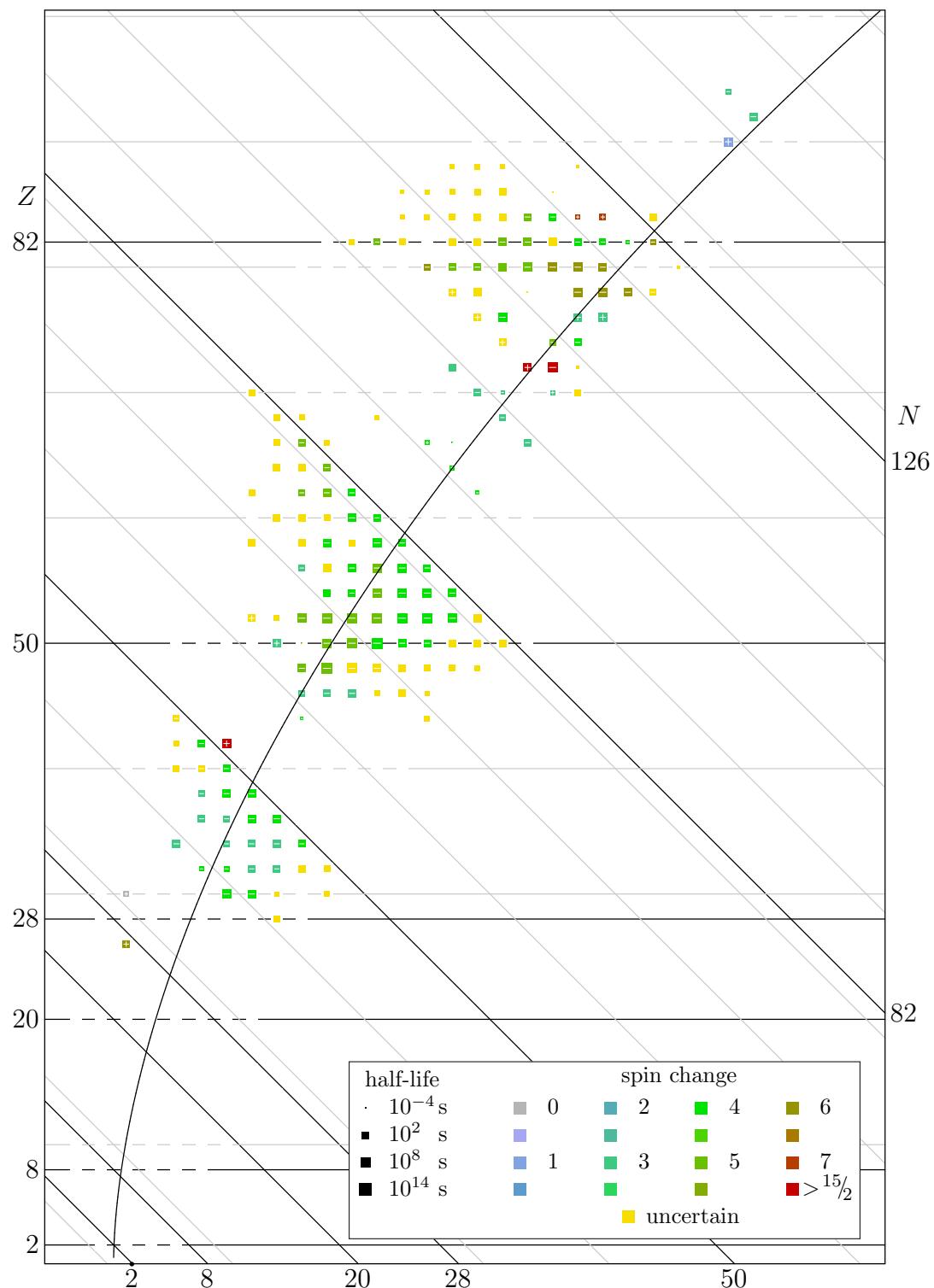


Figure 11.55: Half-life of the longest-lived even-odd isomers.

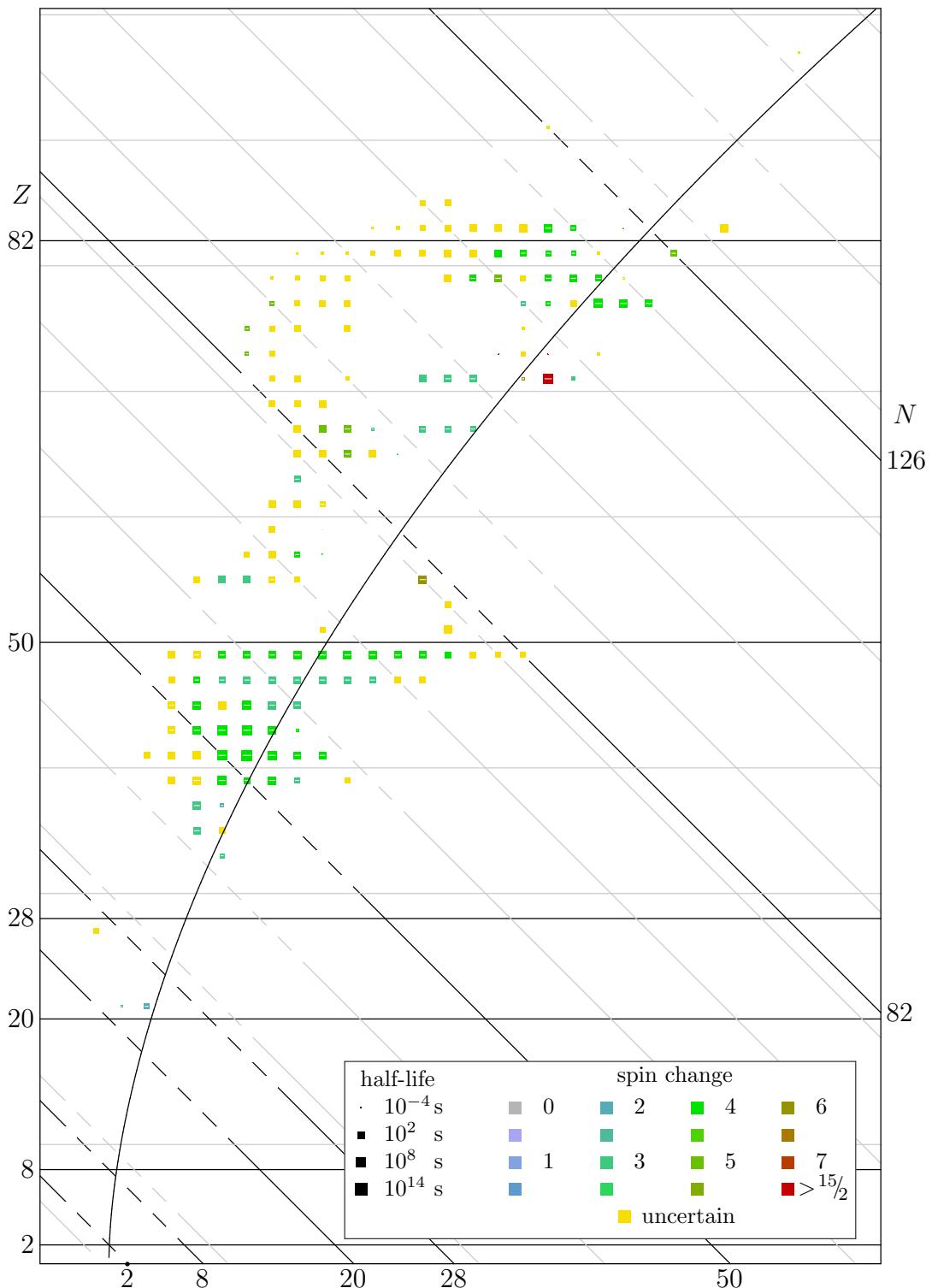


Figure 11.56: Half-life of the longest-lived odd-even isomers.

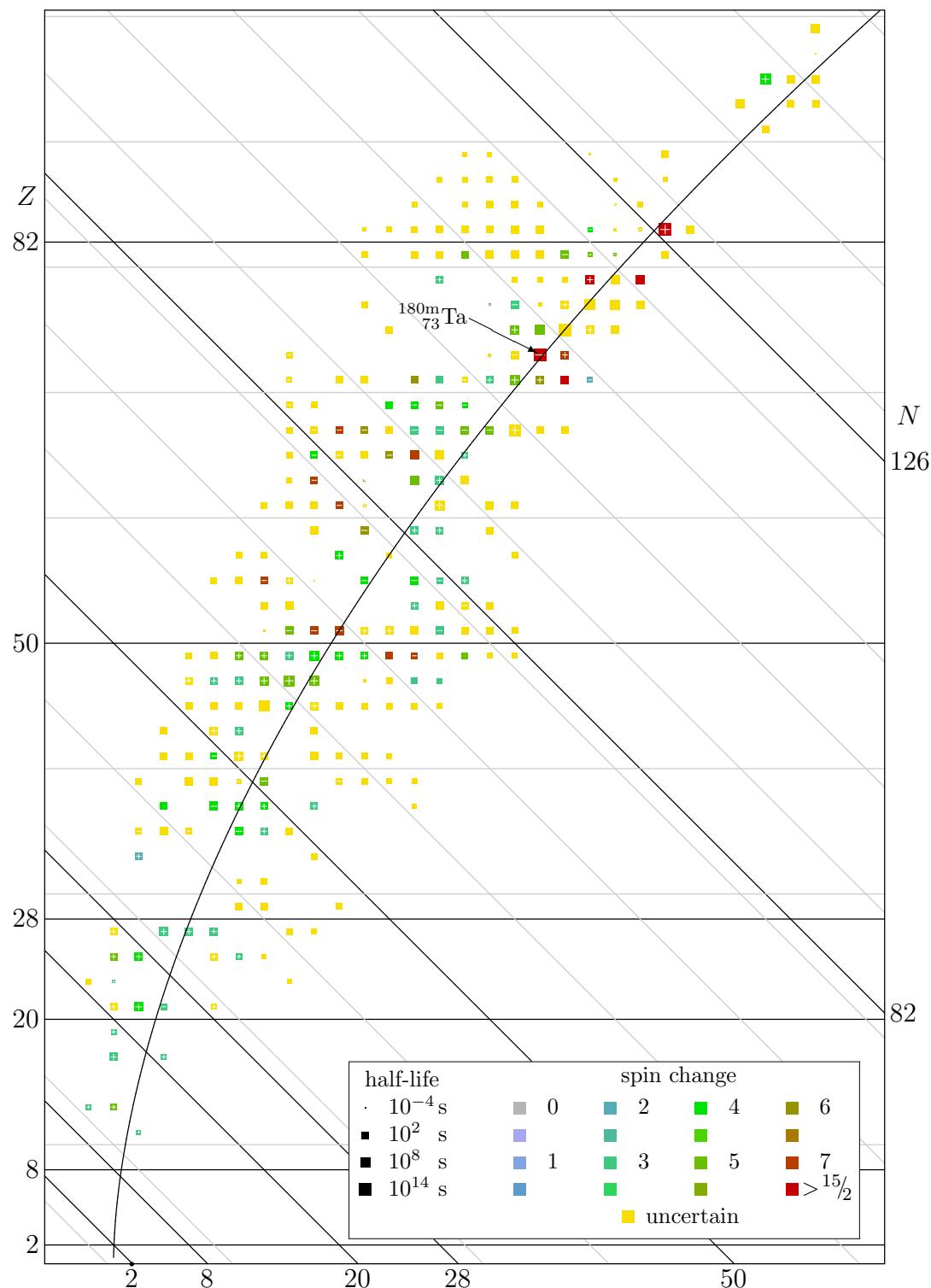


Figure 11.57: Half-life of the longest-lived odd-odd isomers.

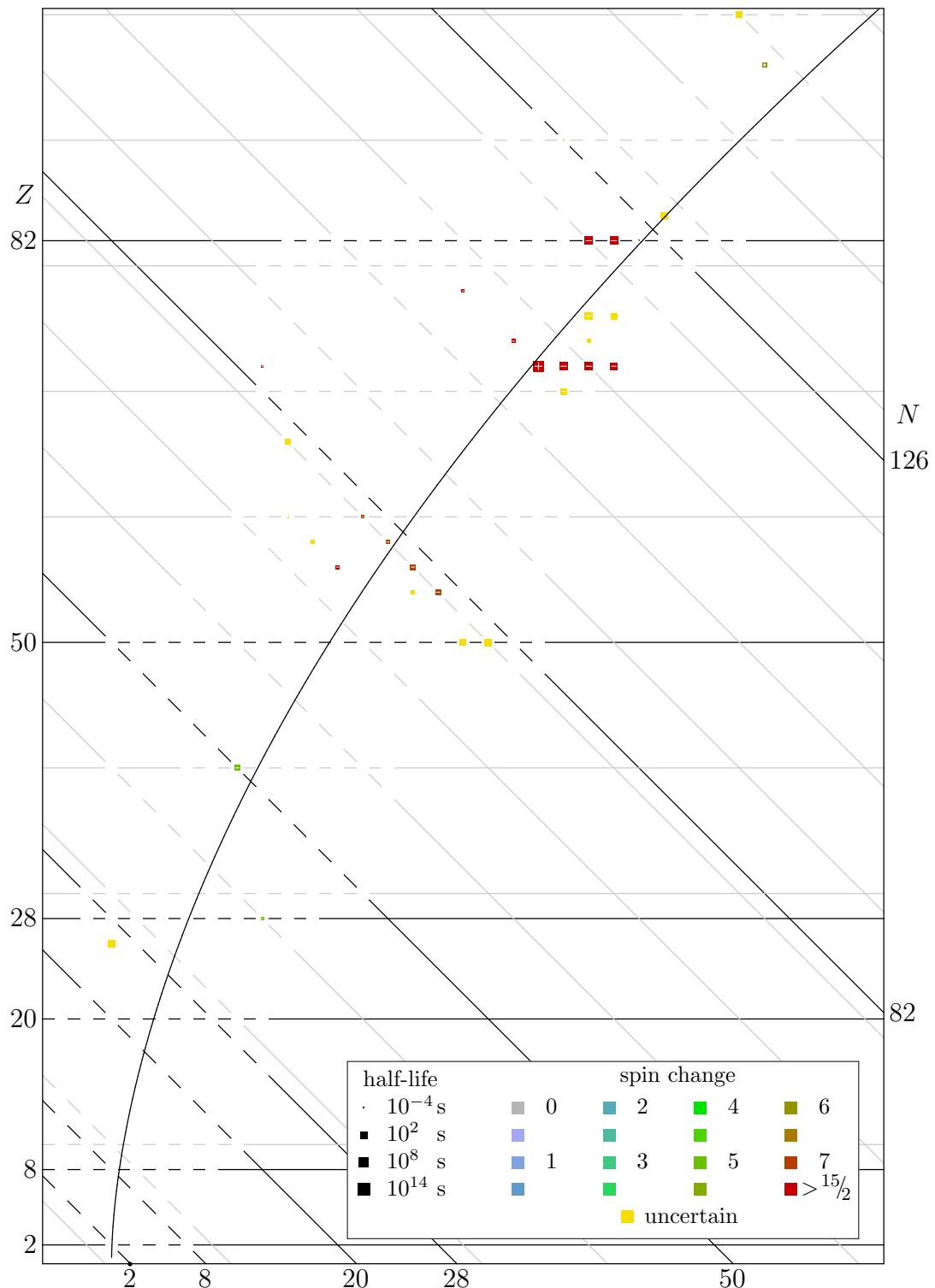


Figure 11.58: Half-life of the longest-lived even-even isomers.

$\ell :$	1	2	3	4	5
$C_{E\ell} :$	$1.0 \cdot 10^{14}$	$7.3 \cdot 10^7$	34	$1.1 \cdot 10^{-5}$	$2.4 \cdot 10^{-12}$
$C_{M\ell} :$	$3.1 \cdot 10^{13}$	$2.2 \cdot 10^7$	10	$3.3 \cdot 10^{-6}$	$7.4 \cdot 10^{-13}$

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 11.20.2, electric transitions 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 11.59. The somewhat more accurate Moszkowski estimates of magnetic transitions are shown in figure 11.60. For more details see note {A.112}. Internal conversion is discussed in the next subsection.

These estimates are very rough ballparks only. They are derived under the assumption that just a single proton is involved in the decay. If there are more nucleons involved, like for the collective motion states of deformed nuclei, the true decay rates may exceed the ballpark by one to two orders of magnitude.

The estimates also assume that the nuclear state produced by the decay Hamiltonian has a good probability of being the correct final nuclear state. This is often untrue, producing decay rate that are orders of magnitude slower than estimated.

If the data is examined in more detail, e.g. [21], it is seen that $E1$ transitions can easily be three or more orders of magnitude slower than estimate. That is similar to what was observed for the ballparks for beta decays given in section 11.19.5.

However, $E2$ transitions are often one to two orders of magnitude faster than estimated. That can be understood from the fact that $E2$ transitions include transitions between the energy levels of rotational bands in deformed nuclei. Such states involve collective motion.

Transitions with large spin changes often agree very well with the Weisskopf estimates. 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, 11.13. Transitions between these states involve a parity change and a large change in spin, leading to $E3$ and $M4$ transitions. They are single-particle transitions as the Weisskopf 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, is that these tend to heavy nuclei. For heavy nuclei the restrictions put on by isospin are less confining.

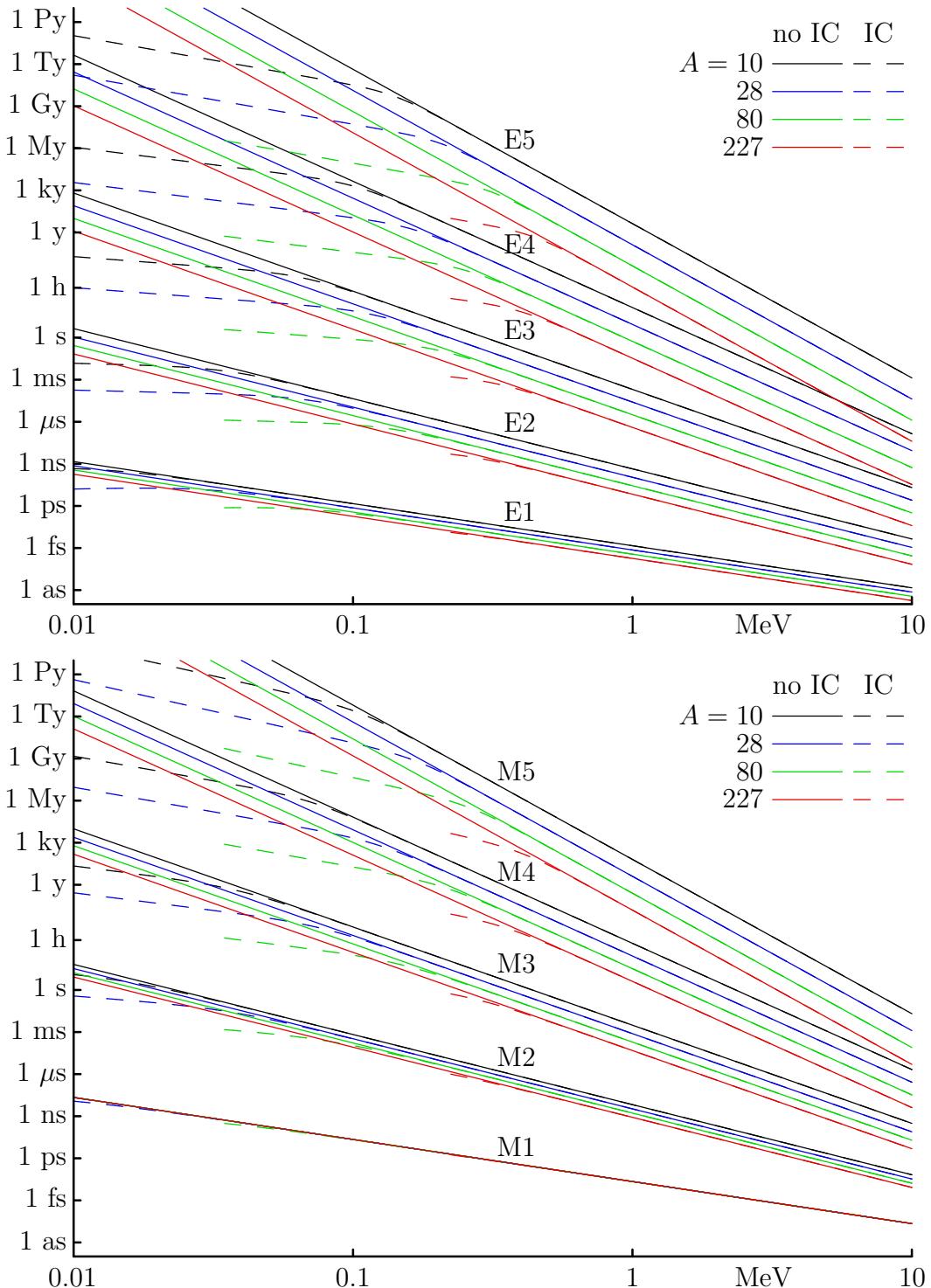


Figure 11.59: Weisskopf ballpark half-lives for electromagnetic transitions versus energy release. Broken lines include ballparked internal conversion.

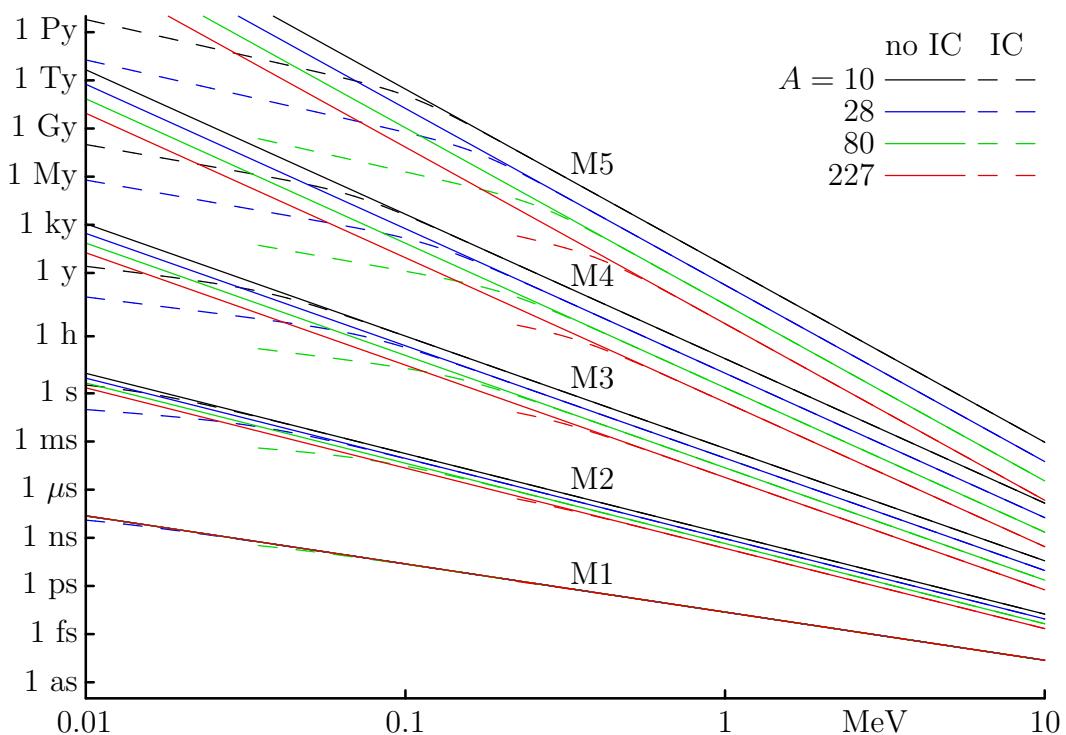


Figure 11.60: Moszkowski ballpark half-lives for magnetic transitions versus energy release. Broken lines include ballparked internal conversion.

11.20.5 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 come out with zero angular momentum. However, the ejected electron, called the “conversion electron,” can keep whatever spin it has. If the excitation energy exceeds the combined rest mass energy of an electron and positron, pair creation is another possibility. It may dominate at high excitation energies. Other possible decay processes such as two-photon or photon-conversion electron emission are normally much slower.

In transitions other than between states of zero spin, internal conversion and pair production compete with gamma decay. They are especially important in highly forbidden gamma decays. The internal conversion coefficient α_ℓ gives the internal conversion rate of a transition as a fraction of its gamma decay rate:

$$\boxed{\alpha_\ell = \frac{\lambda_{\text{IC}}}{\lambda_\gamma}} \quad (11.62)$$

The following ballpark values for the internal conversion coefficient in electric and magnetic transitions can be derived ignoring relativistic effects:

$$\alpha_{E\ell} = \alpha \frac{(\alpha Z)^3}{n^3} \left(\frac{2m_e c^2}{Q} \right)^{\ell+5/2} \frac{\ell}{\ell+1} \quad (11.63)$$

$$\alpha_{M\ell} = \alpha \frac{(\alpha Z)^3}{n^3} \left(\frac{2m_e c^2}{Q} \right)^{\ell+3/2} \quad (11.64)$$

where ℓ is the multipole order of the decay, Q the nuclear energy release, n the principal quantum number of the atomic shell that the conversion electron comes from, and $\alpha = e^2/4\pi\epsilon_0\hbar c \approx 1/137$ the fine structure constant. 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 not really justified, for one because relativistic effects can definitely not be ignored. However, they do predict a few correct trends. Internal conversion is relatively more important compared to gamma decay if the energy release of the decay Q 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.

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.

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, {A.113}).

Chapter 12

Some Additional Topics

Below are some additional topics that are intended to make the coverage of this book fairly complete. They may not be that important to most engineers, (but that depends on what you work on), or they may require a much more advanced description than can be given in a single book.

12.1 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 $H_0 + H_1$.

12.1.1 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

$$H_{\vec{n}\vec{n},1} \equiv \langle \psi_{\vec{n},0} | H_1 \psi_{\vec{n},0} \rangle \quad (12.1)$$

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 12.1.3.

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, {A.114}; just add the corresponding Hamiltonian perturbation coefficient:

$$E_{\vec{n}} = E_{\vec{n},0} + H_{\vec{n}\vec{n},1} + \dots \quad (12.2)$$

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:”

$$E_{\vec{n}} = E_{\vec{n},0} + H_{\vec{n}\vec{n},1} - \sum_{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}} \frac{|H_{\underline{\vec{n}}\vec{n},1}|^2}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} + \dots \quad (12.3)$$

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

$$\psi_{\vec{n}} = \psi_{\vec{n},0} - \sum_{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}} \frac{H_{\underline{\vec{n}},1}}{E_{\underline{\vec{n}},0} - E_{\vec{n},0}} \psi_{\vec{n},0} + \sum_{\substack{E_{\underline{\vec{n}},0} = E_{\vec{n},0} \\ \underline{\vec{n}} \neq \vec{n}}} c_{\underline{\vec{n}}} \psi_{\underline{\vec{n}},0} + \dots \quad (12.4)$$

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 note {A.114}.

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

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},1} = -(H_1 - E_{\vec{n},1})\psi_{\vec{n},0} \quad E_{\vec{n},2} = \langle \psi_{\vec{n},0} | (H_1 - E_{\vec{n},1})\psi_{\vec{n},1} \rangle \quad (12.5)$$

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

$$\frac{\partial E_{\vec{n}}}{\partial \lambda} = \left\langle \psi_{\vec{n},0} \left| \frac{\partial H}{\partial \lambda} \psi_{\vec{n},0} \right. \right\rangle \quad (12.6)$$

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, {A.118}. Of course, always make sure the eigenfunction $\psi_{\vec{n},0}$ is a good one for the derivative of the Hamiltonian.

12.1.2 Ionization energy of helium

One prominent deficiency in the approximate analysis of the heavier atoms in chapter 4.9 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

$$H_{\text{He ion}} = -\frac{\hbar^2}{2m_e} \nabla_1^2 - 2\frac{e^2}{4\pi\epsilon_0} \frac{1}{r_1}$$

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 3.2, 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,

$$a = \frac{4\pi\epsilon_0\hbar^2}{m_e 2e^2} = \frac{1}{2}a_0$$

and so its energy and wave function become

$$E_{\text{gs,ion}} = -\frac{\hbar^2}{2m_e a^2} = 4E_1 \quad \psi_{\text{gs,ion}}(\vec{r}) = \frac{1}{\sqrt{\pi a^3}} e^{-r/a}$$

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 (4.33):

$$H_{\text{He}} = -\frac{\hbar^2}{2m_e} \nabla_1^2 - 2\frac{e^2}{4\pi\epsilon_0} \frac{1}{r_1} - \frac{\hbar^2}{2m_e} \nabla_2^2 - 2\frac{e^2}{4\pi\epsilon_0} \frac{1}{r_2} + \frac{e^2}{4\pi\epsilon_0} \frac{1}{|\vec{r}_2 - \vec{r}_1|}$$

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

$$\psi_{\text{gs},0} = \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \quad E_{\text{gs},0} = 8E_1$$

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 section 4.9 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

$$E_{\text{gs}} = E_{\text{gs},0} + \langle \psi_{\text{gs},0} | H_1 \psi_{\text{gs},0} \rangle$$

or substituting in from above,

$$E_{\text{gs}} = 8E_1 + \frac{e^2}{4\pi\epsilon_0} \left\langle \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \left| \frac{1}{|\vec{r}_2 - \vec{r}_1|} \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \right. \right\rangle$$

The inner product of the final term can be written out as

$$\frac{e^2}{4\pi\epsilon_0} \frac{1}{\pi^2 a^6} \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} \frac{e^{-2(r_1+r_2)/a}}{|\vec{r}_2 - \vec{r}_1|} d^3 \vec{r}_1 d^3 \vec{r}_2$$

This integral can be done analytically. Try it, if you are so inclined; integrate $d^3 \vec{r}_1$ 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 dr_1 . Then integrate $d^3 \vec{r}_2$.

The result of the integration is

$$\frac{e^2}{4\pi\epsilon_0} \left\langle \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \left| \frac{1}{|\vec{r}_2 - \vec{r}_1|} \frac{1}{\pi a^3} e^{-(r_1+r_2)/a} \right. \right\rangle = \frac{e^2}{4\pi\epsilon_0} \frac{5}{8a} = \frac{5}{2} |E_1|$$

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 :

$$\psi_{\text{gs}} = \psi_a(\vec{r}_1)\psi_a(\vec{r}_2) \quad \psi_a(\vec{r}) \equiv \frac{1}{\sqrt{\pi a^3}} e^{-r/a}$$

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

$$E_{\text{gs}} = \langle \psi_{\text{gs}} | H_{\text{He}} \psi_{\text{gs}} \rangle$$

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 6.1.7 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: 3.2,

$$\begin{aligned} \left\langle \psi_a \left| -\frac{\hbar^2}{2m_e} \nabla^2 \right. \psi_a \right\rangle &= \frac{\hbar^2}{2m_e a^2} \\ -\frac{e^2}{4\pi\epsilon_0} \left\langle \psi_a \left| \frac{1}{r} \right. \psi_a \right\rangle &= -\frac{\hbar^2}{m_e a_0} \left\langle \psi_a \left| \frac{1}{r} \right. \psi_a \right\rangle = -\frac{\hbar^2}{m_e a_0} \frac{1}{a} \end{aligned}$$

and this subsection added

$$\left\langle \psi_a \psi_a \left| \frac{1}{|\vec{r}_2 - \vec{r}_1|} \right. \psi_a \psi_a \right\rangle = \frac{5}{8a}$$

Using these results with the helium Hamiltonian, the expectation energy of the helium atom can be written out to be

$$\langle \psi_a \psi_a | H_{\text{He}} \psi_a \psi_a \rangle = \frac{\hbar^2}{m_e a^2} - \frac{27}{8} \frac{\hbar^2}{m_e a_0 a}$$

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 $-3^6\hbar^2/2^8m_ea_0^2$, or $3^6E_1/2^7$. Putting in the numbers, the ionization energy is now found as 23.1 eV, in quite good agreement with the experimental 24.6 eV.

12.1.3 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

$$\psi_{\vec{n}_1,0}, \psi_{\vec{n}_2,0}, \dots, \psi_{\vec{n}_d,0}$$

Then as far as H_0 is concerned, any combination

$$\psi_{\vec{n},0} = c_1\psi_{\vec{n}_1,0} + c_2\psi_{\vec{n}_2,0} + \dots + c_d\psi_{\vec{n}_d,0}$$

with arbitrary coefficients c_1, c_2, \dots, 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 3.2, 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 $0 \leq l \leq n - 1$ corresponding to the square orbital angular momentum $L^2 = l(l+1)\hbar^2$; regardless of what is the magnetic quantum number $|m| \leq l$ 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 $\psi_{200}\uparrow$, $\psi_{211}\uparrow$, $\psi_{210}\uparrow$, and $\psi_{21-1}\uparrow$, and so on for higher values of n .

There are two important rules to identify the good eigenfunctions, {A.114}:

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 $\psi_{nlm}\uparrow$, the quantum numbers l , m , and m_s make the eigenfunctions at a given unperturbed energy level n unique. They correspond to the operators \hat{L}^2 , \hat{L}_z , and \hat{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 eigenfunctions

$$\psi_{\vec{n}_1,0}, \psi_{\vec{n}_2,0}, \dots$$

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, \dots of each good eigenfunction

$$c_1 \psi_{\vec{n}_1,0} + c_2 \psi_{\vec{n}_2,0} + \dots$$

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 note {A.114} 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.

12.1.4 The Zeeman effect

If you put an hydrogen atom in an external magnetic field \vec{B}_{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 10.6, the Hamiltonian of the hydrogen atom acquires an additional term

$$H_1 = \frac{e}{2m_e} B_{\text{ext}} (\hat{L}_z + 2\hat{S}_z) \quad (12.7)$$

beyond the basic hydrogen atom Hamiltonian H_0 of chapter 3.2.1. 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 $\psi_{nml}\downarrow$ energy eigenfunctions are already good ones, because H_1 commutes with all of \hat{L}^2 , \hat{L}_z and \hat{S}_z . So, according to perturbation theory, the energy eigenvalues of an hydrogen atom in a magnetic field are approximately

$$E_n + \langle \psi_{nml}\downarrow | H_1 | \psi_{nml}\downarrow \rangle = E_n + \frac{e}{2m_e} B_{\text{ext}} (m + 2m_s)\hbar$$

Actually, this is not approximate at all; it is the exact eigenvalue of $H_0 + H_1$ corresponding to the exact eigenfunction $\psi_{nml}\downarrow$.

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,

$$E_{100\downarrow} = E_1 - \frac{e\hbar}{2m_e} B_{\text{ext}} \quad E_{100\uparrow} = E_1 + \frac{e\hbar}{2m_e} B_{\text{ext}} \quad E_1 = -13.6 \text{ eV}$$

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 10.6, so it is not like the spectrum would become unrecognizable. The single spectral line of the eight $\psi_{2l m}\uparrow$ “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 3.2 due to relativistic effects, and so the theoretical lines are already split. That is discussed in subsection 12.1.6. 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.

12.1.5 The Stark effect

If an hydrogen atom is placed in an external electric field \vec{E}_{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:

$$H_1 = eE_{\text{ext}}z \quad (12.8)$$

It is much like the potential energy mgh of gravity, with the electron charge e taking the place of the mass m , E_{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 $eE_{\text{ext}}a_0$. A strong laboratory electric field might have $eE_{\text{ext}}a_0$ of the order of 0.0005 eV, [17, 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 $eE_{\text{ext}}a_0$ 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 $\langle \psi_{nlm}\uparrow | H_1 | \psi_{nlm}\uparrow \rangle$. You will then find that this inner product is zero for whatever energy eigenfunction you take:

$$\langle \psi_{nlm}\uparrow | eE_{\text{ext}}z | \psi_{nlm}\uparrow \rangle = 0 \text{ for all } n, l, m, \text{ and } m_s$$

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 \hat{L}^2 , \hat{L}_z , and S_z that make the energy eigenfunctions $\psi_{nlm}\downarrow$ unique. If $H_1 = eE_{\text{ext}}z$ commuted with them all, the $\psi_{nlm}\downarrow$ would be good eigenfunctions. Unfortunately, while z commutes with \hat{L}_z and S_z , it does not commute with \hat{L}^2 , see chapter 3.4.4. The quantum number l is bad.

Still, the two states $\psi_{100}\uparrow$ 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 $\psi_{100}\uparrow$ 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 $\psi_{211}\uparrow$ and $\psi_{21-1}\uparrow$ 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

$$\begin{pmatrix} \langle \psi_{200}\uparrow | H_1 \psi_{200}\uparrow \rangle & \langle \psi_{200}\uparrow | H_1 \psi_{210}\uparrow \rangle \\ \langle \psi_{210}\uparrow | H_1 \psi_{200}\uparrow \rangle & \langle \psi_{210}\uparrow | H_1 \psi_{210}\uparrow \rangle \end{pmatrix} \quad H_1 = eE_{\text{ext}}z$$

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, (1.16), 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 (3.18), looking up the spherical harmonics in table 3.1 and the radial functions in table

3.3, and integrating it all against $eE_{\text{ext}}z$. The resulting matrix is

$$\begin{pmatrix} 0 & -3eE_{\text{ext}}a_0 \\ -3eE_{\text{ext}}a_0 & 0 \end{pmatrix}$$

The eigenvectors of this matrix are simple enough to guess; they have either equal or opposite coefficients c_1 and c_2 :

$$\begin{pmatrix} 0 & -3eE_{\text{ext}}a_0 \\ -3eE_{\text{ext}}a_0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{pmatrix} = -3eE_{\text{ext}}a_0 \begin{pmatrix} \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} \end{pmatrix}$$

$$\begin{pmatrix} 0 & -3eE_{\text{ext}}a_0 \\ -3eE_{\text{ext}}a_0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} \end{pmatrix} = 3eE_{\text{ext}}a_0 \begin{pmatrix} \sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{2}} \end{pmatrix}$$

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 $-3eE_{\text{ext}}a_0$, and the good combination $\sqrt{\frac{1}{2}}\psi_{200}\uparrow - \sqrt{\frac{1}{2}}\psi_{210}\uparrow$ has $+3eE_{\text{ext}}a_0$. The same applies for the spin down states. It follows that to first order the $n = 2$ level splits into three, with energies $E_2 - 3eE_{\text{ext}}a_0$, E_2 , and $E_2 + 3eE_{\text{ext}}a_0$, where the value E_2 applies to the eigenfunctions $\psi_{211}\downarrow$ and $\psi_{21-1}\downarrow$ 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 section 4.11.4. Note from figure 4.13 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 $eE_{\text{ext}}a_0$. 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 (12.3), by integrating $\langle\psi_{100}\uparrow|eE_{\text{ext}}z\psi_{nl0}\uparrow\rangle$ for all values of n and l . But after refreshing your memory about the analytical expression (A.30) for the ψ_{nlm} , you might think again.

It is however possible to find the perturbation in the wave function from the alternate approach (12.5), {A.115}. In that way the second order ground state energy is found to be

$$E_{100} = E_1 - \frac{3eE_{\text{ext}}a_0}{8|E_1|} 3eE_{\text{ext}}a_0 \quad E_1 = -13.6 \text{ eV}$$

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 0.0005 eV or so for $eE_{\text{ext}}a_0$ 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 $\pm 3eE_{\text{ext}}a_0$ 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 minuscule chance of “tunneling out” of the atom through the energy barrier, chapter 6.8.2. 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.

12.1.6 The hydrogen atom fine structure

According to the description of the hydrogen atom given in chapter 3.2, all energy eigenfunctions $\psi_{nlm}\uparrow$ with the same value of n have the same energy E_n , and 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 3.2 must be corrected for a variety of relativistic effects. Before doing so, it is helpful to express the non-relativistic energy levels of that chapter in terms of the “rest mass energy” $m_e c^2$ of the electron, as follows:

$$E_n = -\frac{\alpha^2}{2n^2} m_e c^2 \quad \text{where } \alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137} \quad (12.9)$$

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, [12].

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.
- 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 3.2 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. The transition is highly forbidden in the sense of chapter 6.3, and 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 subsubsections discuss each correction in more detail.

Fine structure

From the Dirac equation, it can be seen that three terms need to be added to the nonrelativistic Hamiltonian of chapter 3.2 to correct the energy levels for relativistic effects. The three terms are worked out in note A.116. 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 $p = m_{e,v}v$, {A.4},

$$T = m_e c^2 \sqrt{1 + \frac{p^2}{m_e^2 c^2}} - m_e c^2 \quad (12.10)$$

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, [28, 22.12], to give:

$$T \approx \frac{p^2}{2m_e} - \frac{p^4}{8m_e^3 c^2} + \dots$$

The first term corresponds to the kinetic energy operator used in the non-relativistic quantum solution of chapter 3.2. (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 $\hat{\vec{p}} = \hbar \nabla / i$. So the Hamiltonian used in chapter 3.2 was a bit relativistic already, because in replacing \vec{p} by $\hbar \nabla / i$, 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

$$H_{1,\text{Einstein}} = -\frac{\alpha^2}{4|E_1|} \left(\frac{\hat{\vec{p}}^2}{2m_e} \right)^2 \quad (12.11)$$

The second correction that must be added to the non-relativistic 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 $\vec{\phi}$ in an electric field \vec{E} is

$$E_{\text{1,spin-orbit}} = -\vec{\phi} \cdot \vec{E}$$

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 $\vec{\phi}$ 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{E} of the nucleus is minus the gradient of its potential $e/4\pi\epsilon_0 r$, so

$$E_{\text{1,spin-orbit}} \propto (\vec{v} \times \vec{\mu}) \cdot \frac{e}{4\pi\epsilon_0 r^3} \vec{r}$$

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 $-e/m_e$, so

$$E_{\text{1,spin-orbit}} \propto \frac{e^2}{m_e 4\pi\epsilon_0 r^3} (\vec{r} \times \vec{v}) \cdot \vec{S}$$

The expression between the parentheses is the angular momentum \vec{L} save for the electron mass. The constant of proportionality is worked out in note {A.117}, giving the spin-orbit Hamiltonian as

$$H_{\text{1,spin-orbit}} = \alpha^2 |E_1| \left(\frac{a_0}{r} \right)^3 \frac{1}{\hbar^2} \hat{\vec{L}} \cdot \hat{\vec{S}} \quad (12.12)$$

The final correction that must be added to the non-relativistic Hamiltonian is the so-called “Darwin term.”

$$H_{\text{1,Darwin}} = \alpha^2 |E_1| \pi a_0^3 \delta^3(\vec{r}) \quad (12.13)$$

According to its derivation in note A.116, it is a crude fix-up for an interaction with a virtual positron that simply cannot be included correctly in a non-relativistic 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 non-relativistic 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 $\hbar/m_e c$. 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 non-relativistically, 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 10.13 or (10.46). 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 $\psi_{nlm}\uparrow$ are made unique by the square angular momentum operator \hat{L}^2 , giving l , the z -angular momentum operator \hat{L}_z , giving m , and the spin angular momentum operator \hat{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

$$H_{1,\text{spin-orbit}} = \alpha^2 |E_1| \left(\frac{a_0}{r}\right)^3 \frac{1}{\hbar^2} (\hat{L}_x \hat{S}_x + \hat{L}_y \hat{S}_y + \hat{L}_z \hat{S}_z)$$

The radial factor is no problem; it commutes with every orbital angular momentum component, since these are purely angular derivatives, chapter 3.1.2. It also commutes with every component of spin because all spatial functions and operators do, chapter 4.5.3. As far as the dot product is concerned, it commutes with \hat{L}^2 since all the components of \hat{L} do, chapter 3.4.4, and since all the components of $\hat{\vec{S}}$ commute with any spatial operator. But unfortunately, \hat{L}_x and \hat{L}_y do not commute with \hat{L}_z , and \hat{S}_x and \hat{S}_y do not commute with \hat{S}_z (chapters 3.4.4 and 4.5.3):

$$[\hat{L}_x, \hat{L}_z] = -i\hbar \hat{L}_y \quad [\hat{L}_y, \hat{L}_z] = i\hbar \hat{L}_x \quad [\hat{S}_x, \hat{S}_z] = -i\hbar \hat{S}_y \quad [\hat{S}_y, \hat{S}_z] = i\hbar \hat{S}_x$$

The quantum numbers m and m_s are bad.

Fortunately, $\hat{\vec{L}} \cdot \hat{\vec{S}}$ does commute with the net z -angular momentum \hat{J}_z , defined as $\hat{L}_z + \hat{S}_z$. Indeed, using the commutators above and the rules of chapter 3.4.4 to take apart commutators:

$$\begin{aligned} [\hat{L}_x \hat{S}_x, \hat{L}_z + \hat{S}_z] &= [\hat{L}_x, \hat{L}_z] \hat{S}_x + \hat{L}_x [\hat{S}_x, \hat{S}_z] = -i\hbar \hat{L}_y \hat{S}_x - i\hbar \hat{L}_x \hat{S}_y \\ [\hat{L}_y \hat{S}_y, \hat{L}_z + \hat{S}_z] &= [\hat{L}_y, \hat{L}_z] \hat{S}_y + \hat{L}_y [\hat{S}_y, \hat{S}_z] = i\hbar \hat{L}_x \hat{S}_y + i\hbar \hat{L}_y \hat{S}_x \\ [\hat{L}_z \hat{S}_z, \hat{L}_z + \hat{S}_z] &= [\hat{L}_z, \hat{L}_z] \hat{S}_z + \hat{L}_z [\hat{S}_z, \hat{S}_z] = 0 \end{aligned}$$

and adding it all up, you get $[\hat{\vec{L}} \cdot \hat{\vec{S}}, \hat{J}_z] = 0$. The same way of course $\hat{\vec{L}} \cdot \hat{\vec{S}}$ commutes with the other components of net angular momentum \hat{J} , since the z -axis is arbitrary. And if $\hat{\vec{L}} \cdot \hat{\vec{S}}$ commutes with every component of \hat{J} , then it commutes with their sum of squares \hat{J}^2 . So, eigenfunctions of \hat{L}^2 , \hat{J}^2 , and \hat{J}_z are good eigenfunctions.

Such good eigenfunctions can be constructed from the $\psi_{nlm}\uparrow$ 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 10.5. 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 3.2.1 $\hat{\vec{p}}^2$ in the Einstein term consists of radial functions and radial derivatives plus \hat{L}^2 . These commute with the angular derivatives that make up the components of \hat{L} , and as spatial functions and operators, they commute with the components of spin. So the Einstein Hamiltonian commutes with all components of \hat{L} and $\hat{J} = \hat{L} + \hat{S}$, hence with \hat{L}^2 , \hat{J}^2 , and \hat{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

$$\langle m_j l n | H_{1,\text{Einstein}} + H_{1,\text{spin-orbit}} + H_{1,\text{Darwin}} | nljm_j \rangle$$

must be found. Starting with the Einstein term, it is

$$\langle m_j l n | H_{1,\text{Einstein}} | nljm_j \rangle = -\frac{\alpha^2}{4|E_1|} \langle m_j l n | \frac{\hat{p}^4}{4m_e^2} | nljm_j \rangle$$

Unlike what you may have read elsewhere, \hat{p}^4 is indeed a Hermitian operator, but $\hat{p}^4|nljm_j\rangle$ may have a delta function at the origin, (10.46), 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 non-relativistic energy eigenvalue problem:

$$\begin{aligned}\langle m_j jln | \frac{\hat{p}^2}{2m_e} \left| \frac{\hat{p}^2}{2m_e} \right| nljm_j \rangle &= \langle m_j jln | E_n - V | E_n - V | nljm_j \rangle \\ &= \langle m_j jln | E_n^2 - 2VE_n + V^2 | nljm_j \rangle\end{aligned}$$

Noting from chapter 3.2 that $E_n = E_1/n^2$, $V = 2E_1a_0/r$ and that the expectation values of a_0/r and $(a_0/r)^2$ are given in note {A.118}, 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

$$\langle m_j jln | H_{1,\text{spin-orbit}} | nljm_j \rangle = \alpha^2 |E_1| \langle m_j jln | \left(\frac{a_0}{r} \right)^3 \frac{1}{\hbar^2} \hat{L} \cdot \hat{S} | nljm_j \rangle$$

For states with no orbital angular momentum, all components of \hat{L} produce zero, so there is no contribution. Otherwise, the dot product $\hat{L} \cdot \hat{S}$ can be rewritten by expanding

$$\hat{J}^2 = (\hat{L} + \hat{S})^2 = \hat{L}^2 + \hat{S}^2 + 2\hat{L} \cdot \hat{S}$$

to give

$$\hat{L} \cdot \hat{S} |nljm_j\rangle = \frac{1}{2} (\hat{J}^2 - \hat{L}^2 - \hat{S}^2) |nljm_j\rangle = \frac{1}{2} \hbar^2 (j(j+1) - l(l+1) - \frac{1}{2}(1+\frac{1}{2})) |nljm_j\rangle$$

That leaves only the expectation value of $(a_0/r)^3$ to be determined, and that can be found in note {A.118}. 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 6.4.1. Note {A.17}, (A.31), implies that the value of the wave functions at the origin is zero

unless $l = 0$, and then the value is given in (A.32). So the Darwin contribution becomes

$$\langle m_j j l n | H_{1,\text{Darwin}} | n l j m_j \rangle = \frac{\alpha^2}{4n^2} 4n |E_n| \quad \text{if } l = 0$$

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:

$$E_{nljm_j,1} = -\left(\frac{1}{n(j + \frac{1}{2})} - \frac{3}{4} \frac{1}{n^2}\right) \alpha^2 |E_n|$$

(12.14)

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.

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 subsubsection are degenerate with respect to l and m_j .

The Zeeman Hamiltonian

$$H_1 = \frac{e}{2m_e} B_{\text{ext}} (\hat{L}_z + 2\hat{S}_z)$$

commutes with both \hat{L}^2 and $\hat{J}_z = \hat{S}_z + \hat{L}_z$, so the eigenfunctions $|nljm_j\rangle$ are good. Therefore, the energy perturbations can be found as

$$\frac{e}{2m_e} B_{\text{ext}} \langle m_j j l n | \hat{L}_z + 2\hat{S}_z | n l j m_j \rangle$$

To evaluate this rigorously would require that the $|nljm_j\rangle$ state be converted into the one or two $\psi_{nlm} \uparrow$ states with $-l \leq m = m_j \pm \frac{1}{2} \leq l$ and $m_s = \mp \frac{1}{2}$ using the appropriate Clebsch-Gordan coefficients from figure 10.5.

However, the following simplistic derivation is usually given instead, including in this book. First get rid of L_z by replacing it by $\hat{J}_z - \hat{S}_z$. The inner product with \hat{J}_z can then be evaluated as being $m_j\hbar$, giving the energy change as

$$\frac{e}{2m_e} B_{\text{ext}} \left[m_j\hbar + \langle m_j j | \hat{S}_z | nljm_j \rangle \right]$$

For the final inner product, make a semi-classical argument that only the component of $\hat{\vec{S}}$ 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 \hat{L} and $\hat{\vec{S}}$ must add up to \hat{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

$$\hat{S}_J = \frac{1}{J^2} (\hat{\vec{S}} \cdot \hat{\vec{J}}) \hat{J}$$

and the dot product in it can be found from expanding

$$\hat{L}^2 = \hat{\vec{L}} \cdot \hat{\vec{L}} = (\hat{\vec{J}} - \hat{\vec{S}}) \cdot (\hat{\vec{J}} - \hat{\vec{S}}) = J^2 - 2\hat{\vec{J}} \cdot \hat{\vec{S}} + S^2$$

to give

$$\hat{S}_J = \frac{J^2 - \hat{L}^2 + S^2}{2J^2} \hat{J}$$

For a given eigenfunction $|nljm_j\rangle$, $J^2 = \hbar^2 j(j+1)$, $\hat{L}^2 = \hbar^2 l(l+1)$, and $S^2 = \hbar^2 s(s+1)$ with $s = \frac{1}{2}$.

If the z -component of \hat{S}_J is substituted for \hat{S}_z in the expression for the Hamiltonian perturbation coefficients, the energy changes are

$$\left[1 + \frac{j(j+1) - l(l+1) + s(s+1)}{2j(j+1)} \right] \frac{e\hbar}{2m_e} B_{\text{ext}} m_j \quad (12.15)$$

(Rigorous analysis using figure 10.5, or more generally item 2 in chapter 10.1.7, 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 \hat{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 $|21\frac{3}{2}\frac{1}{2}\rangle$ and $|21\frac{1}{2}\frac{1}{2}\rangle$ 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 subsection 12.1.3, to find the good combinations and their energy changes. And the same for the $|21\frac{3}{2}-\frac{1}{2}\rangle$ and $|21\frac{1}{2}-\frac{1}{2}\rangle$ eigenfunctions. To obtain the matrix coefficients, use the Clebsch-Gordan coefficients from figure 10.5 to evaluate the effect of the Zeeman part. The fine structure contributions to the matrices are given by (12.14) 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.

Lamb shift

A famous experiment by Lamb & Rutherford 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 non technical level is given by Feynman [12].

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

$$E_{\vec{n},1,\text{Lamb}} = -\frac{\alpha^3}{2n} k(n, 0) E_n \quad \text{if } l = 0 \quad (12.16)$$

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,

$$E_{\vec{n},1,\text{Lamb}} = -\frac{\alpha^3}{2n} \left[k(n, l) \pm \frac{1}{\pi(j + \frac{1}{2})(l + \frac{1}{2})} \right] E_n \quad \text{if } l \neq 0 \text{ and } j = l \pm \frac{1}{2} \quad (12.17)$$

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 $n = 0$, 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, (A.31). 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 .

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

$$\vec{\mu}_p = \frac{g_p e}{2m_p} \hat{\vec{S}}_p \quad \vec{\mu}_e = -\frac{g_e e}{2m_e} \hat{\vec{S}}_e$$

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 $-\vec{\mu}_e \cdot \vec{B}_p$ of the electron in the magnetic field \vec{B}_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 10.4. The perturbation Hamiltonian then becomes

$$H_{1,\text{spin-spin}} = \frac{g_p g_e e^2}{4m_e m_p \epsilon_0 c^2} \left[\frac{3(\hat{\vec{S}}_p \cdot \vec{r})(\hat{\vec{S}}_e \cdot \vec{r}) - (\hat{\vec{S}}_p \cdot \hat{\vec{S}}_e)r^2}{4\pi r^5} + \frac{2(\hat{\vec{S}}_p \cdot \hat{\vec{S}}_e)}{3} \delta^3(\vec{r}) \right]$$

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 4.5.6:

$$\text{triplet: } |\psi_{100}|1\ 1\rangle \quad |\psi_{100}|1\ 0\rangle \quad |\psi_{100}|1\ -1\rangle \quad \text{singlet: } |\psi_{100}|0\ 0\rangle$$

or $|\psi_{100}|s_{\text{net}}m_{\text{net}}\rangle$ for short, where s_{net} and m_{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

$$\langle \underline{m}_{\text{net}} \underline{s}_{\text{net}} | \psi_{100} | H_{1,\text{spin-spin}} | \psi_{100} | s_{\text{net}} m_{\text{net}} \rangle$$

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 $3\hat{S}_{p,x}\hat{S}_{e,y}xy$ 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 $3\hat{S}_{p,x}\hat{S}_{e,x}x^2 - \hat{S}_{p,x}\hat{S}_{e,x}r^2$ 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

$$\frac{g_p g_e e^2}{6m_e m_p \epsilon_0 c^2} \langle \underline{m}_{\text{net}} \underline{s}_{\text{net}} | \psi_{100} | (\hat{\vec{S}}_p \cdot \hat{\vec{S}}_e) \delta^3(\vec{r}) | \psi_{100} | s_{\text{net}} m_{\text{net}} \rangle$$

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. The dot product of the spins can be found by expanding

$$\hat{\vec{S}}_{\text{net}}^2 = (\hat{\vec{S}}_p + \hat{\vec{S}}_e) \cdot (\hat{\vec{S}}_p + \hat{\vec{S}}_e) = \hat{S}_p^2 + 2\hat{S}_p \cdot \hat{S}_e + \hat{S}_e^2$$

to give

$$\hat{\vec{S}}_p \cdot \hat{\vec{S}}_e = \frac{1}{2} (\hat{S}_{\text{net}}^2 - \hat{S}_p^2 - \hat{S}_e^2)$$

The spin states $|s_{\text{net}}m_{\text{net}}\rangle$ are eigenvectors of this operator,

$$\hat{\vec{S}}_p \cdot \hat{\vec{S}}_e |s_{\text{net}}m_{\text{net}}\rangle = \frac{1}{2}\hbar^2 (s_{\text{net}}(s_{\text{net}}+1) - s_p(s_p+1) - s_e(s_e+1)) |s_{\text{net}}m_{\text{net}}\rangle$$

where both proton and electron have spin $s_p = s_e = \frac{1}{2}$. Since the triplet and singlet spin states are orthonormal, only the Hamiltonian perturbation coefficients for which $\underline{s}_{\text{net}} = s_{\text{net}}$ and $\underline{m}_{\text{net}} = m_{\text{net}}$ 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:

$$\text{triplet: } E_{1,\text{spin-spin}} = \frac{1}{3} g_p g_e \frac{m_e}{m_p} \alpha^2 |E_1| \quad \text{singlet: } E_{1,\text{spin-spin}} = -g_p g_e \frac{m_e}{m_p} \alpha^2 |E_1| \quad (12.18)$$

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_e/m_p , this will be a very low energy photon. Its wave length is as big as 0.21 m, producing the 21 cm hydrogen line.

12.2 Quantum Field Theory in a Nanoshell

The “classical” quantum theory discussed in this book has major difficulties describing really relativistic effects such as particle creation and destruction. Einstein’s $E = mc^2$ allows particles to be destroyed as long as their mass times the square speed of light shows up as energy elsewhere. They can also be created when enough energy is available. Indeed, as the Dirac equation, section 10.2, first showed, electrons and positrons can annihilate one another, or they can be created by a very energetic photon near a heavy nucleus.

And quantum field theory is not just for esoteric conditions. The photons of light are routinely created under normal conditions. Still more basic to an engineer, so are their equivalents in solids, the phonons. Then there is the band theory of solids: electrons are “created” within the conduction band, if they pick up enough energy, or “annihilated” when they lose it. And similarly for the real-life equivalent of positrons, holes in the valence band.

Such phenomena are routinely described within the framework of quantum field theory, and almost unavoidably you will run into it in literature, [13, 19]. Electron-phonon interactions are particularly important for engineering applications, leading to electrical resistance (along with crystal defects and impurities), and the combination of electrons into Cooper pairs that act as bosons and so give rise to superconductivity. The intention of this section is to explain enough of the ideas so that you can recognize it when you see it. What to do about it after you recognize it is another matter.

Especially the relativistic applications are very involved. To explain quantum field theory in a nutshell takes 500 pages, [35]. You will also need to pick up linear algebra, tensor algebra, and group theory. However, if you are just interested in relativistic quantum mechanics from an intellectual point of view,

rather than for practical applications, the answer is all good. Feynman gave a set of lectures on “quantum electrodynamics” for a general audience around 1983, and the text is readily available at low cost. Here, freed from the constraint of his lecture notes to cover a standard fare of material, Feynman truly gets it right. 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, technology 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, [12], 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. In the absence of 1 000 pages and a willing author, the following discussion will truly be quantum field theory in a nanoshell. I thank Wikipedia for the basic approach. As far as the rest is concerned, it has been pieced together from, in order of importance, [[18]], [34], [[3]], [13, 19]. Any mistakes in doing so are mine.

12.2.1 Occupation numbers

Consider once more systems of weakly interacting particles like the ones that were studied in section 9.2. The energy eigenfunctions of such a system can be written in terms of whatever are the single-particle energy eigenfunctions $\psi_1^p(\vec{r}, S_z), \psi_2^p(\vec{r}, S_z), \dots$. A completely arbitrary example of such a system eigenfunction for a system of $I = 36$ distinguishable particles is:

$$\psi_q^S = \psi_{24}^p(\vec{r}_1, S_{z1})\psi_4^p(\vec{r}_2, S_{z2})\psi_7^p(\vec{r}_3, S_{z3})\psi_1^p(\vec{r}_4, S_{z4})\psi_6^p(\vec{r}_5, S_{z5}) \dots \psi_{54}^p(\vec{r}_{36}, S_{z36}) \quad (12.19)$$

This system eigenfunction has an energy that is the sum of the 36 single-particle eigenstate energies involved:

$$E_q^S = E_{24}^p + E_4^p + E_7^p + E_1^p + E_6^p + \dots + E_{54}^p$$

Instead of writing out the example eigenfunction mathematically as done in (12.19), it can be graphically depicted as in figure 12.1. 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_{24}^p box, particle 2 is inside the ψ_4^p one, etcetera. It is just the reverse from the mathematical expression (12.19): the mathematical expression shows for each particle in turn what the single-particle eigenstate of that particle is.

E_8^P	ψ_{60}^P	ψ_{61}^P	ψ_{62}^P ⑨	ψ_{63}^P	ψ_{64}^P	ψ_{65}^P	ψ_{66}^P	ψ_{67}^P	ψ_{68}^P	ψ_{69}^P	ψ_{70}^P	ψ_{71}^P	ψ_{72}^P	ψ_{73}^P
E_7^P	ψ_{47}^P	ψ_{48}^P	ψ_{49}^P	ψ_{50}^P	ψ_{51}^P	ψ_{52}^P	ψ_{53}^P	ψ_{54}^P ⑩	ψ_{55}^P	ψ_{56}^P	ψ_{57}^P	ψ_{58}^P	ψ_{59}^P	
E_6^P	ψ_{35}^P	ψ_{36}^P	ψ_{37}^P	ψ_{38}^P	ψ_{39}^P	ψ_{40}^P	ψ_{41}^P	ψ_{42}^P	ψ_{43}^P	ψ_{44}^P	ψ_{45}^P	ψ_{46}^P ⑪ ⑫		
E_5^P	ψ_{24}^P ⑬	ψ_{25}^P	ψ_{26}^P	ψ_{27}^P	ψ_{28}^P	ψ_{29}^P	ψ_{30}^P	ψ_{31}^P ⑭	ψ_{32}^P	ψ_{33}^P	ψ_{34}^P ⑮			
E_4^P	ψ_{15}^P	ψ_{16}^P ⑯	ψ_{17}^P ⑰	ψ_{18}^P ⑱	ψ_{19}^P ⑲	ψ_{20}^P ⑳	ψ_{21}^P	ψ_{22}^P	ψ_{23}^P					
E_3^P	ψ_8^P ⑳	ψ_9^P ⑳	ψ_{10}^P	ψ_{11}^P	ψ_{12}^P ⑳ ⑳	ψ_{13}^P ⑳	ψ_{14}^P ⑳ ⑳							
E_2^P	ψ_3^P ⑳	ψ_4^P ⑳ ⑳ ⑳	ψ_5^P ⑳ ⑳ ⑳	ψ_6^P ⑳ ⑳	ψ_7^P ⑳ ⑳									
E_1^P	ψ_1^P ⑳ ⑳ ⑳ ⑳ ⑳ ⑳	ψ_2^P ⑳ ⑳ ⑳ ⑳ ⑳ ⑳												

Figure 12.1: Graphical depiction of an arbitrary system energy eigenfunction for 36 distinguishable particles.

E_8^P	ψ_{60}^P	ψ_{61}^P	ψ_{62}^P ○	ψ_{63}^P	ψ_{64}^P	ψ_{65}^P	ψ_{66}^P	ψ_{67}^P	ψ_{68}^P	ψ_{69}^P	ψ_{70}^P	ψ_{71}^P	ψ_{72}^P	ψ_{73}^P
E_7^P	ψ_{47}^P	ψ_{48}^P	ψ_{49}^P	ψ_{50}^P	ψ_{51}^P	ψ_{52}^P	ψ_{53}^P	ψ_{54}^P ○	ψ_{55}^P	ψ_{56}^P	ψ_{57}^P	ψ_{58}^P	ψ_{59}^P	
E_6^P	ψ_{35}^P	ψ_{36}^P	ψ_{37}^P	ψ_{38}^P	ψ_{39}^P	ψ_{40}^P	ψ_{41}^P	ψ_{42}^P	ψ_{43}^P	ψ_{44}^P	ψ_{45}^P ○ ○	ψ_{46}^P ○ ○		
E_5^P	ψ_{24}^P ○	ψ_{25}^P	ψ_{26}^P	ψ_{27}^P	ψ_{28}^P	ψ_{29}^P	ψ_{30}^P	ψ_{31}^P ○	ψ_{32}^P	ψ_{33}^P	ψ_{34}^P ○			
E_4^P	ψ_{15}^P	ψ_{16}^P ○	ψ_{17}^P ○	ψ_{18}^P ○	ψ_{19}^P ○	ψ_{20}^P ○	ψ_{21}^P	ψ_{22}^P	ψ_{23}^P					
E_3^P	ψ_8^P ○	ψ_9^P ○	ψ_{10}^P	ψ_{11}^P	ψ_{12}^P ○ ○	ψ_{13}^P ○	ψ_{14}^P ○ ○							
E_2^P	ψ_3^P ○	ψ_4^P ○ ○ ○	ψ_5^P ○ ○	ψ_6^P ○ ○	ψ_7^P ○ ○									
E_1^P	ψ_1^P ○ ○ ○ ○ ○ ○	ψ_2^P ○ ○ ○ ○ ○ ○												

Figure 12.2: Graphical depiction of an arbitrary system energy eigenfunction for 36 identical bosons.

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 (12.19) and corresponding depiction figure 12.1 is unacceptable. Eigenfunctions for bosons must be unchanged if two particles are swapped. As chapter 4.7 explained, in terms of the mathematical expression (12.19) it means that all wave functions that can be obtained from (12.19) by swapping particle numbers must be combined together equally into a *single* wave function. There is no way to actually list such a massive mathematical expression here. It is much easier in terms of the graphical depiction figure 12.1: 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; the numbers 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 12.2.

E_8^P	ψ_{60}^P	ψ_{61}^P	ψ_{62}^P	ψ_{63}^P	ψ_{64}^P	ψ_{65}^P	ψ_{66}^P	ψ_{67}^P	ψ_{68}^P	ψ_{69}^P	ψ_{70}^P	ψ_{71}^P	ψ_{72}^P	ψ_{73}^P
E_7^P	ψ_{47}^P	ψ_{48}^P	ψ_{49}^P	ψ_{50}^P	ψ_{51}^P	ψ_{52}^P	ψ_{53}^P	ψ_{54}^P	ψ_{55}^P	ψ_{56}^P	ψ_{57}^P	ψ_{58}^P	ψ_{59}^P	
E_6^P	ψ_{35}^P	ψ_{36}^P	ψ_{37}^P	ψ_{38}^P	ψ_{39}^P	ψ_{40}^P	ψ_{41}^P	ψ_{42}^P	ψ_{43}^P	ψ_{44}^P	ψ_{45}^P	ψ_{46}^P		
E_5^P	ψ_{24}^P	ψ_{25}^P	ψ_{26}^P	ψ_{27}^P	ψ_{28}^P	ψ_{29}^P	ψ_{30}^P	ψ_{31}^P	ψ_{32}^P	ψ_{33}^P	ψ_{34}^P			
E_4^P	ψ_{15}^P	ψ_{16}^P	ψ_{17}^P	ψ_{18}^P	ψ_{19}^P	ψ_{20}^P	ψ_{21}^P	ψ_{22}^P	ψ_{23}^P					
E_3^P	ψ_8^P	ψ_9^P	ψ_{10}^P	ψ_{11}^P	ψ_{12}^P	ψ_{13}^P	ψ_{14}^P							
E_2^P	ψ_3^P	ψ_4^P	ψ_5^P	ψ_6^P	ψ_7^P									
E_1^P	ψ_1^P	ψ_2^P												

Figure 12.3: Graphical depiction of an arbitrary system energy eigenfunction for 33 identical fermions.

For a system of identical fermions, (like electrons, protons, or neutrons,) the eigenfunctions must change sign if two particles are swapped. As chapter 4.7 showed, that means that you cannot create an eigenfunction for a system of 36 fermions from the example eigenfunction (12.19) and the swapped versions of it. Various single-particle eigenfunctions appear multiple times in (12.19), like ψ_4^P , which is occupied by particles 2, 31, and 33. A system eigenfunction for 36 identical fermions requires 36 different single-particle eigenfunctions. Graph-

ically, the example figure 12.2, which is fine for a system of identical bosons, is completely unacceptable for a system of identical fermions; there cannot be more than one fermion in a given type of single-particle state. A depiction of an arbitrary energy eigenfunction that is acceptable for a system of 33 identical fermions is in figure 12.3.

As explained in chapter 4.7, a neat way of writing down the system energy eigenfunction of the pictured example is to form a Slater determinant from the “occupied states”

$$\psi_1^P, \psi_2^P, \psi_3^P, \dots, \psi_{43}^P, \psi_{45}^P, \psi_{56}^P.$$

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 argument list of your wave function, where the electron and positron used to be? 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 arguments for the new particles into your mathematical wave function? Scribble in another row and column in the Slater determinant for your electrons? That is voodoo mathematics.

And if positrons are too weird for you, consider photons, the particles of electromagnetic radiation, like ordinary light. As chapter 9.14.5 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 arguments for the photons into your wave function when your black box heats up. Then there are solids; as section 9.14.6 showed, 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 (12.19) or the Slater determinants. Instead, the graphical depictions, such as the examples in figures 12.2 and 12.3, 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, by listing the single-state “occupation numbers.”

Consider the example bosonic eigenfunction of figure 12.2. The occupation numbers for that state would be

$$\vec{i} = |3, 4, 1, 3, 2, 2, 2, 1, 1, 0, 0, 2, 1, 2, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, \dots\rangle$$

indicating that there are 3 bosons in single-particle state ψ_1^P , 4 in ψ_2^P , 1 in ψ_3^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 12.3 would be

$$\vec{\imath} = |1, \dots\rangle$$

Such sets of occupation numbers are called “Fock states.” Each describes one system energy eigenfunction.

The most general wave function for a set of I particles is a linear combination of all the Fock states whose occupation numbers add up to I . In relativistic applications like photons in a box, there is no constraint on the number of particles and all states are possible. The set of all possible wave functions that can be formed from linear combinations of all the Fock states regardless of number of particles is called the “Fock space.”

How about the case of distinguishable particles as in figure 12.1? 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 one-particle excitation at that energy state; no particle numbers are involved.

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 are an additional ingredient, added to explain the data. In quantum field theory, it comes naturally: particles that are not indistinguishable 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 5.8 already testified to the usefulness of the Fock space. If you check the derivations in chapter 9 leading to it, they were all conducted based on occupation numbers. A classical wave function for a system of photons was never written down; in fact, that cannot be done.



Figure 12.4: 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 12.4. In nonrelativistic quantum mechanics, it would be a completely trivial quantum system. In the case of identical bosons, all I of them would have to go into the only state there is. In the case of identical fermions, 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).

A relativistic system with just one type of single-particle state does seem very artificial, raising the question how esoteric the example is. But there are in fact two very well established classical systems that behave just like this:

1. The one-dimensional harmonic oscillator 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 angular 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. Any whole number particles can go into that state, each contributing energy $\hbar\omega$ to the system. (The half particle representing the ground state energy is in this interpretation considered to be a build-in part of the single-particle-state box in figure 12.4.) Reformulating the results of chapter 2.6.2 in quantum field theory terms: the harmonic oscillator ground state h_0 is the state $|0\rangle$ with zero particles, the excited state h_1 is the state $|1\rangle$ with one particle $\hbar\omega$, the excited state h_2 is the state $|2\rangle$ with two particles $\hbar\omega$, etcetera. The general wave function, either way, is a linear combination of these states, expressing an uncertainty in energy. Oops, excuse very much, 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 zero or one quantum \hbar of angular momentum in the 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 admittedly is quite poor, 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”

$$|\vec{0}\rangle \equiv |0, 0, 0, \dots\rangle \quad (12.20)$$

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 ; for the electron-spin example, it is the spin-down state. These are completely normal eigenstates that the system can be in. They are *not* zero wave functions, which would be unable to give any probabilities.

12.2.2 Annihilation and creation operators

The key to relativistic quantum mechanics is that particles can be annihilated or created. So it may not be surprising that it is very helpful to define operators that “annihilate” and “create” 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 12.4. However, there can be an arbitrary number of particles in that box.

Definition

The desired actions of the creation and annihilation operators are sketched in figure 12.5. An annihilation operator \hat{a} turns a state $|i\rangle$ with i particles into a state $|i - 1\rangle$ with $i - 1$ particles, and a creation operator \hat{a}^\dagger turns a state $|i\rangle$ with i particles into a state $|i + 1\rangle$ with $i + 1$ particles.

Mathematically, the operators are defined by the relations

$$\hat{a}|i\rangle = \alpha_i|i - 1\rangle \text{ but } \hat{a}|0\rangle = 0 \quad \hat{a}^\dagger|i\rangle = \alpha_i^\dagger|i + 1\rangle \text{ but } \hat{a}^\dagger|1\rangle = 0 \text{ for fermions} \quad (12.21)$$

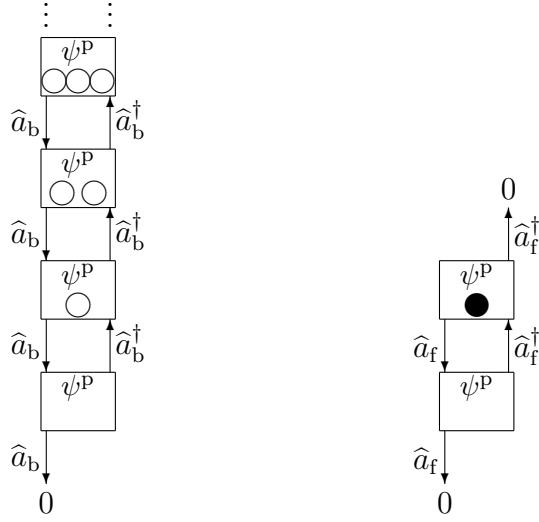


Figure 12.5: Annihilation and creation operators for a system with just one type of single particle state. Left: identical bosons; right: identical fermions.

where the α_i and α_i^\dagger are numerical constants still to be chosen. To avoid having to write the special cases separately each time, α_0 will be defined to be zero and, if it is a fermionic system, so will α_1^\dagger ; then you do not really need to worry about the fact that $| -1 \rangle$ and $| 2 \rangle$ do not exist.

Note that it is mathematically perfectly OK to *define* linear operators by specifying what they do to the basis states of a system. But you must hope that they will turn out to be operators that are mathematically helpful. To help achieve that, you want to choose the numerical constants appropriately. Consider what happens if the operators are applied in sequence:

$$\hat{a}^\dagger \hat{a}|i\rangle = \alpha_{i-1}^\dagger \alpha_i |i\rangle$$

Reading from right to left, the order in which the operators act on the state, first \hat{a} destroys a particle, then \hat{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 $\hat{a}^\dagger \hat{a}$ with eigenvalue $\alpha_{i-1}^\dagger \alpha_i$.

If the constants α_{i-1}^\dagger and α_i are chosen to make the eigenvalue a real number, then the operator $\hat{a}^\dagger \hat{a}$ will be Hermitian. More specifically, if they are chosen to make the eigenvalue equal to i , then $\hat{a}^\dagger \hat{a}$ 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

$$\alpha_i = \sqrt{i} \quad \alpha_{i-1}^\dagger = \sqrt{i} \quad \Rightarrow \quad \alpha_i^\dagger = \sqrt{i+1} \text{ except } \alpha_1^\dagger = 0 \text{ for fermions} \quad (12.22)$$

This choice of constants is particularly convenient since it makes the operators \hat{a} and \hat{a}^\dagger Hermitian conjugates. That means that if you take them to the other side in an inner product, they turn into each other:

$$\langle \underline{i} | \hat{a} | i \rangle = \langle \hat{a}^\dagger | \underline{i} \rangle | i \rangle \quad \langle | \underline{i} \rangle | \hat{a}^\dagger | i \rangle = \langle \hat{a} | \underline{i} \rangle | i \rangle$$

To see why, first note that states with different occupation numbers are taken to be orthonormal in Fock space. If the total number of particles is given, that follows from the classical form of the wave function. And the simple harmonic oscillator and spin examples of the previous subsection illustrate that it still applies when particles can be created or destroyed: these examples were just rewrites of orthonormal wave functions.

It follows that in the first equality above, the inner products are only nonzero if $i = \underline{i} + 1$: after lowering the particle number with \hat{a} , or raising it with \hat{a}^\dagger , the particle numbers must be the same at both sides of the inner product. When $i = \underline{i} + 1$, $\alpha_i = \alpha_{\underline{i}}^\dagger = \sqrt{i}$ so the equality still applies. The second equality is just the complex conjugate of the first, with a change in notations. It remains true for fermions, despite the fact that $\alpha_{f1}^\dagger = 0$ instead of $\sqrt{2}$, because there is no $|2\rangle$ state for which it would make a difference. Also, if it is true for the basis states, it is true for any combination of them.

You may well wonder why $\hat{a}^\dagger \hat{a}$ is the particle count operator; why not $\hat{a} \hat{a}^\dagger$? The reason is that $\hat{a} \hat{a}^\dagger$ would not work for the state $|0\rangle$ unless you took α_0^\dagger or α_1 zero, and then they could no longer create or annihilate the corresponding state. Still, it is interesting to see what the effect of $\hat{a} \hat{a}^\dagger$ is; according to the chosen definitions, for bosons

$$\hat{a}_b \hat{a}_b^\dagger |i\rangle = \sqrt{i+1} \sqrt{i+1} |i\rangle$$

So the operator $\hat{a}_b \hat{a}_b^\dagger$ has eigenvalues one greater than the number of particles. That means that if you subtract $\hat{a}_b \hat{a}_b^\dagger$ and $\hat{a}_b^\dagger \hat{a}_b$, you get the unit operator that leaves all states unchanged. And the difference between $\hat{a}_b \hat{a}_b^\dagger$ and $\hat{a}_b^\dagger \hat{a}_b$ is by definition the commutator of \hat{a}_b and \hat{a}_b^\dagger , indicated by square brackets:

$$[\hat{a}_b, \hat{a}_b^\dagger] \equiv \hat{a}_b \hat{a}_b^\dagger - \hat{a}_b^\dagger \hat{a}_b = 1 \quad (12.23)$$

Isn't that cute! Of course, $[\hat{a}_b, \hat{a}_b]$ and $[\hat{a}_b^\dagger, \hat{a}_b^\dagger]$ are zero since everything commutes with itself.

It does not work for fermions, because $\alpha_{f1}^\dagger = 0$ instead of $\sqrt{2}$. But for fermions, the only state for which $\hat{a}_f \hat{a}_f^\dagger$ produces something nonzero is $|0\rangle$ and then it leaves the state unchanged. Similarly, the only state for which $\hat{a}_f^\dagger \hat{a}_f$ produces something nonzero is $|1\rangle$ and then it leaves that state unchanged. That means that if you add $\hat{a}_f \hat{a}_f^\dagger$ and $\hat{a}_f^\dagger \hat{a}_f$ together, it reproduces the same state whether it is $|0\rangle$ or $|1\rangle$ (or any combination of them). The sum of $\hat{a}_f \hat{a}_f^\dagger$ and

$\hat{a}_f^\dagger \hat{a}_f$ is by definition called the “anticommutator” of \hat{a}_f and \hat{a}_f^\dagger and is indicated by curly brackets:

$$\{\hat{a}_f, \hat{a}_f^\dagger\} \equiv \hat{a}_f \hat{a}_f^\dagger + \hat{a}_f^\dagger \hat{a}_f = 1 \quad (12.24)$$

Isn’t that neat? Note also that $\{\hat{a}_b, \hat{a}_b\}$ and $\{\hat{a}_b^\dagger, \hat{a}_b^\dagger\}$ are zero since applying either operator twice ends up in a non-existing state.

How about the Hamiltonian? 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 $\hat{a}^\dagger \hat{a}$, that is $E^p \hat{a}^\dagger \hat{a}$. So, the total Hamiltonian for noninteracting particles becomes:

$$H = E^p \hat{a}^\dagger \hat{a} + E_{ve} \quad (12.25)$$

where E_{ve} stands for the vacuum, or ground state, energy of the system where there are no particles. This then allows the Schrödinger equation to be written in terms of occupation numbers and creation and annihilation operators.

The caHermitians

It is important to note that the creation and annihilation operators are not Hermitian, and therefore cannot correspond to physically measurable quantities. But since they are Hermitian conjugates, it is easy to form Hermitian operators from them:

$$\hat{P} \equiv \frac{1}{2}(\hat{a}^\dagger + \hat{a}) \quad \hat{Q} \equiv \frac{1}{2}i(\hat{a}^\dagger - \hat{a}) \quad (12.26)$$

Conversely, the creation and annihilation operators can be written as

$$\hat{a} = \hat{P} + i\hat{Q} \quad \hat{a}^\dagger = \hat{P} - i\hat{Q} \quad (12.27)$$

In lack of a better name that the author knows of, this book will call \hat{P} and \hat{Q} the caHermitians.

For bosons, the following commutators follow from the ones for the creation and annihilation operators:

$$[\hat{P}_b, \hat{Q}_b] = \frac{1}{2}i \quad [\hat{a}_b^\dagger \hat{a}_b, \hat{P}_b] = -i\hat{Q}_b \quad [\hat{a}_b^\dagger \hat{a}_b, \hat{Q}_b] = i\hat{P}_b \quad (12.28)$$

Therefore \hat{P}_b and \hat{Q}_b neither commute with each other, nor with the Hamiltonian. It follows that whatever physical variables they may stand for will not be certain at the same time, and will develop uncertainty in time if initially certain.

The Hamiltonian (12.25) for noninteracting particles may be written in terms of the caHermitians using (12.27) and (12.28), to give

$$H = E^p \left(\hat{P}_b^2 + \hat{Q}_b^2 - \frac{1}{2} \right) + E_{ve} \quad (12.29)$$

Often the energy turns out to be simply proportional to $\hat{P}_b^2 + \hat{Q}_b^2$; then the vacuum energy must be half a particle.

For fermions, the following useful relations follow from the anticommutators for the creation and annihilation operators

$$\hat{P}_f^2 = \frac{1}{4} \quad \hat{Q}_f^2 = \frac{1}{4} \quad (12.30)$$

The Hamiltonian then becomes

$$H = E^p \left(i[\hat{P}_f, \hat{Q}_f] + \frac{1}{2} \right) + E_{ve} \quad (12.31)$$

Examples

It is interesting to see how these ideas work out for the two example systems with just one single-particle state as described at the end of subsection 12.2.1.

Consider first the example of bosons that are energy quanta of a one-dimensional harmonic oscillator. The following discussion will *derive* the harmonic oscillator solution from scratch using the creation and annihilation operators. It provides an alternative to the much more algebraic derivation of chapter 2.6 and its note {A.12}.

The classical Hamiltonian can be written, in the notations of chapter 2.6,

$$H = \frac{1}{2m} \hat{p}^2 + \frac{1}{2} m\omega^2 \hat{x}^2$$

or in terms of $\hbar\omega$:

$$H = \hbar\omega \left(\frac{\hat{p}^2}{2\hbar m\omega} + \frac{m\omega \hat{x}^2}{2\hbar} \right)$$

From comparison with (12.29), it looks like maybe the caHermitians are

$$\hat{P} = \sqrt{\frac{m\omega}{2\hbar}} x \quad \hat{Q} = \sqrt{\frac{1}{2\hbar m\omega}} \hat{p}$$

For now just *define* them that way; also define $E_{ve} = \frac{1}{2}\hbar\omega$ and

$$\hat{a} = \hat{P} + i\hat{Q} \quad \hat{a}^\dagger = \hat{P} - i\hat{Q}$$

It has not yet been shown that \hat{a} and \hat{a}^\dagger are annihilation and creation operators. Nor that the Hamiltonian can be written in terms of them, instead of using \hat{P} and \hat{Q} .

However, the commutator $[\hat{P}, \hat{Q}]$ is according to the canonical commutator $i\frac{1}{2}$, which is just as it should be. That in turn implies that $[\hat{a}, \hat{a}^\dagger] = 1$. Then the Hamiltonian can indeed be written as

$$H = \hbar\omega(\hat{a}^\dagger \hat{a} + \frac{1}{2})$$

To see whether \hat{a}^\dagger is a creation operator, apply the Hamiltonian on a state $\hat{a}^\dagger|i\rangle$ where $|i\rangle$ is some arbitrary energy state whose further properties are still unknown:

$$H\hat{a}^\dagger|i\rangle = \hbar\omega(\hat{a}^\dagger\hat{a}\hat{a}^\dagger + \frac{1}{2}\hat{a}^\dagger)|i\rangle$$

But $\hat{a}\hat{a}^\dagger$ can be written as $\hat{a}^\dagger\hat{a} + 1$ because of the unit commutator, so

$$H\hat{a}^\dagger|i\rangle = \hbar\omega\hat{a}^\dagger(\hat{a}^\dagger\hat{a} + \frac{1}{2})|i\rangle + \hbar\omega\hat{a}^\dagger|i\rangle$$

The first term is just \hat{a}^\dagger times the Hamiltonian applied on $|i\rangle$, so this term multiplies $\hat{a}^\dagger|i\rangle$ with the energy E_i of the original $|i\rangle$ state. The second term multiplies $\hat{a}^\dagger|i\rangle$ also by a constant, but now $\hbar\omega$. It follows that $\hat{a}^\dagger|i\rangle$ is an energy eigenstate like $|i\rangle$, but with an energy that is one quantum $\hbar\omega$ higher. That does assume that $\hat{a}^\dagger|i\rangle$ is nonzero, because eigenfunctions may not be zero. Similarly, it is seen that $\hat{a}|i\rangle$ is an energy eigenstate with one quantum $\hbar\omega$ less in energy, if nonzero. So \hat{a}^\dagger and \hat{a} are indeed creation and annihilation operators.

Keep applying \hat{a} on the state $|i\rangle$ to lower its energy even more. This must eventually terminate in zero because the energy cannot become negative. (That assumes that the eigenfunctions are mathematically reasonably well behaved, as the solution of chapter 2.6 verifies they are. That can also be seen without using these solutions, so it is not cheating.) Call the final nonzero state $|0\rangle$. Solve the fairly trivial equation $\hat{a}|0\rangle = 0$ to find the lowest energy state $|0\rangle = h_0(x)$ and note that it is unique. (And that is the same one as derived in chapter 2.6.) Use \hat{a}^\dagger to go back up the energy scale and find the other energy states $|i\rangle = h_i$ for $i = 1, 2, 3, \dots$. Verify using the Hamiltonian that going down in energy with \hat{a} and then up again with \hat{a}^\dagger brings you back to a multiple of the original state, not to some other state with the same energy or to zero. Conclude therefore that all energy states have now been found. And that their energies are spaced apart by whole multiples of the quantum $\hbar\omega$.

While doing this, it is convenient to know not just that $\hat{a}|i\rangle$ produces a multiple of the state $|i-1\rangle$, but also what multiple α_i that is. Now the α_i can be made real and positive by a suitable choice of the normalization factors of the various energy eigenstates $|i\rangle$. Then the α_i^\dagger are positive too because $\alpha_{i-1}^\dagger\alpha_i$ produces the number of energy quanta in the Hamiltonian. The magnitude can be deduced from the square norm of the state produced. In particular, for $\hat{a}|i\rangle$:

$$\alpha_i^* \alpha_i \langle \hat{a}|i\rangle | \hat{a}|i\rangle \rangle = \langle |i\rangle | \hat{a}^\dagger \hat{a}|i\rangle \rangle = i$$

the first because the Hermitian conjugate of \hat{a} is \hat{a}^\dagger and the latter because $\hat{a}^\dagger\hat{a}$ must give the number of quanta. So $\alpha_i = \sqrt{i}$, and then $\alpha_i^\dagger = \sqrt{i+1}$ from the above. The harmonic oscillator has been solved. This derivation using the ideas of quantum field theory is much neater than the classical one; just compare the very logical story above to the algebraic mess in note {A.12}.

It should be noted, however, that a completely equivalent derivation can be given using the classical description of the harmonic oscillator. Many books do in fact do it that way, e.g. [17]. In the classical treatment, the creation and annihilation operators are called the “ladder” operators. But without the ideas of quantum field theory, it is difficult to justify the ladder operators by anything better than as a weird mathematical trick that just turns out to work.

If you have read the advanced section on angular momentum, the example system for fermions is also interesting. In that model system, the Hamiltonian is a multiple of the angular momentum in the z -direction of an electron. The state $|0\rangle$ is the spin-down state \downarrow and the state $|1\rangle$ is the spin-up state \uparrow . Now the annihilation operator must turn \uparrow into \downarrow and \downarrow into zero. In terms of the so-called Pauli spin matrices of section 10.1.9, the operator that does that is $\frac{1}{2}(\sigma_x - i\sigma_y)$. Similarly, the creation operator is $\frac{1}{2}(\sigma_x + i\sigma_y)$. That makes the caHermitians $\frac{1}{2}\sigma_x$ and $-\frac{1}{2}\sigma_y$. The commutator $i[P, Q]$ that appears in the Hamiltonian is then $\frac{1}{2}\sigma_z$, which is a multiple of the angular momentum in the z -direction as it should be.

More single particle states

Now consider the case that there is more than one type of single-particle state. Graphically there is now more than one particle box, as in figures 12.2 and 12.3. Then an annihilation operator \hat{a}_n and a creation operator \hat{a}_n^\dagger must be defined for each type of single-particle state ψ_n^p . In other words, there is one for each occupation number i_n . The mathematical definition of these operators for bosons is

$$\begin{aligned} \hat{a}_{b,n}|i_1, i_2, \dots, i_{n-1}, i_n, i_{n+1}, \dots\rangle &= \sqrt{i_n} |i_1, i_2, \dots, i_{n-1}, i_n - 1, i_{n+1}, \dots\rangle \\ \hat{a}_{b,n}^\dagger|i_1, i_2, \dots, i_{n-1}, i_n, i_{n+1}, \dots\rangle &= \sqrt{i_n + 1}|i_1, i_2, \dots, i_{n-1}, i_n + 1, i_{n+1}, \dots\rangle \end{aligned} \quad (12.32)$$

The commutation relations are

$$[\hat{a}_{b,n}, \hat{a}_{b,\underline{n}}] = 0 \quad [\hat{a}_{b,n}^\dagger, \hat{a}_{b,\underline{n}}^\dagger] = 0 \quad [\hat{a}_{b,n}, \hat{a}_{b,\underline{n}}^\dagger] = \delta_{n\underline{n}} \quad (12.33)$$

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, so it does not make a difference in which order you apply them. For example, $\hat{a}_{b,n}\hat{a}_{b,\underline{n}} = \hat{a}_{b,\underline{n}}\hat{a}_{b,n}$, so the commutator $\hat{a}_{b,n}\hat{a}_{b,\underline{n}} - \hat{a}_{b,\underline{n}}\hat{a}_{b,n}$ is zero. For $n = \underline{n}$, they 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 12.3 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 logical ones, 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, but the occupation-number wave function will not unless you do something. What you can do is define the annihilation and creation operators for fermions as follows:

$$\begin{aligned}\hat{a}_{f,n}|i_1, i_2, \dots, i_{n-1}, 0, i_{n+1}, \dots\rangle &= 0 \\ \hat{a}_{f,n}|i_1, i_2, \dots, i_{n-1}, 1, i_{n+1}, \dots\rangle &= (-1)^{i_1+i_2+\dots+i_{n-1}}|i_1, i_2, \dots, i_{n-1}, 0, i_{n+1}, \dots\rangle \\ \hat{a}_{f,n}^\dagger|i_1, i_2, \dots, i_{n-1}, 0, i_{n+1}, \dots\rangle &= (-1)^{i_1+i_2+\dots+i_{n-1}}|i_1, i_2, \dots, i_{n-1}, 1, i_{n+1}, \dots\rangle \\ \hat{a}_{f,n}^\dagger|i_1, i_2, \dots, i_{n-1}, 1, i_{n+1}, \dots\rangle &= 0\end{aligned}\tag{12.34}$$

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 $(-1)^{\dots}$. 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 $\hat{a}_n^\dagger \hat{a}_n$ 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:

$$\left\{ \hat{a}_{f,n}, \hat{a}_{f,\underline{n}} \right\} = 0 \quad \left\{ \hat{a}_{f,n}^\dagger, \hat{a}_{f,\underline{n}}^\dagger \right\} = 0 \quad \left\{ \hat{a}_{f,n}, \hat{a}_{f,\underline{n}}^\dagger \right\} = \delta_{nn}\tag{12.35}$$

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 non interacting particles is now found by summing over all types of single-particle states:

$$H = \sum_n E_n^p \hat{a}_n^\dagger \hat{a}_n + E_{ve,n}\tag{12.36}$$

This Hamiltonian conserves the number of particles in each state; particles destroyed by the annihilation operator are immediately restored by the creation

operator. Phrased differently, this Hamiltonian commutes with the operator $\hat{a}_n^\dagger \hat{a}_n$ giving the number of particles in a state ψ_n^p , so energy eigenstates can be taken to be states with given numbers of particles in each single-particle state. Such systems can be described by classical quantum mechanics. But the next subsection will give an example of particles that do interact, making quantum field theory essential.

12.2.3 Quantization of radiation

In the discussion of the emission and absorption of radiation in chapter 6.3.3, the electromagnetic field was the classical Maxwell one. However, that cannot be right. According to the Planck-Einstein relation, the electromagnetic field comes in discrete quanta of energy $\hbar\omega$ called photons. A classical electromagnetic field cannot explain that.

The electromagnetic field must be described by operators that act nontrivially on a wave function that includes photons. To identify these operators will involve two steps. First a more careful look needs to be taken at the classical electromagnetic field, and in particular at its energy. By comparing that with the Hamiltonian in terms of creation and annihilation operators, as given in the previous section, the operators corresponding to the electromagnetic field can then be inferred.

To make the process more intuitive, it helps to initially assume that the electromagnetic field is not in the form of a traveling wave, but of radiation confined to a suitable box of volume V .

Classical energy

The discussion on emission and absorption of radiation in chapter 6.3.3 assumed the electromagnetic field to be the single traveling wave

$$\vec{E} = \hat{k}E_0 \cos(\omega t - \phi - ky) \quad c\vec{B} = \hat{i}E_0 \cos(\omega t - \phi - ky)$$

where \vec{E} and \vec{B} are the electric and magnetic fields, respectively, ω is the frequency, k is the wave number ω/c , and ϕ is an arbitrary constant phase angle. This wave is moving in the positive y direction with the speed of light c . To get a standing wave, add a second wave going the other way:

$$\vec{E} = -\hat{k}E_0 \cos(\omega t - \phi + ky) \quad c\vec{B} = \hat{i}E_0 \cos(\omega t - \phi + ky).$$

If these are added together, using the trig formula for separating the cosines into time-only and space-only sines and cosines, the result is

$$\vec{E} = \hat{k}e_k(t) \sin(ky) \quad c\vec{B} = \hat{i}b_k(t) \cos(ky) \quad (12.37)$$

$$e_k(t) = 2E_0 \sin(\omega t - \phi) \quad b_k(t) = 2E_0 \cos(\omega t - \phi) \quad (12.38)$$

Alternatively, just plug the assumption (12.37) directly into Maxwell's equations. That produces

$$\frac{de_k}{dt} = \omega b_k \quad \frac{db_k}{dt} = -\omega e_k \quad (12.39)$$

The solution to those equations are the coefficients (12.38) above.

Such a standing wave solution is appropriate for a box with perfectly conducting walls at $y = 0$ and $y = \ell_y$, where $\sin(k\ell_y) = 0$. On perfectly conducting walls the electric field \vec{E} must be zero on behalf of Ohm's law. For the other surfaces of the box, just assume periodic boundary conditions over some chosen periods ℓ_x and ℓ_z , their length does not make a difference here.

Next, it is shown in basic textbooks on electromagnetics that the energy in an electromagnetic field is

$$E_V = \frac{1}{2}\epsilon_0 \int_V \vec{E}^2 + c^2 \vec{B}^2 d^3\vec{r} \quad (12.40)$$

where $\epsilon_0 = 8.85 \cdot 10^{-12} \text{ C}^2/\text{J m}$ is the permittivity of space. This energy is typically derived by comparing the energy obtained from discharging a condenser with the electric field it holds when charged, and from a coil compared to its magnetic field. Note that this expression implies that the energy of the electric field of a point charge is infinite.

As an aside, if the energy E_V is differentiated with respect to time, substituting in Maxwell's equations to get rid of the time derivatives, and cleaning up, the result is

$$\frac{dE_V}{dt} = -\epsilon_0 c \int_V \nabla \cdot (\vec{E} \times c\vec{B}) d^3\vec{r}$$

From the divergence theorem it can now be seen that the flow rate of electromagnetic energy is given by the "Poynting vector"

$$\epsilon_0 c^2 \vec{E} \times \vec{B} \quad (12.41)$$

This was included because other books have Poynting vectors and you would be very disappointed if yours did not.

The important point here is that in terms of the coefficients e_k and b_k , the energy is found to be

$$E_V = \frac{\epsilon_0 V}{4} (e_k^2 + b_k^2) \quad (12.42)$$

Quantization

Following the Planck-Einstein relation, electromagnetic radiation should come in photons, each with one unit $\hbar\omega$ of energy. This indicates that the energy in

the electromagnetic field is not a classical value, but corresponds to the discrete eigenvalues of some as yet unknown Hamiltonian operator. The question then is, what is that Hamiltonian?

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, in chapter 6, that it did produce the correct classical momentum for macroscopic systems, as well as correct quantum results like the energy levels of the hydrogen atom, in chapter 3.2. A similar leap of faith will be needed to quantize the electromagnetic field.

Whatever the details of the Hamiltonian, it is clear that the appropriate mathematical tool is here quantum field theory. After all, photons are not conserved particles. Atoms readily absorb them or radiate new ones when they heat up. Now the system considered here involves only one mode of radiation; therefore the wave function can be indicated by the simple Fock space ket $|i\rangle$, where i is the number of photons present in the mode. Also, since the single-photon energy is $\hbar\omega$, the quantum field Hamiltonian operator (12.29) becomes

$$H = \hbar\omega \left(\hat{P}^2 + \hat{Q}^2 - \frac{1}{2} \right) + E_{ve}$$

Somehow then, the caHermitian operators \hat{P} and \hat{Q} must be identified. They must produce the classical equations of motion in the macroscopic limit, including the classical energy (12.42),

$$E_V = \frac{\epsilon_0 V}{4} (e_k^2 + b_k^2)$$

Comparing the two under macroscopic conditions in which $\frac{1}{2}\hbar\omega$ and the ground state energy E_{ve} can be ignored, the very simplest assumption is that the caHermitians are scaled versions of the coefficients e_k and b_k :

$$e_k \rightarrow 2\varepsilon_p P = \varepsilon_p (\hat{a}^\dagger + a) \quad b_k \rightarrow 2\varepsilon_p Q = \varepsilon_p \mathrm{i}(\hat{a}^\dagger - \hat{a}) \quad (12.43)$$

where the scaling factor must be

$$\varepsilon_p = \sqrt{\frac{\hbar\omega}{\epsilon_0 V}} \quad (12.44)$$

This scaling factor can be thought of as the square root of the nominal mean square electric field per photon.

If this association of mode coefficients with caHermitian operators is indeed correct even under non macroscopic conditions, one immediate consequence is that the ground state energy must be equal to that of half a photon. And that is just for the single mode considered here. Since there are infinitely many modes of radiation, the total vacuum energy is infinite.

While that is certainly counterintuitive, it may be noted that even classically, the energy in the electromagnetic field is infinite, assuming that electrons are indeed point charges. On the other hand, the caHermitians are the Hermitian components of the very logically and simply defined creation and annihilation operators, and you would really *expect* them to be physically meaningful. They certainly were for the harmonic oscillator and spin system examples of subsection 12.2.2.

Therefore, the assumption will be made that the scaled caHermitians appear in the quantized electromagnetic field where the measurable quantities e_k and b_k appear in the classical electromagnetic field:

$$\hat{\vec{E}}(\vec{r}) = \hat{k}\varepsilon_p \sin(ky)(\hat{a}^\dagger + \hat{a}) \quad c\hat{\vec{B}}(\vec{r}) = i\varepsilon_p \cos(ky)i(\hat{a}^\dagger - \hat{a}) \quad (12.45)$$

This process of replacing the coefficients of the modes by operators is called “second quantization.” No, there was no earlier quantization of the electromagnetic field involved. The word “second” is there for historical reasons: physicists have historically found it hysterical to confuse students.

Note that just like the time-dependent momentum of a classical particle $p(t)$ becomes the time-independent operator $\hbar\partial/\mathrm{i}\partial x$, the creation and annihilation operators are taken to be time-independent. In quantum mechanics, the time dependence is in the wave function, not the operators:

$$|\Psi\rangle = \sum_i c_i e^{\mathrm{i}E_i t/\hbar} |i\rangle$$

where the energy E_i of the state with i photons is $(i + \frac{1}{2})\hbar\omega$. (The Heisenberg picture absorbs the time dependence in the operator; that is particularly convenient for relativistic applications. However, true relativity is beyond the scope of this book and this discussion will stay as close to the normal Schrödinger picture as possible.)

To see whether this quantization of the electromagnetic field does indeed make sense, its immediate consequences will now be explored. First consider the Hamiltonian according to the Newtonian (or is that Maxwellian?) analogy:

$$H = \frac{1}{2}\epsilon_0 \int_V \hat{\vec{E}}^2 + c^2 \hat{\vec{B}}^2 \mathrm{d}^3\vec{r}$$

Substituting in the quantized electromagnetic field, (12.44) and (12.45), and integrating,

$$H = \frac{1}{2}\epsilon_0 \frac{\hbar\omega}{\epsilon_0 V} \left(\frac{1}{2}V(\hat{a}^\dagger + \hat{a})^2 - \frac{1}{2}V(\hat{a}^\dagger - \hat{a})^2 \right)$$

Multiplying out gives

$$H = \frac{1}{4}\hbar\omega(2\hat{a}^\dagger\hat{a} + 2\hat{a}\hat{a}^\dagger)$$

and using the commutator $\hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a} = 1$ to get rid of $\hat{a}\hat{a}^\dagger$ gives

$$H = \hbar\omega\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)$$

which is just as it should be.

Also consider the equation for the expectation value $\langle P \rangle$:

$$\frac{d\langle P \rangle}{dt} = \frac{i}{\hbar}\langle[H, P]\rangle = \omega\langle Q \rangle$$

using (12.28) and (12.29). Similarly for the expectation value $\langle Q \rangle$:

$$\frac{d\langle Q \rangle}{dt} = \frac{i}{\hbar}\langle[H, Q]\rangle = -\omega\langle P \rangle$$

Those are the same equations as satisfied by e_k and b_k . They are also the equations for the position and linear momentum of a harmonic oscillator, when suitably scaled. It is often said that the mode amplitudes of the electromagnetic field are mathematically modelled as quantum harmonic oscillators.

Now suppose there are exactly i photons, what is the expectation value of the electric field at a given position and time? Well, the wave function will be

$$|\Psi\rangle = c_i e^{i(i+\frac{1}{2})\omega t}|i\rangle$$

and so

$$\langle \vec{E} \rangle(\vec{r}, t) = \hat{k}\varepsilon_p \sin(ky) c_i^* e^{i(i+\frac{1}{2})\omega t} c_i e^{-i(i+\frac{1}{2})\omega t} \langle i | \hat{a}^\dagger + \hat{a} | i \rangle$$

That is zero because \hat{a}^\dagger and \hat{a} applied on $|i\rangle$ create states $|i+1\rangle$ and $|i-1\rangle$ that are orthogonal to the $|i\rangle$ in the left side of the inner product. The same way the expectation magnetic field is zero!

Oops, not quite as expected. In fact, the previous subsection pointed out that the caHermitians do not commute with the Hamiltonian. If the number of photons, hence the energy, is certain, then the electromagnetic field is not. And the caHermitians also do not commute with each other; if the electric field is certain, then the magnetic field is not, and vice-versa.

To get something resembling a classical electric field, there must be uncertainty in energy. In particular, if the coefficients of multiple energy states are nonzero, then the expectation values of the electric and magnetic fields become

$$\langle \vec{E} \rangle(\vec{r}, t) = \hat{k}\varepsilon_p \sin(ky) \sum_i \sqrt{i} (c_i^* c_{i-1} e^{i\omega t} + c_i c_{i-1}^* e^{-i\omega t})$$

$$\langle \vec{B} \rangle(\vec{r}, t) = \hat{i}\varepsilon_p \cos(ky) i \sum_i \sqrt{i} (c_i^* c_{i-1} e^{i\omega t} - c_i c_{i-1}^* e^{-i\omega t})$$

To check this, observe that for any i , $\langle i | \hat{a}^\dagger | i \rangle$ is only nonzero if $\underline{i} = i - 1$ and the same for its complex conjugate $\langle i | \hat{a} | i \rangle$. Renotating

$$\sum_i \sqrt{i} c_i c_{i-1}^* \equiv i C_1 e^{i\phi_1}$$

where C_1 and ϕ_1 are real constants produces

$$\begin{aligned}\langle \vec{E} \rangle(\vec{r}, t) &= \hat{k} 2\varepsilon_p C_1 \sin(ky) \sin(\omega t - \phi_1) \\ \langle c\vec{B} \rangle(\vec{r}, t) &= i 2\varepsilon_p C_1 \sin(ky) \cos(\omega t - \phi_1)\end{aligned}$$

That is in the form of the classical electromagnetic field (12.37).

Similarly it may be found that

$$\begin{aligned}\langle E^2 \rangle(\vec{r}, t) &= 2\varepsilon_p^2 \sin^2(ky) \left[\langle i \rangle + \frac{1}{2} + C_2 \sin(2\omega t - 2\phi_2) \right] \\ \langle c^2 B^2 \rangle(\vec{r}, t) &= 2\varepsilon_p^2 \cos^2(ky) \left[\langle i \rangle + \frac{1}{2} - C_2 \sin(2\omega t - 2\phi_2) \right]\end{aligned}$$

where the expectation number of photons and the constants are defined by

$$\sum_i |c_i|^2 i \equiv \langle i \rangle \quad \sum_i c_i c_{i-2}^* \sqrt{i(i-1)} \equiv i C_2 e^{i2\phi_2}$$

Note that the mean square expectation electric field is ε_p^2 per photon, with half a photon left in the ground state.

Consider now a “photon packet” in which the numbers of photons with nonzero probabilities are restricted to a relatively narrow range. However, assume that the range is still large enough so that the coefficients can vary slowly from one number of photons to the next, except for possibly a constant phase difference:

$$c_{i-1} \approx c_i e^{-i(\phi - \frac{\pi}{2})}$$

Then

$$C_1 \approx \sqrt{\langle i \rangle} \quad C_2 \approx \langle i \rangle \quad \phi_1 \approx \phi \quad 2\phi_2 \approx 2\phi + \frac{\pi}{2}$$

In that case, the expectation electromagnetic field above becomes the classical one, with an energy of about $\langle i \rangle$ photons.

Photon spin

If photons act as particles, they should have a value of spin. Physicists have concluded that the appropriate spin operators are

$$\hat{S}_x = -i\hbar(\hat{j}\hat{k} \cdot - \hat{k}\hat{j} \cdot) \quad \hat{S}_y = -i\hbar(\hat{k}\hat{i} \cdot - \hat{i}\hat{k} \cdot) \quad \hat{S}_z = -i\hbar(\hat{i}\hat{j} \cdot - \hat{j}\hat{i} \cdot) \quad (12.46)$$

Note the dots. These operators need to act on vectors and then produce vectors times dot products. (In terms of linear algebra, the unit vectors above are multiplied as tensor products.)

If the above operators are correct, they should satisfy the fundamental commutation relations (4.20). They do. For example:

$$[\hat{S}_x, \hat{S}_y] = -\hbar^2(\hat{j}\hat{k} \cdot - \hat{k}\hat{j} \cdot)(\hat{k}\hat{i} \cdot - \hat{i}\hat{k} \cdot) + \hbar^2(\hat{k}\hat{i} \cdot - \hat{i}\hat{k} \cdot)(\hat{j}\hat{k} \cdot - \hat{k}\hat{j} \cdot)$$

Since the unit vectors are mutually orthogonal, when multiplying out the first term only the dot product $\hat{k} \cdot \hat{k} = 1$ is nonzero, and the same for the second term. So

$$[\hat{S}_x, \hat{S}_y] = -\hbar^2\hat{j}\hat{i} + \hbar^2\hat{i}\hat{j} = i\hbar\hat{S}_z$$

The next question is how to apply them, given that the wave function of photons is a Fock space ket without a precise physical interpretation. However, following the ideas of quantum mechanics, presumably photon wave functions can be written as linear combinations of wave functions with definite electric fields. All the ones for the single mode considered here have an electric field that is proportional to \hat{k} , and the operators above can be applied on that. But unfortunately, \hat{k} is not an eigenvector of any of the spin components above.

However, even given the direction of wave propagation and wave number, there are still two different modes: the electric field could be fluctuating in the x -direction instead of the z -direction. (Fluctuating in an oblique direction is just a linear combination of these two independent modes, and not another possibility.) In short, any of the considered electromagnetic modes can be rotated 90 degrees around the y -axis to give a second mode with an electric field proportional to \hat{i} . If these modes are combined pairwise in the combination $\hat{k} + i\hat{i}$ it produces an eigenstate of \hat{S}_y with spin \hbar , while $\hat{k} - i\hat{i}$ produces an eigenstate with spin $-\hbar$. That can easily be checked by direct substitution in the eigenvalue problem.

Note that there are only two independent states, so that is it. There is no third state with spin zero in the y -direction, the direction of wave propagation. The missing state reflects the classical limitation that the electric field cannot have a component in the direction of wave propagation, chapter 10.4. However, it can be seen using the analysis of chapter 10.1 that suitable combinations of equal amounts of spin forward and backward in y can produce photons with zero spin in a direction normal to the direction of propagation.

One thing should still be checked: that the magnetic field does not conflict. Now, if the electric field is rotated from \hat{k} to \hat{i} , the magnetic field rotates from \hat{i} to $-\hat{k}$. So the eigenstates have magnetic fields proportional to $\hat{i} - i\hat{k}$ and $\hat{i} + i\hat{k}$. That is just $-i$, respectively i times the vectors of the electric field, so even including the magnetic field the states remain eigenstates.

Traveling waves

To get the quantized electromagnetic field of traveling waves, the quickest way to get there is to take the standing wave apart using

$$\sin ky = \frac{e^{iky} - e^{-iky}}{2i} \quad \cos ky = \frac{e^{iky} + e^{-iky}}{2}$$

Then the parts that propagate forwards in y must be of the form $\hat{a}e^{iky}$ or $\hat{a}^\dagger e^{-iky}$. To see why, just look at the expectation values of these terms for the generic wave function $\sum_i c_i e^{-i\omega t} |i\rangle$.

A single wave of wave number k moving in the positive y -direction and polarized in the z -direction therefore takes the form

$$\hat{\vec{E}}(\vec{r}) = \hat{k}\varepsilon'_p i(\hat{a}^\dagger e^{-iky} - \hat{a}e^{iky}) \quad c\hat{\vec{B}}(\vec{r}) = i\varepsilon'_p (\hat{a}^\dagger e^{-iky} - \hat{a}e^{iky}) \quad (12.47)$$

where the scaling constant is

$$\varepsilon'_p = \sqrt{\frac{\hbar\omega}{2\epsilon_0 V'}} \quad (12.48)$$

For a traveling wave, it is more physical to assume that it is in a box that is periodic in the y -direction; that is true for a box with twice the length, hence a volume $V' = 2V$. The constant ε'_p can be thought of as the square root of half the mean square electric field per photon.

However, in general there will be a similar wave polarized in the x -direction. And then there will be pairs of such waves for different directions of propagation and different wave numbers k . To describe all these waves, it is convenient to combine wave number and direction of propagation into a wave number vector \vec{k} that has the magnitude of k and the direction of wave propagation. Then the complete electromagnetic field operators become

$$\hat{\vec{E}}(\vec{r}) = \varepsilon'_p \sum_{\mu=1}^2 (-1)^\mu \hat{i}_\mu \sum_{\vec{k}} i(\hat{a}_{\vec{k},\mu}^\dagger e^{-i\vec{k}\cdot\vec{r}} - \hat{a}_{\vec{k},\mu} e^{i\vec{k}\cdot\vec{r}}) \quad (12.49)$$

$$c\hat{\vec{B}}(\vec{r}) = \varepsilon'_p \sum_{\mu=1}^2 \hat{i}_{3-\mu} \sum_{\vec{k}} i(\hat{a}_{\vec{k},\mu}^\dagger e^{-i\vec{k}\cdot\vec{r}} - \hat{a}_{\vec{k},\mu} e^{i\vec{k}\cdot\vec{r}}) \quad (12.50)$$

where the unit vector \hat{i}_1 must be chosen normal to the direction of wave propagation and $\hat{i}_2 = \hat{i}_1 \times \vec{k}/k$.

12.2.4 Spontaneous emission

In this subsection, the spontaneous emission rate of excited atoms will be derived. It may be recalled that this was done in chapter 6.3.10 following an

argument given by Einstein. However, that was cheating: it peeked at the answer for blackbody radiation. This section will verify that a quantum treatment gives the same answer.

Like in chapter 6.3, consider the interaction of an atom with electromagnetic radiation, but this time, do it right, using the quantized electromagnetic field instead of the classical one. The approach will again be to consider the interaction for a two state system involving a single electromagnetic wave and two energy levels of the atom. The total effects should then again follow from summation over all waves and energy levels.

The most appropriate energy states are now

$$\psi_1 = \psi_L |i+1\rangle \quad \psi_2 = \psi_H |i\rangle \quad (12.51)$$

In state ψ_1 the atom is in the lower energy state and there are $i+1$ photons in the wave; in state ψ_2 , the atom is excited, but one photon has disappeared. Note that these wave functions depend on both the atom energy level and on the number of photons.

The Hamiltonian is now

$$H = H_0 + \hbar\omega(\hat{a}^\dagger\hat{a} + \frac{1}{2}) + ez\varepsilon'_p i(\hat{a}^\dagger - \hat{a}) \quad (12.52)$$

where H_0 is the Hamiltonian of the unperturbed atom, the second term is the Hamiltonian of the electromagnetic field, and the final term is the $e\hat{E}_z z$ interaction energy of the electron with the electric field, but now quantized as in (12.47). It was again assumed that interaction with the magnetic field can be ignored and that the atom is small enough compared to the electromagnetic wave length that it can be taken to be at the origin.

Next the Hamiltonian matrix coefficients are needed. The first one is

$$H_{11} = \langle i+1 | \psi_L | (H_0 + \hbar\omega(\hat{a}^\dagger\hat{a} + \frac{1}{2}) + ez\varepsilon'_p i(\hat{a}^\dagger - \hat{a})) \psi_L | i+1 \rangle$$

The first term of the Hamiltonian acts on the spatial state and produces the lower atom energy level. The second term acts on the Fock space ket to produce the electromagnetic energy of $i+1$ photons. The final term produces zero, because of the symmetry of the atom eigenstates. Alternatively, it produces zero because \hat{a}^\dagger and \hat{a} produce states $|i+2\rangle$ and $|i\rangle$ that are orthogonal to the $|i+1\rangle$ in the left side of the inner product. So

$$H_{11} = E_L + (i+1+\frac{1}{2})\hbar\omega$$

and similarly

$$H_{22} = E_H + (i+\frac{1}{2})\hbar\omega$$

The remaining Hamiltonian matrix coefficient is

$$H_{12} = \langle i+1 | \psi_L | (H_0 + \hbar\omega(\hat{a}^\dagger\hat{a} + \frac{1}{2}) + ez\varepsilon'_p i(\hat{a}^\dagger - \hat{a})) \psi_H | i \rangle$$

Here the first two terms produce zero, because ψ_1 and ψ_2 are orthonormal eigenstates of these operators. The third term produces

$$H_{12} = i\varepsilon'_p \langle \psi_L | ez | \psi_H \rangle \sqrt{i+1}$$

because the creation operator \hat{a}^\dagger turns $|i\rangle$ into $\sqrt{i+1}|i+1\rangle$.

As in chapter 6.3, a Hamiltonian of a simplified two-state system may be defined as

$$\bar{H}_{12} \equiv H_{12} e^{-i \int (H_{22} - H_{11}) dt / \hbar}$$

where from the above

$$H_{22} - H_{11} = E_H - E_L - \hbar\omega = \hbar(\omega_p - \omega)$$

where ω_p is the frequency of the photon released in a transition from the higher to the lower atom energy state. The simplified two state system becomes

$$\dot{\bar{a}} = \frac{\varepsilon'_p}{\hbar} \langle \psi_L | ez | \psi_H \rangle \sqrt{i+1} e^{i(\omega-\omega_p)t} \bar{b} \quad \dot{\bar{b}} = -\frac{\varepsilon'_p}{\hbar} \langle \psi_L | ez | \psi_H \rangle^* \sqrt{i+1} e^{-i(\omega-\omega_p)t} \bar{a}$$

where \bar{a} and \bar{b} give the probabilities of states ψ_1 and ψ_2 respectively.

Consider a system that starts out with the atom in the excited state, $a_0 = 0$ and $b_0 = 1$. Then, if the perturbation is weak over the time that it acts, \bar{b} can be approximated as one in the first equation, and the transition probability to the lower atom energy state ψ_2 is found to be

$$|\bar{a}|^2 = \frac{1}{4} \frac{4(i+1)\varepsilon'^2_p}{\hbar^2} |\langle \psi_L | ez | \psi_H \rangle|^2 t^2 \left(\frac{\sin \frac{1}{2}(\omega - \omega_p)t}{\frac{1}{2}(\omega - \omega_p)t} \right)^2$$

For a large number of photons i , this is the same as the classical result (6.35), because $4\varepsilon'^2_p$ is the peak square electric field per photon.

But now consider the electromagnetic ground state where the number of photons i is zero. The transition probability above is as if there is still one photon of electromagnetic energy left. And as noted in chapter 6.3.10, that is exactly what is needed to explain spontaneous emission using the quantum equations.

Some additional observations may be interesting. While you may think of it as excitation by the ground state electromagnetic field, the actual energy of the ground state was earlier seen to be 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: the Hamiltonian coefficient H_{12} is the energy if the atom is not excited if the atom is excited. Think of it in classical common sense terms. There is an excited atom and no photons around it. (Or if you prefer, the number of photons is as low as it can

ever get. Classical common sense would make that zero.) Why would things ever change? But in quantum mechanics, the twilight term allows the excited atom to interact with the electromagnetic radiation of the photon that would be there if it was not excited. Sic.

12.2.5 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 best to work in terms of position coordinates instead of single-particle energy eigenfunctions. Relativistic applications must make sure that matter does not exceed the speed of light, and that coordinate systems moving at different speeds are physically equivalent and related through the Lorentz transformation. 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” that annihilate or create particles at a given position in space. Now classically, a particle at a given position \vec{r}_0 corresponds to a wave function that is a delta function, $\Psi = \delta(\vec{r} - \vec{r}_0)$, chapter 6.4. A delta function can be written in terms of the single-particle eigenfunctions ψ_n as $\sum c_n \psi_n$. Here the constants can be found from taking inner products; $c_n = \langle \psi_n | \Psi \rangle$, and that gives $c_n = \psi_n^*(\vec{r}_0)$ because of the property of the delta function to pick out that value of any function that it is in an inner product with. Since c_n is the amount of eigenfunction ψ_n that must be annihilated/created to annihilate/create the delta function at \vec{r}_0 , the field operators become

$$\hat{a}(\vec{r}) = \sum_n \psi_n^*(\vec{r}) \hat{a}_n \quad \hat{a}^\dagger(\vec{r}) = \sum_n \psi_n^*(\vec{r}) \hat{a}_n^\dagger \quad (12.53)$$

where the subscript zero was dropped from \vec{r} since it is no longer needed to distinguish from the independent variable of the delta function. It means that \vec{r} is now the position at which the particle is annihilated/created.

In the case of particles in free space, the energy eigenfunctions are the momentum eigenfunctions $e^{i\vec{p}\cdot\vec{r}/\hbar}$, and the sums become integrals called the Fourier transforms; see chapter 6.4 and 6.5.1 for more details. In fact, unless you are particularly interested in converting the expression (12.36) for the Hamiltonian, basing the field operators on the momentum eigenfunctions works fine even if the particles are not in free space.

A big advantage of the way the annihilation and creation operators were defined now shows up: their (anti)commutation relations are effectively unchanged

in taking linear combinations. In particular

$$\boxed{\begin{aligned} [\hat{a}_b(\vec{r})\hat{a}_b(\vec{r}')] &= 0 & [\hat{a}_b^\dagger(\vec{r})\hat{a}_b^\dagger(\vec{r}')] &= 0 & [\hat{a}_b(\vec{r})\hat{a}_b^\dagger(\vec{r}')] &= \delta^3(\vec{r} - \vec{r}') \end{aligned}} \quad (12.54)$$

$$\boxed{\begin{aligned} \{\hat{a}_f(\vec{r})\hat{a}_f(\vec{r}')\} &= 0 & \{\hat{a}_f^\dagger(\vec{r})\hat{a}_f^\dagger(\vec{r}')\} &= 0 & \{\hat{a}_f(\vec{r})\hat{a}_f^\dagger(\vec{r}')\} &= \delta^3(\vec{r} - \vec{r}') \end{aligned}} \quad (12.55)$$

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.

The field operators help solve a vexing problem in relativistic quantum mechanics; how to put space and time on equal footing as relativity needs. The classical Schrödinger equation $i\hbar\psi_t = H\psi$ 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 allows you to think of the time coordinate as simply a label on different spatial wave functions, one for each time, and application of the *spatial* Hamiltonian produces the change from one spatial wave function to the next one, a time dt later. Of course, you cannot think of the spatial coordinates in the same way; even if there was only one spatial coordinate instead of three: the second order spatial derivatives do not represent a change of wave function from one position to the next.

As section 10.2 discussed, for spinless particles, the Schrödinger equation can be converted into the Klein-Gordon equation, which turns the time derivative to second order by adding the rest mass energy to the Hamiltonian, and for electrons, the Schrödinger equation can be converted into the Dirac equation by switching to a vector wave function, which turns the spatial derivatives to first order. But there are problems; for example, the Klein-Gordon equation does not naturally conserve probabilities unless the solution is a simple wave; the Dirac equation has energy levels extending to minus infinity that must be thought of as being already filled with electrons to prevent an explosion of energy when the electrons fall down those states. Worse, filling the negative energy states would not help for bosons, since bosons do not obey the exclusion principle.

The field operators turn out to provide a better option, because they allow both the spatial coordinates and time to be converted into labels on annihilation and creation operators. It allows relativistic theories to be constructed that treat space and time in a uniform way.

12.2.6 An example using field operators

This example exercise from Srednicki [30, p. 11] compares quantum field theory to the classical formulation of quantum mechanics. The objective is to convert

the classical spatial Schrödinger equation for I particles,

$$i\hbar \frac{\partial \Psi}{\partial t} = \left[\sum_{i=1}^I \left(\frac{\hbar^2}{2m} \nabla_i^2 + V_{\text{ext}}(\vec{r}_i) \right) + \frac{1}{2} \sum_{i=1}^I \sum_{\underline{i}=1}^I V(\vec{r}_i - \vec{r}_{\underline{i}}) \right] \Psi \quad (12.56)$$

into quantum field form. The classical wave function has the positions of the numbered particles and time as arguments:

$$\text{classical quantum mechanics: } \Psi = \Psi(\vec{r}_1, \vec{r}_2, \vec{r}_3, \dots, \vec{r}_I; t) \quad (12.57)$$

where \vec{r}_1 is the position of particle 1, \vec{r}_2 is the position of particle 2, etcetera.

In quantum field theory, the wave function for exactly I particles takes the form

$$|\Psi\rangle = \int_{\text{all } \vec{r}_1} \dots \int_{\text{all } \vec{r}_I} \Psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_I; t) \hat{a}^\dagger(\vec{r}_1) \hat{a}^\dagger(\vec{r}_2) \dots \hat{a}^\dagger(\vec{r}_I) |\vec{0}\rangle d^3\vec{r}_1 \dots d^3\vec{r}_I \quad (12.58)$$

where ket $|\Psi\rangle$ is the wave function expressed in Fock space kets, and plain $\Psi(\dots)$ is to be shown to be equivalent to the classical wave function above. The Fock space Hamiltonian is

$$\begin{aligned} H &= \int_{\text{all } \vec{r}} \hat{a}^\dagger(\vec{r}) \left[-\frac{\hbar^2}{2m} \nabla_{\vec{r}}^2 + V_{\text{ext}}(\vec{r}) \right] \hat{a}(\vec{r}) d^3\vec{r} \\ &\quad + \frac{1}{2} \int_{\text{all } \vec{r}} \int_{\text{all } \vec{r}} \hat{a}^\dagger(\vec{r}) \hat{a}^\dagger(\vec{r}) V(\vec{r} - \vec{r}) \hat{a}(\vec{r}) \hat{a}(\vec{r}) d^3\vec{r} d^3\vec{r} \end{aligned} \quad (12.59)$$

It is to be shown that the Fock space Schrödinger equation for $|\Psi\rangle$ produces the classical Schrödinger equation (12.56) for function $\Psi(\dots)$, whether it is a system of bosons or 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

$$\Psi(x) = \int_{\text{all } \underline{x}} \underbrace{\Psi(\underline{x})}_{\text{coefficients}} \underbrace{\delta(x - \underline{x}) d\underline{x}}_{\text{basis states}}$$

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 $\vec{v} = v_1 \hat{i} + v_2 \hat{j} + v_3 \hat{k}$ or a vector in 30 dimensions as $\vec{v} = \sum_{i=1}^{30} v_i \hat{i}_i$. The $\Psi(\underline{x})$ are the coefficients, corresponding to the v_i -components of the vectors. The $\delta(x - \underline{x}) d\underline{x}$ are the basis states, just like the unit vectors \hat{i}_i . If you want a graphical illustration, each $\delta(x - \underline{x}) d\underline{x}$ would correspond to one spike of unit height at a position \underline{x} in figure 1.3, and you need to sum (integrate) over them all, with their coefficients, to get the total vector.

Now $H\Psi(x)$ is just another function of x , so it can be written similarly:

$$\begin{aligned} H\Psi(x) &= \int_{\text{all } \underline{x}} H\Psi(\underline{x})\delta(x - \underline{x}) d\underline{x} \\ &= \int_{\text{all } \underline{x}} \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(\underline{x})}{\partial \underline{x}^2} + V_{\text{ext}}(\underline{x})\Psi(\underline{x}) \right] \delta(x - \underline{x}) d\underline{x} \end{aligned}$$

Note that the Hamiltonian acts on the *coefficients*, not on the basis states. You may be surprised by this since if you straightforwardly apply the Hamiltonian, in terms of x , on the integral expression for $\Psi(x)$, you get

$$H\Psi(x) = \int_{\text{all } \underline{x}} \Psi(\underline{x}) \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \delta(x - \underline{x})}{\partial x^2} + V_{\text{ext}}(x)\delta(x - \underline{x}) \right] d\underline{x}$$

in which the Hamiltonian acts on the *basis states*, not the coefficients. However, the two expressions are indeed the same. (You can see that using a couple of integrations by parts in the latter, after recognizing that differentiation of the delta function with respect to x or \underline{x} is the same save for a sign change. Much better, make the change of integration variable $u = \underline{x} - x$ before applying the Hamiltonian to the integral.)

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. 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. First consider the expression for the wave function

$$|\Psi\rangle = \int_{\text{all } \vec{r}_1} \dots \int_{\text{all } \vec{r}_I} \underbrace{\Psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_I; t)}_{\text{coefficients}} \underbrace{\hat{a}^\dagger(\vec{r}_1)\hat{a}^\dagger(\vec{r}_2)\dots\hat{a}^\dagger(\vec{r}_I)|\vec{0}\rangle}_{\text{Fock space basis state kets}} d^3\vec{r}_1 \dots d^3\vec{r}_I$$

The ket $|\vec{0}\rangle$ is the vacuum state, but the preceding creation operators \hat{a}^\dagger create particles at positions $\vec{r}_1, \vec{r}_2, \dots$. So the net state becomes a Fock state where 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 $\Psi(\dots)$ determines the probability for the particles to actually be at those states, so it is the coefficient of that Fock state ket. The integration gives the combined wave function $|\Psi\rangle$ as a ket state in Fock space.

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 \hat{a}^\dagger creation operators flips the sign, compare subsection 12.2.2.) This will become important at the end of the derivation.

As far as understanding the Fock space Hamiltonian, for now you may just note a superficial similarity in form with the expectation value of energy. Its appropriateness will follow from the fact that the correct classical Schrödinger equation is obtained from it.

The left hand side of the Fock space Schrödinger equation is evaluated by pushing the time derivative inside the integral as a partial:

$$\mathrm{i}\hbar \frac{d|\Psi\rangle}{dt} = \int_{\text{all } \vec{r}_1} \dots \int_{\text{all } \vec{r}_I} \mathrm{i}\hbar \frac{\partial \Psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_I; t)}{\partial t} \hat{a}^\dagger(\vec{r}_1) \hat{a}^\dagger(\vec{r}_2) \dots \hat{a}^\dagger(\vec{r}_I) |\vec{0}\rangle d^3\vec{r}_1 \dots d^3\vec{r}_I$$

so the time derivative drops down on the classical wave function in the normal way.

Applying the Fock-space Hamiltonian (12.59) on the wave function is quite a different story, however. It is best to start with just a single particle:

$$H|\Psi\rangle = \int_{\text{all } \vec{r}} \int_{\text{all } \vec{r}_1} \hat{a}^\dagger(\vec{r}) \left[-\frac{\hbar^2}{2m} \nabla_{\vec{r}}^2 + V_{\text{ext}}(\vec{r}) \right] \hat{a}(\vec{r}) \Psi(\vec{r}_1; t) \hat{a}^\dagger(\vec{r}_1) |\vec{0}\rangle d^3\vec{r}_1 d^3\vec{r}$$

The field operator $\hat{a}(\vec{r})$ may be pushed past the classical wave function $\Psi(\dots)$; $\hat{a}(\vec{r})$ is defined by what it does to the Fock basis states while leaving their coefficients, here $\Psi(\dots)$, unchanged:

$$H|\Psi\rangle = \int_{\text{all } \vec{r}} \int_{\text{all } \vec{r}_1} \hat{a}^\dagger(\vec{r}) \left[-\frac{\hbar^2}{2m} \nabla_{\vec{r}}^2 + V_{\text{ext}}(\vec{r}) \right] \Psi(\vec{r}_1; t) \hat{a}(\vec{r}) \hat{a}^\dagger(\vec{r}_1) |\vec{0}\rangle d^3\vec{r}_1 d^3\vec{r}$$

It is now that the (anti)commutator relations become useful. The fact that for bosons $[\hat{a}(\vec{r})\hat{a}^\dagger(\vec{r}_1)]$ or for fermions $\{\hat{a}(\vec{r})\hat{a}^\dagger(\vec{r}_1)\}$ equals $\delta^3(\vec{r} - \vec{r}_1)$ means that you can swap the order of the operators as long as you add a delta function term:

$$\hat{a}_b(\vec{r})\hat{a}_b^\dagger(\vec{r}_1) = \hat{a}_b^\dagger(\vec{r}_1)\hat{a}_b(\vec{r}) + \delta^3(\vec{r} - \vec{r}_1) \quad \hat{a}_f(\vec{r})\hat{a}_f^\dagger(\vec{r}_1) = -\hat{a}_f^\dagger(\vec{r}_1)\hat{a}_f(\vec{r}) + \delta^3(\vec{r} - \vec{r}_1)$$

But when you swap the order of the operators in the expression for $H|\Psi\rangle$, you get a factor $\hat{a}(\vec{r})|\vec{0}\rangle$, and that is zero, because applying an annihilation operator on the vacuum state produces zero, figure 12.5. So the delta function is all that remains:

$$H|\Psi\rangle = \int_{\text{all } \vec{r}} \int_{\text{all } \vec{r}_1} \hat{a}^\dagger(\vec{r}) \left[-\frac{\hbar^2}{2m} \nabla_{\vec{r}}^2 + V_{\text{ext}}(\vec{r}) \right] \Psi(\vec{r}_1; t) \delta^3(\vec{r} - \vec{r}_1) |\vec{0}\rangle d^3\vec{r}_1 d^3\vec{r}$$

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

$$H|\Psi\rangle = \int_{\text{all } \vec{r}} \hat{a}^\dagger(\vec{r}) \left[-\frac{\hbar^2}{2m} \nabla_{\vec{r}}^2 + V_{\text{ext}}(\vec{r}) \right] \Psi(\vec{r}; t) |\vec{0}\rangle d^3\vec{r}$$

The creation operator $\hat{a}^\dagger(\vec{r})$ can be pushed over the coefficient $H\Psi(\vec{r}; t)$ of the vacuum state ket for the same reason that $\hat{a}(\vec{r})$ 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, renimating \vec{r} to \vec{r}_1 , the grand total Fock state Schrödinger equation for a system of one particle becomes

$$\begin{aligned} \int_{\text{all } \vec{r}_1} i\hbar \frac{\partial \Psi(\vec{r}_1; t)}{\partial t} \hat{a}^\dagger(\vec{r}_1) |\vec{0}\rangle d^3\vec{r}_1 = \\ \int_{\text{all } \vec{r}_1} \left[-\frac{\hbar^2}{2m} \nabla_{\vec{r}_1}^2 + V_{\text{ext}}(\vec{r}_1) \right] \Psi(\vec{r}_1; t) \hat{a}^\dagger(\vec{r}_1) |\vec{0}\rangle d^3\vec{r}_1 \end{aligned}$$

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-state 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.

If there is more than one particle, however, the 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, $a\hat{i} = a'\hat{i}$ implies that $a = a'$, but no, $(a + b)\hat{i} = (a' + b')\hat{i}$ 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, $(a + a)\hat{i} = (a' + a')\hat{i}$ does imply that $a = a'$.

Operator swapping like in the derivation above also helps to understand why the Fock-space Hamiltonian has an appearance similar to an energy expectation value. For example, consider the effect of placing the one-particle Hamiltonian between $\langle \vec{0} | \hat{a}(\vec{r}_1) \Psi^*(\vec{r}_1; t) \rangle$ and $\langle \Psi(\vec{r}_1; t) \hat{a}^\dagger(\vec{r}_1) | \vec{0} \rangle$ and integrating over all \vec{r}_1 and \vec{r}_1 .

So 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.

Chapter 13

The Interpretation 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, a Caltech quantum physicist who received a Nobel Prize for the creation of quantum electrodynamics with Schwinger and Tomonaga, and who pioneered nanotechnology with his famous talk “There’s Plenty of Room at the Bottom,” 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) 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.

The sections of this chapter are based on the general ideas of quantum mechanics as discussed in part I of this book.

13.1 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 13.5, 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.

13.2 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 13.1. You might send one proton off to your



Figure 13.1: 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 figure 13.2:



Figure 13.2: 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 13.2. 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:

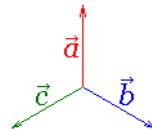


Figure 13.3: Spin measurement directions.

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, \dots . 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 *subluminal 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 show 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 13.4 shows how a Bohm experiment appears to an observer on earth. The spins



Figure 13.4: 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 13.4, {A.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 13.5. 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:

$$|2(f_3 + f_4 + f_5 + f_6) - 2(f_2 + f_4 + f_5 + f_7)| \leq 2(f_2 + f_3 + f_6 + f_7)$$

So the discussion cheated. And Bell allowed general directions of measurement

not just 120 degree ones. See [17, pp. 423-426]. The above discussion seems a lot less messy, even though not historically accurate.

13.3 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 13.1, 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 chapter 12.2.1, the idea has become a fundamental tenet in quantum field theory.

13.4 Failure of the Schrödinger Equation?

Section {13.2} mentioned sending half of the wave function of an electron to Venus, and half to Mars. A scattering setup as described in chapter 6.8 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, …, with a supercolossal wave function:

$$\Psi(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \vec{r}_3, S_{z3}, \vec{r}_4, S_{z4}, \vec{r}_5, S_{z5}, \dots)$$

Particle 1 will be taken to be the scattered electron. It will be assumed that the wave function satisfies the Schrödinger equation:

$$i\hbar \frac{\partial \Psi}{\partial t} = - \sum_i \sum_{j=1}^3 \frac{\hbar^2}{2m_i} \frac{\partial^2 \Psi}{\partial r_{i,j}^2} + V(\vec{r}_1, S_{z1}, \vec{r}_2, S_{z2}, \vec{r}_3, S_{z3}, \vec{r}_4, S_{z4}, \dots) \Psi \quad (13.1)$$

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:

$$i\hbar \frac{\partial \Psi^* \Psi}{\partial t} = - \sum_i \sum_{j=1}^3 \frac{\hbar^2}{2m_i} \frac{\partial}{\partial r_{i,j}} \left(\Psi^* \frac{\partial \Psi}{\partial r_{i,j}} - \Psi \frac{\partial \Psi^*}{\partial r_{i,j}} \right) \quad (13.2)$$

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 non-conserved 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.

13.5 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 13.2, where a positron is sent to Venus and an entangled electron to Mars, as in figure 13.5. The spin states are



Figure 13.5: Bohm's version of the Einstein, Podolski, Rosen Paradox.

uncertain when the two are send 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 4.5.6. 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 13.6. You would still not want to write down the supercolossal



Figure 13.6: Non entangled 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 13.7. It will produce a wave function of everything describing that the observer on



Figure 13.7: Non entangled positron and electron spins; down and up.

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 non entangled ones above, as shown in figure 13.8. If the wave function in each non entangled case describes

$$\text{Venus} \quad \text{Mars}$$

	$\frac{1}{\sqrt{2}}$	
	$-\frac{1}{\sqrt{2}}$	

Figure 13.8: The wave functions of two universes combined

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:

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 send from earth. In the other universe, it was send out as up. Similarly, for the case of the last subsection, where half the wave function of an electron was send 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 send 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 13.9. 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 good record of what was found. Figure 13.10 shows what happens.

When the electron is send from Earth, two universes can be distinguished,

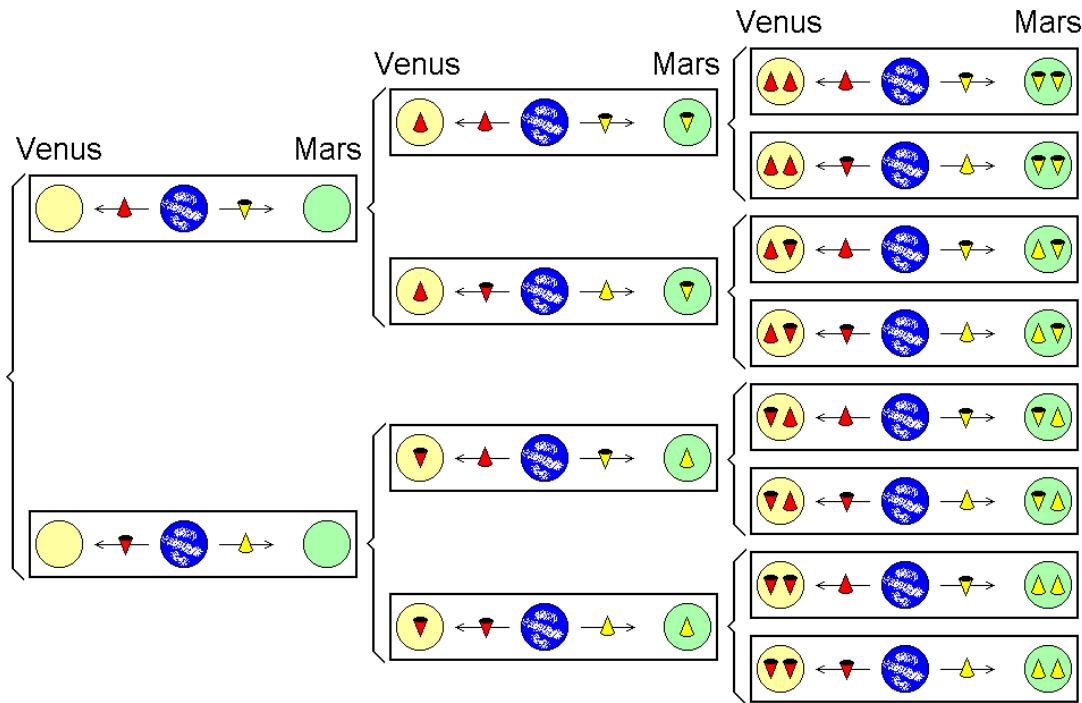


Figure 13.9: The Bohm experiment repeated.

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 13.10 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

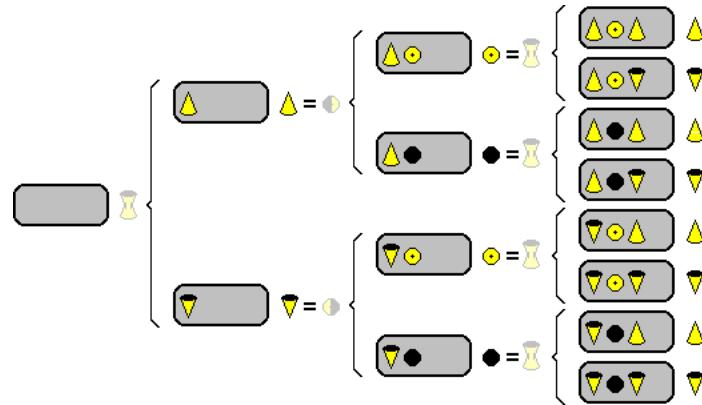


Figure 13.10: Repeated experiments on the same electron.

are summed together, the net spin is still down (the two down spins have equal 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 6.8.1 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 {A.119}.

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.

13.6 The Arrow of Time

This section has some further musings on the many worlds interpretation.

The many worlds interpretation is unscientific in the sense that it is not empirically testable. The linearity of the quantum equations makes it so. By the same token, neither are the competing theories scientific. Maybe that is good news: if you do not want to believe in alternate realities, you can do so without fear of being proven wrong.

The main reason to regard the many worlds interpretation as the most plau-

sible one is that it meets Occam's razor. While some authors have argued that the introduction of infinitely many worlds hardly does that, they are mistaken. Their mistake is that it is not the many worlds interpretation, but the fundamental concepts of quantum mechanics that introduce the infinitely many worlds. The Copenhagen interpretation simply adds the additional assumption that the wave function amplitude for all but one single-observer point of view is somehow made very small.

Wikipedia (9/2009) offers the following description of Occam's razor:

*Occam's razor or Ockham's razor, attributed to 14th-century English logician and Franciscan friar, William of Ockham is the principle that "entities should not be multiplied unnecessarily" or, popularly applied, "when you have two competing theories that make exactly the same predictions, the simpler one is the better." The principle states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory. The principle is often expressed in Latin as the *lex parsimoniae* ("law of parsimony", "law of economy", or "law of succinctness"): *entia non sunt multiplicanda praeter necessitatem*, roughly translated as "entities must not be multiplied beyond necessity." An alternative version *Pluralitas non est ponenda sine necessitate* translates "plurality should not be posited without necessity." [[18, 9/09]]*

The many worlds interpretation uses only the solidly established, fully quantitative, Schrödinger equation, or its relativistic version, that has been tested to great accuracy. It explains the experimental observations logically and comprehensively based on that single theory.

The main competing theory, the Copenhagen interpretation, must add a further massive machinery, wave function collapse, to get rid of the quantum amplitude of all observer views besides one. The required additional effort is as enormous as the infinitely many "worlds" (observer views) that the Schrödinger equation creates continuously when left to itself. Is there any proof that this additional massive machinery exists? No, because the Schrödinger equation already explains the observations fully. The only thing that the collapse mechanism adds is to allow human beings to be right if they want to feel unique. Maybe there is indeed a God somewhere out there who worries about that. A Deus ex Machina would certainly explain a few puzzling points. Such as why there are no equations, let alone empirically verified ones like the Schrödinger equation, for the collapse. In fact, the "collapse" itself is so vaguely and inconsistently defined that even if there were equations, they could not possibly be used to verify or refute it.

However unprovable the competing theories may be, it is still interesting to see how their ideas compare. One theory might provide an explanation for certain phenomena that may seem more plausible than the other, even if neither theory can be rigorously refuted.

So what is the difference in thinking? 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 point when considering "virtual particles," particles that have a fleeting existence. Relativistic quantum mechanics shows that normal matter constantly interacts with large amounts of virtual particles. They determine the interactions and decay processes of matter. The problem is that there are infinitely many of these virtual particles. Even if their number is arbitrarily assumed to be limited by the smallest quantum scales, there still are enormously many of them present at any given time. Their gravitational effect should be gigantic, it should dwarf anything else. Somehow that does not happen. In the Copenhagen interpretation, there is no self-evident explanation. A collapse mechanism provides no straightforward reason to exclude virtual particles from producing gravitational effects. However, in Everett's theory a particle only becomes real when a universe is established in which there is no doubt that the particle exists. Virtual particles cannot meet that requirement because they disappear and get replaced by another one before such a process can take place.

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, which says that a measurable definition of disorder in the observed universe, called entropy, always increases with time. But 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 Copenhagen interpretation answers that by postulating ad-hoc asymmetries in the initial data that "just happened" to be that way. Then it adds evolution equations that magnify that asymmetry. This might at first seem fairly plausible to a reasonable person. However, if you take a closer look at those evolution equations, it becomes a lot less plausible. The mechanism that provides the increase in disorder with time is, you guessed it, exactly that poorly defined collapse mechanism. The collapse mechanism is molded in a form to provide the observed increase. It is easy to simply say that something with no testable definition does exactly what you would like it to do.

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 any better than the other. So how plausible is it really that nature would suddenly apparently change its mind and create initial asymmetries and evolution equations that would leave only one arbitrarily chosen solution?

According to Everett's theory, there is zero reason to assume 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. And reversal of the direction of time brings in a preference for antimatter versus matter, and a preference for the opposite handedness. So these components of the global wave function going backwards compared to our time end up observing a universe composed of antimatter, with left-handed biases changed into right-handed ones and vice versa.

It is interesting to learn that string theorists have been exploring gravity as an expression of thermodynamics. The above arguments do much the same based on classical quantum mechanics; they make both the second law and the magnitude of gravity a consequence of the limited part of the wave function that evolves into existing for an observer.

Perhaps, if we spend less time on developing crackpot theories for what we *want* to believe, and more time on listening to what nature is really telling us, we would now understand those processes a lot more clearly.

Appendix A

Notes

The notes in this section give background information, various derivations of claims made, and other material that is not essential to understand quantum mechanics. Use it when curious, or when a ambiguous issue arises.

A.1 Why another book on quantum mechanics?

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 avoids the highly abstract Kronig-Penney (Heaviside functions) or Dirac combs (delta functions) mathematical models 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.

A.2 History and wish list

- Oct 24, 2004. The first version of this manuscript was posted.
- 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.
- 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.
- 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.
- 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.
- Mid April, 2006. Various minor fixes. Also I changed the format from the “article” to the “book” style.
- 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 Feb., 2007. There are now lists of key points and review questions for chapter 1. Answers are in the new solution manual.

- 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.
 - 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.
 - 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.
 - 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 4.7 and subsection 4.9.2 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.
- 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.

- 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.
- 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.
- 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.
- 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.
- 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 I. 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 non-delta 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 A.99 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.
- 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.

- 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 I. Much of this chapter has been scavenged from Part II. 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 II in a single semester. The semiconductor discussion in the chapter is all new.
- July 13, 2010. Version 5.03 alpha. Various rewrites and error corrections. Also a new periodic table.
- July 16, 2010. Version 5.04 alpha. Some rewrites and error corrections.
- July 19, 2010. Version 5.05 alpha. Some error corrections in thermoelectrics and a discussion of the Onsager relations added.
- July 23, 2010. Version 5.06 alpha. Some rewrites and error corrections in thermoelectrics.
- July 28, 2010. Version 5.07 alpha. Better description of the third law. Some rewrites and error corrections in thermoelectrics.
- 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.

Part I is mostly in a fairly good shape. But there are a few recent additions that probably could do with another look.

In Part II various sections sure could do with a few more rewrites. I am expanding the quantity much faster than the quality at the time of this writing. The chapter on nuclei is an incomplete mess. I need a sabbatical.

Somewhat notably missing at this time:

1. Electron split experiments. Do engineers need it??
2. Quantum electrodynamics. Do engineers need it??
3. Density-functional theory.

4. Mössbauer effect.
5. Superfluidity (but there is not really a microscopic theory.)
6. Superconductivity.

Density-functional theory will eventually be added. How old are you, and how is your health?

The index is in a sorry shape. It is being worked on. Once in a while.

I would like to add key points and review questions to all basic sections. I am inching up to it. Very slowly.

After that, the idea is to run all this text through a style checker to eliminate the dead wood. Also, ispell seems to be missing misspelled words. Probably thinks they are TeX.

It would be nice to put frames around all key formulae. Many are already there.

Need to have figure 6.2 of Yariv and explain angular momentum states are assumed equivalent if space has no preferred direction.

A.3 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.3.1 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

$$\mathcal{L} = \sum_j \frac{1}{2} m_j |\vec{v}_j|^2 - V(\vec{r}_1, \vec{r}_2, \dots)$$

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” $p_{j,i}^c$ of particle j in the i direction, (with $i = 1, 2$, or 3 for the x , y , or z component respectively), is defined as

$$p_{j,i}^c \equiv \frac{\partial \mathcal{L}}{\partial v_{j,i}}$$

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

$$\frac{dp_{j,i}^c}{dt} = \frac{\partial \mathcal{L}}{\partial r_{j,i}}$$

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.3.2 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.3.3 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, \dots and their time derivatives $\dot{q}_1, \dot{q}_2, \dots$. 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:

$$\begin{aligned} \mathcal{L}(q_1, q_2, \dots, q_K, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_K, t) \\ \equiv T(q_1, q_2, \dots, q_K, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_K, t) - V(q_1, q_2, \dots, q_K, t) \end{aligned}$$

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

$$p_k^c \equiv \frac{\partial \mathcal{L}}{\partial \dot{q}_k} \tag{A.1}$$

for each individual generalized coordinate q_k . The equations of motion are

$$\frac{dp_k^c}{dt} = \frac{\partial \mathcal{L}}{\partial q_k} + Q_k \quad (\text{A.2})$$

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 unifying concept is that of “action,” defined as the time integral of the Lagrangian. The action integral is unchanged by infinitesimal temporary displacements of the system, and that is all that is needed for the Lagrangian equations of motion to apply.

Derivation

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

$$m_j \frac{d^2\vec{r}_j}{dt^2} = -\frac{\partial V}{\partial \vec{r}_j} + \vec{F}'_j$$

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:

$$0 = \int_{t_1}^{t_2} \sum_j \left[-m_j \frac{d^2\vec{r}_j}{dt^2} - \frac{\partial V}{\partial \vec{r}_j} + \vec{F}'_j \right] \cdot \delta\vec{r}_j dt$$

Multiply out and integrate the first term by parts:

$$0 = \int_{t_1}^{t_2} \sum_j \left[m_j \frac{d\vec{r}_j}{dt} \cdot \delta \frac{d\vec{r}_j}{dt} - \frac{\partial V}{\partial \vec{r}_j} \cdot \delta\vec{r}_j + \vec{F}'_j \delta\vec{r}_j \right] dt$$

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

$$0 = \delta \int_{t_1}^{t_2} \mathcal{L} dt + \int_{t_1}^{t_2} \sum_j [\vec{F}'_j \cdot \delta\vec{r}_j] dt \quad (\text{A.3})$$

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 $\delta \vec{r}_j$:

$$0 = \int_{t_1}^{t_2} \left[\frac{\partial \mathcal{L}}{\partial q_k} \delta q_k + \frac{\partial \mathcal{L}}{\partial \dot{q}_k} \delta \dot{q}_k \right] dt + \int_{t_1}^{t_2} \sum_j \left[\vec{F}'_j \cdot \frac{\partial \vec{r}_j}{\partial q_k} \delta q_k \right] dt$$

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

$$0 = \int_{t_1}^{t_2} \left[\frac{\partial \mathcal{L}}{\partial q_k} - \frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{q}_k} \right) + Q_k \right] \delta q_k dt$$

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.

A.3.4 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:

$$\frac{dq_k}{dt} = \dot{q}_k$$

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

$$H(q_1, q_2, \dots, q_K, p_1^c, p_2^c, \dots, p_K^c, t) \equiv \sum_{k=1}^K \dot{q}_k p_k^c - \mathcal{L}(q_1, q_2, \dots, q_K, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_K, t) \quad (\text{A.4})$$

In the right hand side expression, you must rewrite all the time derivatives \dot{q}_k in terms of the canonical momenta

$$p_k^c \equiv \frac{\partial \mathcal{L}}{\partial \dot{q}_k}$$

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

$$\frac{dq_k}{dt} = \frac{\partial H}{\partial p_k^c} \quad \frac{dp_k^c}{dt} = -\frac{\partial H}{\partial q_k} + Q_k \quad (\text{A.5})$$

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.

Derivation

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 $d\mathcal{L}$,

$$dH = \left(\sum_k p_k^c d\dot{q}_k \right) + \sum_k \dot{q}_k dp_k^c - \sum_k \frac{\partial \mathcal{L}}{\partial q_k} dq_k - \left(\sum_k \frac{\partial \mathcal{L}}{\partial \dot{q}_k} d\dot{q}_k \right) - \frac{\partial \mathcal{L}}{\partial t} dt$$

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,

$$\frac{\partial H}{\partial q_k} = -\frac{\partial \mathcal{L}}{\partial \dot{q}_k} \quad \frac{\partial H}{\partial p_k^c} = \dot{q}_k \quad \frac{\partial H}{\partial t} = -\frac{\partial \mathcal{L}}{\partial t}$$

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

$$\sum_j \sum_{\underline{k}=1}^K \sum_{\underline{k}=1}^K \frac{1}{2} m_j \frac{\partial \vec{r}_j}{\partial q_{\underline{k}}} \cdot \frac{\partial \vec{r}_j}{\partial q_{\underline{k}}} \dot{q}_{\underline{k}} \dot{q}_{\underline{k}} - V(q_1, q_2, \dots, q_K, t)$$

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.

A.4 Special relativity

Special relativity tends to keep popping up in quantum mechanics. This note gives a brief summary of the relevant points.

A.4.1 History

Special relativity is commonly attributed to Albert Einstein's 1905 papers, even though Einstein swiped the big ideas of relativity from Henri Poincaré, (who developed and named the principle of relativity in 1895 and the 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 work appeared in the same journal as the one in which Einstein published his papers on relativity. 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 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 clearly be credited to Einstein fair and square. But he was a lot less hungry then. (And had a lot more false starts.)

A.4.2 Overview of relativity

The most important result of relativity for this book is Einstein's famous relation $E = mc^2$, where E is energy, m mass, and c the speed of light. (To be precise, this relation traces back to Poincaré, but Einstein generalized the idea.) The kinetic energy of a particle is not $\frac{1}{2}mv^2$, with m the mass and v the velocity, as Newtonian physics says. Instead it is the difference between the energy $m_v c^2$ based on the mass m_v of the particle in motion and the energy $m_0 c^2$ based on the mass m_0 of the particle at rest. According to special relativity the mass in motion is

$$m_v = \frac{m_0}{\sqrt{1 - (v/c)^2}} \quad (\text{A.6})$$

so the true kinetic energy is

$$T = \frac{m_0}{\sqrt{1 - (v/c)^2}} c^2 - m_0 c^2$$

For velocities small compared to the tremendous speed of light, this is equivalent to the classical $\frac{1}{2}m_0 v^2$; 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 only reason that the photons of electromagnetic radiation, (including radio waves, microwaves, light, x-rays, gamma rays, etcetera), can travel at the speed of light is because they have zero rest mass m_0 ; there is no way that they can be brought to a halt, because there would be nothing left.

Quantum mechanics does not use the speed v but the momentum $p = m_v v$; in those terms the square root can be rewritten to give the kinetic energy as

$$T = m_0 c^2 \sqrt{1 + \frac{p^2}{m_0^2 c^2}} - m_0 c^2 \quad (\text{A.7})$$

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

Note that it suggests that a particle at rest still has a “rest mass energy” $m_0 c^2$ 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 difference between the rest masses of two atomic nuclei gives the energy released in nuclear reactions that convert one nucleus into the other.

Another weird relativistic effect is that the speed of light is the same regardless of how fast you are travelling. 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. Michelson & Morley figured that the speed of the earth could similarly be measured 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.

You can think up a hundred lame excuses. (In particular, the sound *inside* a plane does not move 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 in empty space like light.) Or you can be honest. In 1895 Poincaré reasoned that such experiments suggested that it seems to be impossible to detect the absolute motion of matter. In 1900 he proposed the "Principle of Relative Motion," that the laws of movement should be the same in all coordinate systems regardless of their velocity, as long as they are not accelerating. 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 a constant speed relative to each other, it is impossible to say which one, if any, is at rest. (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.)

All this pops up in the thought example of figure 13.4 where an electron is sent at very nearly the speed of light to Mars and a positron in the other

direction to Venus. As far as an observer on Earth is concerned, the electron and positron reach their destinations at almost the same time. For a observer traveling in the direction from Venus to Mars at almost the speed of light, the electron and positron still seem to be going at close to the speed of light, if their kinetic energies are high enough. However, for that observer, it appears that Venus is moving at close to the speed of light away from its positron, while Mars is moving at the same speed towards its electron. Therefore it appears to that observer that the electron reaches Mars much earlier than the positron reaches Venus.

Obviously then, 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 interval between two events slows down according to

$$\Delta t_v = \frac{\Delta t_0}{\sqrt{1 - (v/c)^2}} \quad (\text{A.8})$$

where Δt_v is shorthand for the time interval between the events as perceived by an observer compared to whom the location of the events is moving at speed v , while Δt_0 is the same time interval as perceived by an observer compared to whom the location is at rest.

For example, cosmic rays can create radioactive particles in the upper atmosphere that reach the surface of the earth, even though in a laboratory they do not last long enough to do so by far. Because of the high speed of these particles, for an observer standing on earth the decay process seems to take much longer than normal. Time dilation in action.

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

$$\Delta x_v = \Delta x_0 \sqrt{1 - (v/c)^2} \quad (\text{A.9})$$

where the x -axis is taken to be in the direction of motion and Δx_0 is the distance

in x -direction as perceived by an observer compared to which the object is at rest.

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.”

A.4.3 Lorentz transformation

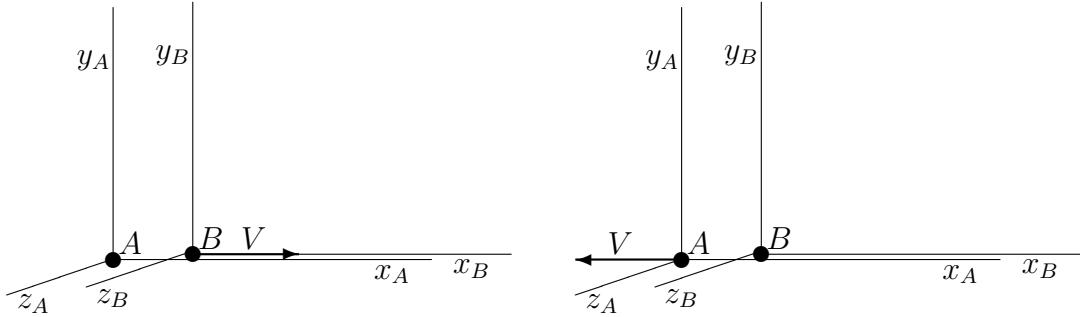


Figure A.1: Coordinate systems for the Lorentz transformation.

The “Lorentz transformation” describes exactly how measurements of the position and time of events change from one observer to the next. Consider two observers A and B that are in motion compared to each other with a relative speed V . To relate their time and position measurements, it is convenient to give each observer its own coordinate system x, y, z, t as shown in figure A.1. Each coordinate system will be taken to have its origin at the location of the observer, while both x -axes are along the line from A to B , both y -axes are parallel, and both z -axes are parallel. As the left side of figure A.1 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.

The Lorentz transformation says that the relation between positions and times of events as perceived by the two observers is

$$ct_B = \frac{ct_A - (V/c)x_A}{\sqrt{1 - (V/c)^2}} \quad x_B = \frac{x_A - (V/c)ct_A}{\sqrt{1 - (V/c)^2}} \quad y_B = y_A \quad z_B = z_A \quad (\text{A.10})$$

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 A.1. In the limit that 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 .

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

$$v_{x,B} = \frac{v_{x,A} - V}{1 - (V/c^2)v_{x,A}} \quad v_{y,B} = \frac{v_{y,A}\sqrt{1 - (V/c)^2}}{1 - (V/c^2)v_{x,A}} \quad v_{z,B} = \frac{v_{z,A}\sqrt{1 - (V/c)^2}}{1 - (V/c^2)v_{x,A}} \quad (\text{A.11})$$

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.

Derivation

The Lorentz transformation can be derived by assuming that the transformation from the coordinates t_A, x_A, y_A, z_A to t_B, x_B, y_B, z_B is linear;

$$\begin{aligned} t_B &= a_{tx}x_A + a_{ty}y_A + a_{tz}z_A + a_{tt}t_A & y_B &= a_{yx}x_A + a_{yy}y_A + a_{yz}z_A + a_{yt}t_A \\ x_B &= a_{xx}x_A + a_{xy}y_A + a_{xz}z_A + a_{xt}t_A & z_B &= a_{zx}x_A + a_{zy}y_A + a_{zz}z_A + a_{zt}t_A \end{aligned}$$

where the $a_{..}$ are constants still to be found. The big 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_B, y_B, z_B, t_B back to x_A, y_A, z_A, t_A should be similar to 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. Note that since the choice what to define as time zero and as the origin is quite arbitrary, it can be arranged that x_B, y_B, z_B, t_B are zero when x_A, y_A, z_A, t_A are.

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_B, t_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_A from the equation for y_B . Similarly, there goes y_A in the equation for z_B . Since the choice of y and z -axes is arbitrary, the remaining a_z coefficients must equal the corresponding a_y 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 , $x_B = y_B = z_B = 0$ should correspond to $x_A = vt_A$. And by the principle of relativity, $x_A = y_A = z_A = 0$ should correspond to $x_B = -vt_B$. You might be able to think up some more constraints, but this will do. Put it all together to get

$$\begin{aligned} t_B &= a_{tx}x_A + a_{xx}t_A & y_B &= a_{yx}x_A + y_A + a_{yt}t_A \\ x_B &= a_{xx}(x_A - vt_A) & z_B &= a_{yx}x_A + z_A + a_{yt}t_A \end{aligned}$$

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 :

$$x_A^2 + y_A^2 + z_A^2 = c^2t_A^2 \quad x_B^2 + y_B^2 + z_B^2 = c^2t_B^2$$

Plug the linearized expressions for x_B, y_B, z_B, t_B in terms of x_A, y_A, z_A, t_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_A, y_A, z_A, t_A in terms of x_B, y_B, z_B, t_B , solve the Lorentz equations for x_A, y_A, z_A , and t_A .

Now assume that two events 1 and 2 happen at the same location x_A, y_A, z_A in system A , then the Lorentz transformation formula (A.10) giving t_B implies the time dilation for the events seen in the coordinate system B . Next assume that two stationary locations in system B are apart by a distance $x_{B,2} - x_{B,1}$ in the direction of relative motion; then the Lorentz transformation formula giving x_B implies that seen in the A system, these now moving locations are at any single time t_A apart by a distance reduced by the Lorentz-Fitzgerald contraction. Take differentials of the Lorentz transformation formulae to derive the given transformations between the velocities seen in the two systems.

A.4.4 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 interval 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 Δs between two events that occur at the same time:

nonrelativistic: Δs is independent of the observer if $\Delta t = 0$

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

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

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 Δs traveled between ticks, (as they would do nonrelativistically), but they also disagree about the time interval 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 interval is called the “proper time” interval Δt_0 . It is shorter than the time interval that an observer actually perceives due to the time dilation:

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

where v is the velocity of the clock as perceived by the observer. To clean this up, take the square root to the other side, move the time interval inside it, multiply by c to get rid of the ratio, and then recognize $v\Delta t$ as the distance traveled. That gives the proper time interval Δt_0 as

$$c\Delta t_0 = \sqrt{(c\Delta t)^2 - (\Delta x)^2 - (\Delta y)^2 - (\Delta z)^2} \quad (\text{A.12})$$

The scaled proper time interval $c\Delta t_0$ is called the “space-time interval,” because it involves both space and time. In particular, it involves the spatial distance between the events too. Note however that the interval 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 space-time interval between those two events is imaginary. To avoid dealing with complex numbers, it is then more convenient to define the “proper distance” between two events as

$$\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 - (c\Delta t)^2} \quad (\text{A.13})$$

If the space-time interval, or proper time interval, is real, it is called “time-like.” If it is imaginary, so that the proper distance is real instead, the space-time interval is called “space-like.” For vanishingly small differences in time and location, all differences Δ become differentials d .

A.4.5 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 tele-communications 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 (A.10), in the coordinate system fixed to the earth, the beam reaches the location of the accident at a position and time equal to

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

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, and at that stage the time interval would have to be imaginary instead of real. 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.

A.4.6 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 to be 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

$$\overleftrightarrow{r} \equiv \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \equiv \begin{pmatrix} r_0 \\ r_1 \\ r_2 \\ r_3 \end{pmatrix} \quad (\text{A.14})$$

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:

$$\overleftrightarrow{r}_1 \cdot \overleftrightarrow{r}_2 \equiv -c^2 t_1 t_2 + x_1 x_2 + y_1 y_2 + z_1 z_2 \quad (\text{A.15})$$

It can be checked by simple substitution that the Lorentz transformation (A.10) 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

$$\Delta s = \sqrt{(\Delta \overleftrightarrow{r}) \cdot (\Delta \overleftrightarrow{r})} \quad (\text{A.16})$$

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 convention, which is still the limit case for velocities small compared to the speed of light. 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.

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.

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.

A.4.7 Index notation

The notations used in the previous subsection are non 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, physicists like to indicate the components of four vectors by x_0, x_1, x_2, x_3 because it is inconsistent with the non relativistic convention. Also, since the letter x is already greatly over-used as it is, it promotes confusion, something that is always hilarious. A generic component may be denoted as x_μ , and an entire four vector can then be indicated by $\{x_\mu\}$ where the brackets indicate the set of all four components. Needless to say, some physicists forget about the brackets, because using a component where a vector is required can have hilarious consequences.

Physicists also like to put the coefficients of the Lorentz transformation (A.10) into a table called a “matrix” or “second-order tensor,” as follows

$$\Lambda \equiv \begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} & \Lambda_{14} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} & \Lambda_{24} \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} & \Lambda_{34} \\ \Lambda_{41} & \Lambda_{42} & \Lambda_{43} & \Lambda_{44} \end{pmatrix} \equiv \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{aligned} \gamma &\equiv \frac{1}{\sqrt{1 - (V/c)^2}} \\ \beta &\equiv \frac{V}{c} \\ \gamma^2 - \beta^2\gamma^2 &= 1 \end{aligned} \tag{A.17}$$

The Lorentz matrix as shown assumes that the axis systems are aligned with the direction of relative motion of the observers. Otherwise, it becomes a lot more messy.

In terms of those notations, the Lorentz transformation (A.10) can be written as

$$x_{\mu,B} = \sum_{\nu=0}^3 \Lambda_{\mu\nu} x_{\nu,A} \quad \text{for all values } \mu = 0, 1, 2, 3$$

The “Einstein summation convention” is now to leave out the sum. Summation is understood to be required whenever the same index appears twice in an expression. So, you will likely find the Lorentz transformation written more concisely as

$$x_{\mu,B} = \Lambda_{\mu\nu} x_{\nu,A}$$

where some of these subscripts may actually appear as superscripts. The basic reason for raising indices is that a quantity like a position differential transforms differently than a quantity like a gradient,

$$dx_{\mu,B} = \frac{\partial x_{\mu,B}}{\partial x_{\nu,A}} dx_{\nu,A} \quad \frac{\partial f}{\partial x_{\mu,B}} = \frac{\partial f}{\partial x_{\nu,A}} \frac{\partial x_{\nu,A}}{\partial x_{\mu,B}}$$

and raising or lowering indices is a means of keeping track of that.

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.

A.4.8 Group property

The derivation of the Lorentz transform 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,

$$\overleftrightarrow{r}_C = \Lambda_{C \leftarrow B} \overleftrightarrow{r}_B = \Lambda_{C \leftarrow B} \Lambda_{B \leftarrow A} \overleftrightarrow{r}_A$$

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.

Derivation

This subsubsection verifies the group property of the Lorentz transformation. It is not recommended unless you have had a solid course in linear algebra.

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 (A.17) 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

$$\Lambda_{C \leftarrow A} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} & 0 \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} & 0 \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

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 Λ_{31} and the spatial coordinate system of observer A can be rotated to eliminate Λ_{13} . 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 Λ_{23} must be zero, because Λ_{21} 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 Λ_{33} 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 Λ_{32} .

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,

$$-\Lambda_{11}\Lambda_{12} + \Lambda_{21}\Lambda_{22} = 0 \quad \Rightarrow \quad \frac{\Lambda_{12}}{\Lambda_{22}} = \frac{\Lambda_{21}}{\Lambda_{11}} \equiv \beta$$

where β is some constant. Also, since the proper lengths of these vectors are minus one, respectively one,

$$-\Lambda_{11}^2 + \Lambda_{21}^2 = -1 \quad -\Lambda_{12}^2 + \Lambda_{22}^2 = 1$$

or substituting in for Λ_{12} and Λ_{21} from the above

$$-\Lambda_{11}^2 + \beta^2\Lambda_{11}^2 = -1 \quad -\beta^2\Lambda_{22}^2 + \Lambda_{22}^2 = 1$$

It follows that Λ_{11} and Λ_{22} 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 (A.17). (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 $x_A = y_A = z_A = 0$ 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.

A.4.9 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.

In the quantum solution of the hydrogen atom, an electrostatic potential is fine when the heavy proton can be assumed to be at rest. However, an observer compared to who the entire atom is in high speed motion does *not* see the electron as moving in an electrostatic field. As far as that observer is concerned, the moving proton also creates a magnetic field. Indeed, for moving charges, electrostatics turns into electromagnetics, and the combined electromagnetic field *does* obey the limitation of the speed of light, as Maxwell's equations show.

But in Maxwell's equations, the speed of light limitation still seems almost accidental; it comes out of the theory, it is not build into it. It is therefore maybe not a coincidence that in the more complicated theory of quantum electrodynamics, electromagnetic interactions end up being described as mediated by a particle, the photon. Collisions between particles inherently avoid erroneous action at a distance, especially if those particles have some uncertainty in position and time. Taking a hint from that, this introduction will restrict itself to collisions between particles, [14]. It allows simple dynamics to be done without the use of a potential between particles that is relativistically suspect.

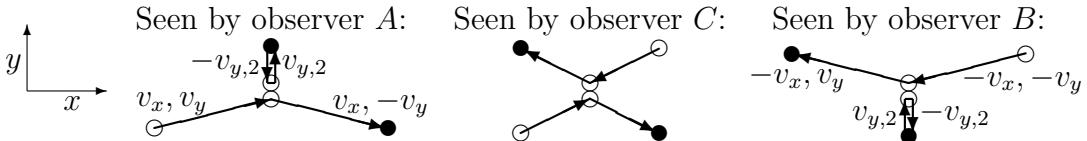


Figure A.2: Example elastic collision seen by different observers.

As a first example, consider two particles of equal mass and opposite speeds that collide as shown in the center of figure A.2. You might think of the particles as two helium atoms. It will be assumed that while the speed of the atoms may 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 (A.11) 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:

$$\vec{p} = m \frac{d\vec{r}}{dt} \quad \Rightarrow \quad \begin{cases} \overset{\leftrightarrow}{p} = m \frac{d\overset{\leftrightarrow}{r}}{dt} ? \\ \overset{\leftrightarrow}{p} = m \frac{d\overset{\leftrightarrow}{r}}{dt_0} ? \end{cases}$$

Indeed, nonrelativistically, all observers agree about time intervals. However, relativistically the question arises whether the right time differential in momentum is dt as perceived by the observer, or the proper time difference dt_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 dt_0 . For, after collisions the sum of the momenta should be the same as before them. However, the Lorentz velocity transformation (A.11) 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 dt_0 transforms like $\vec{d\tilde{r}}$, 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 $j = 1, 2, \dots$ is the same before and after,

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

then so does any other observer B :

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

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

$$p_0 = mc \frac{dt}{dt_0} \quad p_1 = m \frac{dt}{dt_0} \frac{dx}{dt} \quad p_2 = m \frac{dt}{dt_0} \frac{dy}{dt} \quad p_3 = m \frac{dt}{dt_0} \frac{dz}{dt} \quad (\text{A.18})$$

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

$$m_v = m \frac{dt}{dt_0} = \frac{m}{\sqrt{1 - (v/c)^2}} \quad (\text{A.19})$$

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.

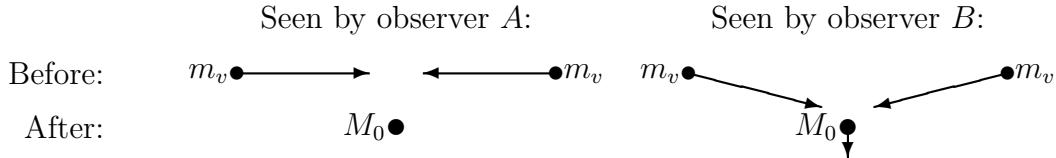


Figure A.3: A completely inelastic collision.

Note that so far, this only indicates that differences between $m_v c^2$ and mc^2 give 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, [14]. Consider two identical rest masses m_0 that are accelerated to high speed and then made to crash into each other head-on, as in the left part of figure A.3. 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_0 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_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_B . The two masses have a slight downward velocity v_B before the collision and so has the mass M_0 after the collision. But then vertical momentum conservation inevitably implies

$$2m_v v_B = M_0 v_B$$

So M_0 must be twice the moving mass m_v . The combined rest mass M_0 is not the sum of the *rest* masses m_0 , 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_0 .

A.4.10 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

$$\mathcal{L} = \frac{1}{2}m|\vec{v}|^2 - q\varphi \quad (\text{A.20})$$

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 a function 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

$$p_i^c = \frac{\partial \mathcal{L}}{\partial v_i} \quad (\text{A.21})$$

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

$$\frac{dp_i^c}{dt} = \frac{\partial \mathcal{L}}{\partial r_i} \quad (\text{A.22})$$

which is found to be

$$\frac{dp_i}{dt} = -q \frac{\partial \varphi}{\partial r_i}$$

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.

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

$$\mathcal{L} = -mc^2 \sqrt{1 - (|\vec{v}|/c)^2} + q \vec{\Phi} \cdot \frac{d\vec{r}}{dt} \quad \vec{\Phi} = \left(\frac{1}{c} \varphi, A_x, A_y, A_z \right) \quad (\text{A.23})$$

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

The canonical momentum is now

$$p_i^c = \frac{\partial \mathcal{L}}{\partial v_i} = m_v v_i + q A_i \quad (\text{A.24})$$

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, after clean up and in vector notation,

$$\frac{d\vec{p}}{dt} = q\vec{E} + q\vec{v} \times \vec{B} \quad (\text{A.25})$$

where

$$\vec{E} = -\nabla\phi - \frac{\partial\vec{A}}{\partial t} \quad \vec{B} = \nabla \times \vec{A}$$

are called the electric and magnetic fields, respectively. The right hand side in the equation of motion is called the Lorentz force.

The so-called “action” $\int \mathcal{L} dt$ is the same for different observers, provided that Φ transforms according to the Lorentz transformation. That then implies that different observers agree about the evolution, {A.3}. From the transformation of Φ , that of the electric and magnetic fields may be found; that will not be a Lorentz transformation.

It may be noted that the field strengths are unchanged in a “gauge transformation” that modifies φ and \vec{A} as

$$\varphi' = \varphi - \frac{\partial\Omega}{\partial t} \quad \vec{A}' = \vec{A} + \nabla\Omega \quad (\text{A.26})$$

where Ω is any arbitrary real function of position and time. In quantum mechanics, this gauge transform adds a phase factor to the wave function, but still leaves its magnitude unchanged, hence is usually inconsequential. Gauge transforms do become very important in advanced quantum mechanics, but that is beyond the scope of this book.

The energy can be found following note {A.3} as

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

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

$$H = mc^2 \sqrt{1 + \frac{(\vec{p}^c - q\vec{A})^2}{m^2 c^2}} + q\varphi$$

using the formula given in the overview subsection.

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

$$\frac{dp_i}{dt} + q \left(\frac{\partial A_i}{\partial t} + \frac{\partial A_i}{\partial x_j} v_j \right) = -q \frac{\partial \varphi}{\partial x_j} + q \frac{\partial A_j}{\partial x_i} v_j$$

using the Einstein convention of suppressing the summation symbols over j . Reorder to get

$$\frac{dp_i}{dt} = q \left(-\frac{\partial \varphi}{\partial x_j} - \frac{\partial A_i}{\partial t} \right) + q \left(\frac{\partial A_j}{\partial x_i} v_j - \frac{\partial A_i}{\partial x_j} v_j \right)$$

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

$$\frac{\partial A_i}{\partial x_i} v_i - \frac{\partial A_i}{\partial x_{\bar{i}}} v_{\bar{i}} + \frac{\partial A_{\bar{i}}}{\partial x_i} v_i - \frac{\partial A_i}{\partial x_{\bar{i}}} v_{\bar{i}} + \frac{\partial A_{\bar{i}}}{\partial x_{\bar{i}}} v_{\bar{i}} - \frac{\partial A_i}{\partial x_{\bar{i}}} v_{\bar{i}}$$

where \bar{i} follows i in the cyclic sequence $\dots, 1, 2, 3, 1, 2, 3, \dots$ and \bar{i} precedes it. This can be recognized as component number i of $\vec{v} \times (\nabla \times \vec{A})$. Defining \vec{B} as $\nabla \times \vec{A}$, the Lorentz force law results.

A.5 Completeness of Fourier modes

The purpose of this note is to show completeness of the “Fourier modes”

$$\dots, \frac{e^{-3ix}}{\sqrt{2\pi}}, \frac{e^{-2ix}}{\sqrt{2\pi}}, \frac{e^{-ix}}{\sqrt{2\pi}}, \frac{1}{\sqrt{2\pi}}, \frac{e^{ix}}{\sqrt{2\pi}}, \frac{e^{2ix}}{\sqrt{2\pi}}, \frac{e^{3ix}}{\sqrt{2\pi}}, \dots$$

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

$$f(x) = \dots + c_{-2} \frac{e^{-2ix}}{\sqrt{2\pi}} + c_{-1} \frac{e^{-ix}}{\sqrt{2\pi}} + c_0 \frac{1}{\sqrt{2\pi}} + c_1 \frac{e^{ix}}{\sqrt{2\pi}} + c_2 \frac{e^{2ix}}{\sqrt{2\pi}} + c_3 \frac{e^{3ix}}{\sqrt{2\pi}} + \dots$$

or

$$f(x) = \sum_{k=-\infty}^{\infty} c_k \frac{e^{kix}}{\sqrt{2\pi}}$$

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

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos(x)}{\sqrt{\pi}}, \frac{\sin(x)}{\sqrt{\pi}}, \frac{\cos(2x)}{\sqrt{\pi}}, \frac{\sin(2x)}{\sqrt{\pi}}, \frac{\cos(3x)}{\sqrt{\pi}}, \frac{\sin(3x)}{\sqrt{\pi}}, \dots$$

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

$$\dots, \frac{e^{-k_2 ix}}{\sqrt{2\ell}}, \frac{e^{-k_1 ix}}{\sqrt{2\ell}}, \frac{1}{\sqrt{2\ell}}, \frac{e^{k_1 ix}}{\sqrt{2\ell}}, \frac{e^{k_2 ix}}{\sqrt{2\ell}}, \dots \quad k_1 = \frac{1\pi}{\ell}, \quad k_2 = \frac{2\pi}{\ell}, \quad k_3 = \frac{3\pi}{\ell}, \dots$$

and similarly the real version of them becomes

$$\frac{1}{\sqrt{2\ell}}, \frac{\cos(k_1 x)}{\sqrt{\ell}}, \frac{\sin(k_1 x)}{\sqrt{\ell}}, \frac{\cos(k_2 x)}{\sqrt{\ell}}, \frac{\sin(k_2 x)}{\sqrt{\ell}}, \frac{\cos(k_3 x)}{\sqrt{\ell}}, \frac{\sin(k_3 x)}{\sqrt{\ell}}, \dots$$

See [28, 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 $0 \leq x \leq \ell$ and must satisfy $f(0) = f(\ell) = 0$, then extend them to the range $-\ell \leq x \leq 0$ by setting $f(x) = -f(-x)$ and take the range $-\ell \leq x \leq \ell$ 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 [28, 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 [28, 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 section 1.6 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

$$f(x) = \sum_{k=-\infty}^{k=\infty} c_k \frac{e^{kix}}{\sqrt{2\pi}},$$

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

$$c_l = \int_{x=0}^{2\pi} \frac{e^{-lix}}{\sqrt{2\pi}} f(x) dx,$$

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 rennotated as

$$c_k = \int_{\bar{x}=0}^{2\pi} \frac{e^{-kix}}{\sqrt{2\pi}} f(\bar{x}) d\bar{x},$$

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 $f_K(x)$,

$$f_K(x) \equiv \sum_{k=-K}^{k=K} c_k \frac{e^{kix}}{\sqrt{2\pi}},$$

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 $\lim_{K \rightarrow \infty} f_K(x) = f(x)$.

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:

$$f_K(x) = \frac{1}{2\pi} \int_{\bar{x}=0}^{2\pi} f(\bar{x}) \left[\sum_{k=-K}^{k=K} e^{ki(x-\bar{x})} \right] d\bar{x}.$$

The sum in the square brackets can be evaluated, because it is a geometric series with starting value $e^{-K\text{i}(x-\bar{x})}$ and ratio of terms $e^{\text{i}(x-\bar{x})}$. Using the formula from [28, item 21.4], multiplying top and bottom with $e^{-\text{i}(x-\bar{x})/2}$, and cleaning up with, what else, the Euler formula, the sum is found to equal

$$\frac{\sin((K + \frac{1}{2})(x - \bar{x}))}{\sin(\frac{1}{2}(x - \bar{x}))}.$$

This expression is called the “Dirichlet kernel”. You now have

$$f_K(x) = \int_{\bar{x}=0}^{2\pi} f(\bar{x}) \frac{\sin((K + \frac{1}{2})(x - \bar{x}))}{2\pi \sin(\frac{1}{2}(x - \bar{x}))} d\bar{x}.$$

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

$$\begin{aligned} f_K(x) &= f(x) \int_{\bar{x}=0}^{2\pi} \frac{\sin((K + \frac{1}{2})(x - \bar{x}))}{2\pi \sin(\frac{1}{2}(x - \bar{x}))} d\bar{x} \\ &\quad + \int_{\bar{x}=0}^{2\pi} (f(\bar{x}) - f(x)) \frac{\sin((K + \frac{1}{2})(x - \bar{x}))}{2\pi \sin(\frac{1}{2}(x - \bar{x}))} d\bar{x}. \end{aligned}$$

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

$$g(\bar{x}) = \frac{f(\bar{x}) - f(x)}{2\pi \sin(\frac{1}{2}(x - \bar{x}))}$$

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

$$f_K(x) = f(x) + \frac{1}{K + \frac{1}{2}} \int_{\bar{x}=0}^{2\pi} g'(\bar{x}) \cos((K + \frac{1}{2})(x - \bar{x})) d\bar{x}.$$

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.

A.6 Derivation of the Euler formula

To verify the Euler formula, write all three functions involved in terms of their Taylor series, [28, p. 136]

A.7 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 $55 + 90i$ and someone else had a $70 + 65i$? It would be logical in a world in which the important operators would not be Hermitian.

A.8 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.

A.9 Are linear momentum operators 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:

$$\begin{aligned}\langle \Psi_1 | \hat{p}_x \Psi_2 \rangle &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \Psi_1^* \frac{\hbar}{i} \frac{\partial \Psi_2}{\partial x} dx dy dz \\ \langle \hat{p}_x \Psi_1 | \Psi_2 \rangle &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \int_{z=-\infty}^{\infty} \left(\frac{\hbar}{i} \frac{\partial \Psi_1}{\partial x} \right)^* \Psi_2 dx dy dz\end{aligned}$$

The two must be equal for \hat{p}_x to be an Hermitian operator. That they are indeed equal may be seen from integration by parts in the x -direction, noting that by definition $i^* = -i$ and that Ψ_1 and Ψ_2 must be zero at infinite x : if they were not, their integral would be infinite, so that they could not be normalized.

A.10 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.

A.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:

$$H_x = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \quad H_y = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial y^2} \quad H_z = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial z^2}$$

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:

$$(H_x + H_y + H_z)\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z) = E_x\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z) + E_y\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z) + E_z\psi_{n_x}(x)\psi_{n_y}(y)\psi_{n_z}(z)$$

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.

Well, the one-dimensional eigenfunctions $\psi_{n_x}(x)$ are complete, see [28, 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.

A.12 Derivation of 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

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi_x}{\partial x^2} + \frac{1}{2} m \omega^2 x^2 \psi_x = E_x \psi_x$$

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

$$x \equiv \ell \xi \quad E_x \equiv E_0 \epsilon.$$

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

$$\frac{\partial^2 \psi_x}{\partial \xi^2} \approx \xi^2 \psi_x + \dots$$

If you write the solution as an exponential, you can ballpark that it must take the form

$$\psi_x = e^{\pm \frac{1}{2}\xi^2 + \dots}$$

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 $e^{-\frac{1}{2}\xi^2+\dots}$ 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 $c_0 + \dots$ or $c_1 \xi + \dots$. 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 ,

$$h = c_q \xi^q + c_{q+2} \xi^{q+2} + \dots$$

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 $(2(q+2)+1-\epsilon)c_{q+2} \xi^{q+2}$

in the right hand side which will have to be cancelled 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

$$h = c_0 + c_2\xi^2 + c_4\xi^4 + c_6\xi^6 + \dots$$

while if it starts with $q = 1$ you will get

$$h = c_1\xi + c_3\xi^3 + c_5\xi^5 + c_7\xi^7 + \dots$$

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 $e^{+\frac{1}{2}\xi^2+\dots}$. (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 [28, 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, $\int_{-\infty}^{\infty} |\psi_x|^2 dx = 1$. You should be able to find the value for the appropriate integral in your table book, like [28, 29.15].

Putting it all together, the generic expression for the eigenfunctions can be found to be:

$$h_n = \frac{1}{(\pi\ell^2)^{1/4}} \frac{H_n(\xi)}{\sqrt{2^n n!}} e^{-\xi^2/2} \quad n = 0, 1, 2, 3, 4, 5, \dots \quad (\text{A.27})$$

where the details of the “Hermite polynomials” H_n can be found in table books like [28, 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 chapter 12.2.2 or equivalently in {A.89}.

A.13 More on 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 not defined for the ground state. However, as explained more fully in chapter 3.3, you can define the “expectation value” of kinetic energy to be the average predicted result for kinetic energy measurements. Similarly, you can define the expectation value of potential energy to be the average predicted result for potential energy measurements. Quantum mechanics does require the total energy of the ground state to be the sum of the kinetic and potential energy expectation values. Now if there would be an almost infinite uncertainty in linear momentum, then typical measurements would find a large momentum, hence a large kinetic energy. So the kinetic energy expectation value would then

be large; that would be nowhere close to any ground state. Similarly, if there would be a large uncertainty in position, then typical measurements will find the particle at large distance from the nominal position, hence at large potential energy. Not good either.

It so happens that the ground state of the harmonic oscillator manages to obtain the absolute minimum in combined position and momentum uncertainty that the uncertainty relationship, given in chapter 3.4.3, allows. (This can be verified using the fact that the two uncertainties, σ_x and σ_{p_x} , as defined in chapter 3.3, are directly related to the expectation values for potential energy, respectively kinetic energy in the x -direction. It follows from the virial theorem of chapter 6.1.7 that the expectation values of kinetic and potential energy for the harmonic oscillator eigenstates are equal. So each must be $\frac{1}{4}\hbar\omega$ since the total energy is $\frac{3}{2}\hbar\omega$ and each coordinate direction contributes an equal share to the potential and kinetic energies.)

A.14 Derivation of a vector identity

The elementary equality required is not in [28] in any form. In the absence of tensor algebra, it is best to just grind it out. Define $\vec{f} \equiv (\vec{r} \times \nabla)\Psi$. Then $(\vec{r} \times \nabla) \cdot \vec{f}$ equals

$$y \frac{\partial f_x}{\partial z} - z \frac{\partial f_x}{\partial y} + z \frac{\partial f_y}{\partial x} - x \frac{\partial f_y}{\partial z} + x \frac{\partial f_z}{\partial y} - y \frac{\partial f_z}{\partial x}$$

On the other hand, $\vec{r} \cdot (\nabla \times \vec{f})$ is

$$x \frac{\partial f_z}{\partial y} - x \frac{\partial f_y}{\partial z} + y \frac{\partial f_x}{\partial z} - y \frac{\partial f_z}{\partial x} + z \frac{\partial f_y}{\partial x} - z \frac{\partial f_x}{\partial y}$$

which is the same.

A.15 Derivation of the spherical harmonics

This analysis will use similar techniques as for the harmonic oscillator solution, {A.12}. The requirement that the spherical harmonics Y_l^m are eigenfunctions of L_z means that they are of the form $\Theta_l^m(\theta)e^{im\phi}$ 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 $\hat{L}^2\psi = L^2\psi$ with \hat{L}^2 as in (3.5) yields an ODE (ordinary differential equation) for $\Theta_l^m(\theta)$:

$$-\frac{\hbar^2}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial \Theta_l^m}{\partial \theta} \right) + \frac{\hbar^2 m^2}{\sin^2 \theta} \Theta_l^m = L^2 \Theta_l^m$$

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

$$-(1 - x^2) \frac{d^2\Theta_l^m}{dx^2} + 2x \frac{d\Theta_l^m}{dx} + \frac{m^2}{1 - x^2} \Theta_l^m = \lambda^2 \Theta_l^m$$

As you may guess from looking at this ODE, the solutions Θ_l^m are likely to be problematic near $x = \pm 1$, (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 section 3.3.3, 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 :

$$-(1 - x^2) \frac{d^2 f_l^m}{dx^2} + 2(1 + m)x \frac{df_l^m}{dx} + (m^2 + m)f_l^m = \lambda^2 f_l^m$$

Plug in a power series, $f_l^m = \sum c_p x^p$, to get, after clean up,

$$\sum p(p - 1)c_p x^{p-2} = \sum [(p + m)(p + m + 1) - \lambda^2] c_p x^p$$

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 $l \geq m$ 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, [28, 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 [28, 28.49], and that the solutions that you need are the associated Legendre functions of the first kind [28, 28.50].

To normalize the eigenfunctions on the surface area of the unit sphere, find the corresponding integral in a table book, like [28, 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, {A.89}, 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

$$Y_l^m(\theta, \phi) = (-1)^{\max(m, 0)} \sqrt{\frac{2l+1}{4\pi} \frac{(l-|m|)!}{(l+|m|)!}} P_l^{|m|}(\cos \theta) e^{im\phi} \quad (\text{A.28})$$

where the properties of the associated Legendre functions of the first kind $P_l^{|m|}$ can be found in table books like [28, pp. 162-166].

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 $e^{im\phi}$ 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 $P_l^{|m|}(\cos \theta)$.

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

$$\sum_{2a+b=l-m} c_{ab} u^{a+m} v^a z^b \quad a, b, m \geq 0$$

$$\sum_{2a+b=l-|m|} c_{ab} u^a v^{a+|m|} z^b \quad a, b, -m \geq 0$$

where the coordinates $u = x + iy$ and $v = x - iy$ 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 non-power terms, to settle completeness. In other words, you must assume that the solution is analytic.

The simplest way of getting the spherical harmonics is probably the one given in note {A.89}.

A.16 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 4.1.

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, and 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:

$$\vec{r}_{\text{cg}} = w_1 \vec{r}_1 + w_2 \vec{r}_2 \quad w_1 = \frac{m_1}{m_1 + m_2} \quad w_2 = \frac{m_2}{m_1 + m_2}$$

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

$$\vec{r}_1 = -w_2 \vec{r}_{21} \quad \vec{r}_2 = w_1 \vec{r}_{21}$$

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

$$m_{\text{red}} = \frac{m_1 m_2}{m_1 + m_2}$$

(A.29)

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:

$$H\psi = E\psi \quad H = -\frac{\hbar^2}{2m_1}\nabla_1^2 - \frac{\hbar^2}{2m_2}\nabla_2^2 + V(\vec{r}_{21})$$

where the two kinetic energy Laplacians in the Hamiltonian H are with respect to the position coordinates of the two particles:

$$\nabla_1^2\psi \equiv \sum_{j=1}^3 \frac{\partial^2\psi}{\partial r_{1,j}^2} \quad \nabla_2^2\psi \equiv \sum_{j=1}^3 \frac{\partial^2\psi}{\partial r_{2,j}^2}$$

Now make a change of variables from \vec{r}_1 and \vec{r}_2 to \vec{r}_{cg} and \vec{r}_{21} where

$$\vec{r}_{\text{cg}} = w_1\vec{r}_1 + w_2\vec{r}_2 \quad \vec{r}_{21} = \vec{r}_2 - \vec{r}_1$$

The derivatives of ψ can be converted using the chain rule of differentiation:

$$\frac{\partial\psi}{\partial r_{1,j}} = \frac{\partial\psi}{\partial r_{\text{cg},j}} \frac{\partial r_{\text{cg},j}}{\partial r_{1,j}} + \frac{\partial\psi}{\partial r_{21,j}} \frac{\partial r_{21,j}}{\partial r_{1,j}} = \frac{\partial\psi}{\partial r_{\text{cg},j}} w_1 - \frac{\partial\psi}{\partial r_{21,j}}$$

or differentiating once more and summing

$$\nabla_1^2\psi = \sum_{j=1}^3 \frac{\partial^2\psi}{\partial r_{1,j}^2} = w_1^2 \sum_{j=1}^3 \frac{\partial^2\psi}{\partial r_{\text{cg},j}^2} - 2w_1 \sum_{j=1}^3 \frac{\partial^2\psi}{\partial r_{\text{cg},j} \partial r_{21,j}} + \sum_{j=1}^3 \frac{\partial^2\psi}{\partial r_{21,j}^2}$$

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

$$H = -\frac{\hbar^2}{2(m_1 + m_2)}\nabla_{\text{cg}}^2 - \frac{\hbar^2}{2m_{\text{red}}}\nabla_{21}^2 + V(\vec{r}_{21})$$

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

$$\psi = \psi_{\text{cg}}(\vec{r}_{\text{cg}})\psi_{21}(\vec{r}_{21})$$

Substituting this and the Hamiltonian above into $H\psi = E\psi$ and dividing by $\psi_{\text{cg}}\psi_{21}$ produces

$$-\frac{\hbar^2}{2(m_1 + m_2)} \frac{1}{\psi_{\text{cg}}} \nabla_{\text{cg}}^2 \psi_{\text{cg}} + \frac{1}{\psi_{21}} \left[-\frac{\hbar^2}{2m_{\text{red}}} \nabla_{21}^2 + V \right] \psi_{21} = E$$

Call the first term in the left hand side E_{cg} and the second E_{21} . By that definition, E_{cg} would normally be a function of \vec{r}_{cg} , because ψ_{cg} is, but since it is equal to $E - E_{21}$ and those do not depend on \vec{r}_{cg} , E_{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_{cg} and E_{21} , two separate eigenvalue problems are obtained:

$$-\frac{\hbar^2}{2(m_1 + m_2)} \nabla_{\text{cg}}^2 \psi_{\text{cg}} = E_{\text{cg}} \psi_{\text{cg}} \quad \left[-\frac{\hbar^2}{2m_{\text{red}}} \nabla_{21}^2 + V \right] \psi_{21} = E_{21} \psi_{21}$$

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_{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 $m_p m_e / (m_p + m_e)$, 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.17 The hydrogen radial wave functions

This will be child's play for harmonic oscillator, {A.12}, and spherical harmonics, {A.15}, veterans. If you replace the angular terms in (3.15) by $l(l+1)\hbar^2$, and then divide the entire equation by \hbar^2 , you get

$$-\frac{1}{R} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) + l(l+1) - 2 \frac{m_e e^2}{4\pi\epsilon_0 \hbar^2} r = \frac{2m_e}{\hbar^2} r^2 E$$

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 $(\hbar^2/2m_e a_0^2)\epsilon$, where ϵ is a nondimensional energy. In terms of these scaled coordinates you get

$$-\frac{1}{R} \frac{d}{d\rho} \left(\rho^2 \frac{dR}{d\rho} \right) + l(l + 1) - 2\rho = \rho^2 \epsilon$$

or written out

$$-\rho^2 R'' - 2\rho R' + [l(l + 1) - 2\rho - \epsilon\rho^2]R = 0$$

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: $\int |\psi|^2 d^3\vec{r}$ 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 $\cos(\sqrt{\epsilon}\rho + \phi)$ 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 $R = e^{-\frac{1}{2}\xi}\bar{R}$ where $\bar{R}(\xi)$ must remain bounded at large ξ . Substituting these new variables, the ODE becomes

$$-\xi^2 \bar{R}'' + \xi(\xi - 2)\bar{R}' + [l(l + 1) - (n - 1)\xi]\bar{R} = 0$$

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 $\bar{R} = \xi^l \tilde{R}$, which turns the ODE into

$$\xi \bar{R}'' + [2(l+1) - \xi] \bar{R}' + [n-l-1] \bar{R} = 0$$

Substituting in a power series $\bar{R} = \sum c_p \xi^p$, 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. [28, 30.26], the $(2l+1)$ -th derivative of Laguerre's differential equation, e.g. [28, 30.1], and the polynomial solutions as the associated Laguerre polynomials L_{n+l}^{2l+1} , e.g. [28, 30.27], the $(2l+1)$ -th derivatives of the Laguerre's polynomials L_{n+l} , e.g. [28, 30.2]. To normalize the wave function use an integral from a table book, e.g. [28, 30.46].

Putting it all together, the generic expression for hydrogen eigenfunctions are, drums please:

$$\psi_{nlm} = -\frac{2}{n^2} \sqrt{\frac{(n-l-1)!}{[(n+l)!a_0]^3}} \left(\frac{2\rho}{n}\right)^l L_{n+l}^{2l+1} \left(\frac{2\rho}{n}\right) e^{-\rho/n} Y_l^m(\theta, \phi) \quad (\text{A.30})$$

The properties of the associated Laguerre polynomials $L_{n+l}^{2l+1}(2\rho/n)$ are in table books like [28, pp. 169-172], and the spherical harmonics were given earlier in section 3.1.3 and in note {A.15}, (A.28).

Do keep in mind that different references have contradictory definitions of the associated Laguerre polynomials. This book follows the notations of [28, pp. 169-172], who define

$$L_n(x) = e^x \frac{d^n}{dx^n} (x^n e^{-x}), \quad L_n^m = \frac{d^m}{dx^m} L_n(x).$$

In other words, L_n^m is simply the m -th derivative of L_n , which certainly tends to simplify things. According to [17, p. 152], the “most nearly standard” notation defines

$$L_n^m = (-1)^m \frac{d^m}{dx^m} L_{n+m}(x).$$

Combine the messy definition of the spherical harmonics (A.28) with the uncertain definition of the Laguerre polynomials in the formulae (A.30) 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 (A.30), it is seen that

$$\psi_{nlm} \propto r^l \quad \text{for } r \rightarrow 0 \quad (\text{A.31})$$

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 $L_n(0) = n!$, [28, 30.19], while the differential equation for these function implies that $L_n'(0) = -nL_n(0)$. Therefore:

$$\psi_{n00}(0) = \frac{1}{\sqrt{n^3 \pi a_0^3}} \quad (\text{A.32})$$

A.18 Inner product for the expectation value

To see that $\langle \Psi | A | \Psi \rangle$ works for getting the expectation value, just write Ψ out in terms of the eigenfunctions α_n of A :

$$\langle c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \dots | A | c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \dots \rangle$$

Now by the definition of eigenfunctions $A\alpha_n = a_n\alpha_n$ for every n , so you get

$$\langle c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \dots | c_1a_1\alpha_1 + c_2a_2\alpha_2 + c_3a_3\alpha_3 + \dots \rangle$$

Since eigenfunctions are orthonormal:

$$\langle \alpha_1 | \alpha_1 \rangle = 1 \quad \langle \alpha_2 | \alpha_2 \rangle = 1 \quad \langle \alpha_3 | \alpha_3 \rangle = 1 \quad \dots$$

$$\langle \alpha_1 | \alpha_2 \rangle = \langle \alpha_2 | \alpha_1 \rangle = \langle \alpha_1 | \alpha_3 \rangle = \langle \alpha_3 | \alpha_1 \rangle = \langle \alpha_2 | \alpha_3 \rangle = \langle \alpha_3 | \alpha_2 \rangle = \dots = 0$$

So, multiplying out produces the desired result:

$$\langle \Psi | A | \Psi \rangle = |c_1|^2 a_1 + |c_2|^2 a_2 + |c_3|^2 a_3 + \dots \equiv \langle A \rangle$$

A.19 Why commuting operators have common eigenvectors

The fact that two operators that commute have a common set of eigenvectors can be seen as follows: assume that \vec{a} is an eigenvector of A with eigenvalue

a. Then since A and B commute, $AB\vec{\alpha} = BA\vec{\alpha} = aB\vec{\alpha}$, so, comparing start and end, $B\vec{\alpha}$ must be an eigenvector of A with eigenvalue a too. If there is no degeneracy of the eigenvalue, that must mean that $B\vec{\alpha}$ equals $\vec{\alpha}$ or is at least proportional to it, which is the same as saying that $\vec{\alpha}$ is an eigenvector of B too. (In the special case that $B\vec{\alpha}$ is zero, α is an eigenvector of B with eigenvalue zero.)

If there is degeneracy, the eigenvectors of A are not unique and you can mess with them until they all do become eigenvectors of B too. The following procedure will construct such a set of common eigenvectors in finite dimensional space. Consider each eigenvalue of A in turn. There will be more than one eigenvector corresponding to a degenerate eigenvalue a . Now by completeness, any eigenvector β can be written as a combination of the eigenvectors of A , and more particularly as $\beta = \beta_n + \beta_a$ where β_a is a combination of the eigenvectors of A with eigenvalue a and β_n a combination of the eigenvectors of A with other eigenvalues.

The vectors β_n and β_a separately are still eigenvectors of B if nonzero, since as noted above, B converts eigenvectors of A into eigenvectors with the same eigenvalue or zero. (For example, if $B\beta_a$ was not $b\beta_a$, $B\beta_n$ would have to make up the difference, and $B\beta_n$ can only produce combinations of eigenvectors of A that do *not* have eigenvalue a .) Now replace the eigenvector β by either β_a or β_n , whichever is independent of the other eigenvectors of B . Doing this for all eigenvectors of B you achieve that the replacement eigenvectors of B are either combinations of the eigenvectors of A with eigenvalue a or of the other eigenvectors of A . The set of new eigenvectors of B that are combinations of the eigenvectors of A with eigenvalue a can now be taken as the replacement eigenvectors of A with eigenvalue a . They are also eigenvectors of B . Repeat for all eigenvalues of A .

Similar arguments can be used recursively to show that more generally, a set of operators that all commute have a common set of eigenvectors.

The operators do not really have to be Hermitian, just “diagonalizable”: they must have a complete set of eigenfunctions.

In the infinite dimensional case the mathematical justification gets much trickier. However, as the hydrogen atom and harmonic oscillator eigenfunction examples indicate, it continues to be relevant in nature.

A.20 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 ,

$$|\langle f|g \rangle| \leq \sqrt{\langle f|f \rangle} \sqrt{\langle g|g \rangle}$$

(See the notations for more on this theorem.) Using the Cauchy-Schwartz inequality in reversed order, you get

$$\sigma_A^2 \sigma_B^2 \geq |\langle A' \Psi | B' \Psi \rangle|^2 = |\langle A' B' \rangle|^2$$

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

$$\sigma_A^2 \sigma_B^2 \geq \left| \frac{\langle A' B' \rangle - \langle B' A' \rangle}{2} \right|^2$$

The quantity in the top is the expectation value of the commutator $[A', B']$. Writing it out shows that $[A', B'] = [A, B]$.

A.21 Derivation of the commutator rules

This note explains where the formulae of section 3.4.4 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 x -position and p_x -linear momentum 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

$$[\hat{L}_x, \hat{L}_y] = [\hat{y}\hat{p}_z - \hat{z}\hat{p}_y, \hat{z}\hat{p}_x - \hat{x}\hat{p}_z]$$

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 “ $xyzxyzx\dots$ ” “cyclic permutation” scheme, as in:

$$x \rightarrow y, \quad y \rightarrow z, \quad z \rightarrow x$$

which produces the other two commutators if you do it twice:

$$[\hat{L}_x, \hat{L}_y] = i\hbar \hat{L}_z \quad \longrightarrow \quad [\hat{L}_y, \hat{L}_z] = i\hbar \hat{L}_x \quad \longrightarrow \quad [\hat{L}_z, \hat{L}_x] = i\hbar \hat{L}_y$$

For the commutators with square angular momentum, work out

$$[\hat{L}_x, \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2]$$

using the manipulation rules and the commutators between angular momentum components.

A commutator like $[\hat{x}, \hat{L}_x] = [\hat{x}, \hat{y}\hat{p}_z - \hat{z}\hat{p}_y]$ is zero because everything commutes in it. However, in a commutator like $[\hat{x}, \hat{L}_y] = [\hat{x}, \hat{z}\hat{p}_x - \hat{x}\hat{p}_z]$, \hat{x} does not commute with \hat{p}_x , so multiplying out and taking the \hat{z} out of $[\hat{x}, \hat{z}\hat{p}_x]$ at its own side, you get $\hat{z}[\hat{x}, \hat{p}_x]$, and the commutator left is the canonical one, which has value $i\hbar$. Plug these results and similar into $[\hat{x}^2 + \hat{y}^2 + \hat{z}^2, \hat{L}_x]$ and you get zero.

For a commutator like $[\hat{x}, \hat{L}^2] = [\hat{x}, \hat{L}_x^2 + \hat{L}_y^2 + \hat{L}_z^2]$, the \hat{L}_x^2 term produces zero because \hat{L}_x commutes with \hat{x} , and in the remaining term, taking the various factors out at their own sides of the commutator produces

$$= \hat{L}_y[\hat{x}, \hat{L}_y] + [\hat{x}, \hat{L}_y]\hat{L}_y + \hat{L}_z[\hat{x}, \hat{L}_z] + [\hat{x}, \hat{L}_z]\hat{L}_z = i\hbar \hat{L}_y \hat{z} + i\hbar \hat{z} \hat{L}_y - i\hbar \hat{L}_z \hat{y} - i\hbar \hat{y} \hat{L}_z$$

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 $\hat{L}_y \hat{z}$ as long as you add the commutator $[\hat{L}_y, \hat{z}]$ to make up for it, and that commutator was already found to be $i\hbar \hat{x}$. The same way the order of $\hat{L}_z \hat{y}$ can be swapped to give

$$= -2\hbar^2 \hat{x} - 2i\hbar(\hat{y} \hat{L}_z - \hat{z} \hat{L}_y)$$

and the parenthetical expression can be recognized as the x -component of $\vec{r} \times \vec{\hat{L}}$, giving one of the expressions claimed. Instead you can work out the parenthetical expression further by substituting in the definitions for \hat{L}_z and \hat{L}_y :

$$= -2\hbar^2 \hat{x} - 2i\hbar \left(\hat{y}(\hat{x}\hat{p}_y - \hat{y}\hat{p}_x) - \hat{z}(\hat{z}\hat{p}_x - \hat{x}\hat{p}_z) - \hat{x}(\hat{x}\hat{p}_x - \hat{x}\hat{p}_x) \right)$$

where the third term added within the big parenthesis 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 (3.55), consider the z -component as the example:

$$[x\hat{L}_y - y\hat{L}_x, \hat{L}^2] = [x, \hat{L}^2]\hat{L}_y - [y, \hat{L}^2]\hat{L}_x$$

because L^2 commutes with \hat{L} , and using (3.50)

$$[x\hat{L}_y - y\hat{L}_x, \hat{L}^2] = -2\hbar^2 x\hat{L}_y - 2i\hbar(y\hat{L}_z\hat{L}_y - z\hat{L}_y^2) + 2\hbar^2 y\hat{L}_x + 2i\hbar(z\hat{L}_x^2 - x\hat{L}_z\hat{L}_x)$$

Now use the commutator $[\hat{L}_y, \hat{L}_z]$ to get rid of $\hat{L}_z\hat{L}_y$ and $[\hat{L}_z, \hat{L}_x]$ to get rid of $\hat{L}_z\hat{L}_x$ and clean up to get

$$[x\hat{L}_y - y\hat{L}_x, \hat{L}^2] = 2i\hbar(-y\hat{L}_y\hat{L}_z + z\hat{L}_y^2 + z\hat{L}_x^2 - x\hat{L}_x\hat{L}_z)$$

Now $\vec{r} \cdot \hat{\vec{L}} = \vec{r} \cdot (\vec{r} \times \hat{\vec{p}}) = 0$ so $x\hat{L}_x + y\hat{L}_y = -z\hat{L}_z$, which gives the claimed expression. To verify the second equation of (3.55), use (3.50), the first of (3.55), and the definition of $[\vec{r}, \hat{L}^2]$.

A.22 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.24}. (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.

A.23 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

$$\langle E \rangle = \langle a\psi_l + b\psi_r | H | a\psi_l + b\psi_r \rangle$$

This can be multiplied out and simplified by noting that ψ_l and ψ_r are eigenfunctions of the partial Hamiltonians. For example,

$$H\psi_l = E_1\psi_l - \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_l} \psi_l$$

where E_1 is the -13.6 eV hydrogen atom ground state energy. The expression can be further simplified by noting that by symmetry

$$\langle \psi_r | r_l^{-1} \psi_r \rangle = \langle \psi_l | r_r^{-1} \psi_l \rangle \quad \langle \psi_l | r_l^{-1} \psi_r \rangle = \langle \psi_r | r_r^{-1} \psi_l \rangle$$

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

$$a^2 + b^2 + 2ab\langle \psi_l | \psi_r \rangle = 1$$

Cleaning up the expectation energy in this way, the result is

$$\begin{aligned} \langle E \rangle &= E_1 - \\ &\frac{e^2}{4\pi\epsilon_0} \left[\langle \psi_l | r_r^{-1} \psi_l \rangle - \frac{1}{d} + 2ab \langle \psi_l | \psi_r \rangle \left\{ \frac{\langle \psi_l | r_l^{-1} \psi_r \rangle}{\langle \psi_l | \psi_r \rangle} - \langle \psi_l | r_r^{-1} \psi_l \rangle \right\} \right] \end{aligned}$$

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

$$r_l = |\vec{r} - \vec{r}_{lp}| = r \quad r_r = |\vec{r} - \vec{r}_{rp}| = \sqrt{d^2 + r^2 - 2dr \cos(\theta)}.$$

In those terms,

$$\psi_l = \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0} \quad \psi_r = \frac{1}{\sqrt{\pi a_0^3}} e^{-\sqrt{d^2+r^2-2dr \cos(\theta)}/a_0}.$$

Then integrate angles first using $d^3\vec{r} = r^2 \sin(\theta) dr d\theta d\phi = -r^2 dr d\cos(\theta) d\phi$. Do not forget that $\sqrt{x^2} = |x|$, not x , e.g. $\sqrt{(-3)^2} = 3$, not -3. More details are in [17, pp. 305-307].

The “overlap integral” turns out to be

$$\langle \psi_l | \psi_r \rangle = e^{-d/a_0} \left[1 + \frac{d}{a_0} + \frac{1}{3} \left(\frac{d}{a_0} \right)^2 \right]$$

and provides a measure of how much the regions of the two wave functions overlap. The “direct integral” is

$$\langle \psi_l | r_r^{-1} \psi_l \rangle = \frac{1}{d} - \left[\frac{1}{a_0} + \frac{1}{d} \right] e^{-2d/a_0}$$

and gives the classical potential of an electron density of strength $|\psi_l|^2$ in the field of the right proton, except for the factor $-e^2/4\pi\epsilon_0$. The “exchange integral” is

$$\langle \psi_l | r_l^{-1} \psi_r \rangle = \left[\frac{1}{a_0} + \frac{d}{a_0^2} \right] e^{-d/a_0}.$$

and is somewhat of a twilight term, since ψ_l suggests that the electron is around the left proton, but ψ_r suggests it is around the right one.

A.24 Accuracy of the variational method

Any approximate ground state solution ψ may always be written as a sum of the eigenfunctions ψ_1, ψ_2, \dots :

$$\psi = c_1 \psi_1 + \varepsilon_2 \psi_2 + \varepsilon_3 \psi_3 + \dots$$

where, if the approximation is any good at all, the coefficient c_1 of the ground state ψ_1 is close to one, while $\varepsilon_2, \varepsilon_3, \dots$ are small.

The condition that ψ is normalized, $\langle \psi | \psi \rangle = 1$, works out to be

$$1 = \langle c_1 \psi_1 + \varepsilon_2 \psi_2 + \dots | c_1 \psi_1 + \varepsilon_2 \psi_2 + \dots \rangle = c_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \dots$$

since the eigenfunctions ψ_1, ψ_2, \dots are orthonormal.

Similarly, the expectation energy $\langle E \rangle = \langle \psi | H \psi \rangle$ works out to be

$$\langle E \rangle = \langle c_1 \psi_1 + \varepsilon_2 \psi_2 + \dots | E_1 c_1 \psi_1 + E_2 \varepsilon_2 \psi_2 + \dots \rangle = c_1^2 E_1 + \varepsilon_2^2 E_2 + \varepsilon_3^2 E_3 + \dots$$

Eliminating c_1^2 using the normalization condition above gives

$$\langle E \rangle = E_1 + \varepsilon_2^2 (E_2 - E_1) + \varepsilon_3^2 (E_3 - E_1) + \dots$$

One of the things this expression shows is that any approximate wave function (not just eigenfunctions) has more expectation energy than the ground

state E_1 . All other terms in the sum above are positive since E_1 is the lowest energy value.

The expression above also shows that while the deviations of the wave function from the exact ground state ψ_1 are proportional to the coefficients $\varepsilon_2, \varepsilon_3, \dots$, the errors in energy are proportional to the *squares* of those coefficients. And the square of any reasonably small quantity is much smaller than the quantity itself. So the approximate ground state energy is much more accurate than would be expected from the wave function errors.

Still, if an approximate system is close to the ground state energy, then the wave function must be close to the ground state wave function. More precisely, if the error in energy is a small number, call it ε^2 , then the amount ε_2 of eigenfunction ψ_2 “polluting” approximate ground state ψ must be no more than $\varepsilon/\sqrt{E_2 - E_1}$. And that is in the worst case scenario that all the error in the expectation value of energy is due to the second eigenfunction.

As a measure of the average combined error in wave function, you can use the magnitude or norm of the combined pollution:

$$\|\varepsilon_2\psi_2 + \varepsilon_3\psi_3 + \dots\| = \sqrt{\varepsilon_2^2 + \varepsilon_3^2 + \dots}$$

That error is no more than $\varepsilon/\sqrt{E_2 - E_1}$. To verify it, note that

$$\varepsilon_2^2(E_2 - E_1) + \varepsilon_3^2(E_2 - E_1) + \dots \leq \varepsilon_2^2(E_2 - E_1) + \varepsilon_3^2(E_3 - E_1) + \dots = \varepsilon^2.$$

(Of course, if the ground state wave function would be degenerate, E_2 would be E_1 . But in that case you do not care about the error in ψ_2 , since then ψ_1 and ψ_2 are equally good ground states, and $E_2 - E_1$ becomes $E_3 - E_1$.)

The bottom line is that the lower you can get your expectation energy, the closer you will get to the true ground state energy, and the small error in energy will reflect in a small error in wave function.

A.25 Positive molecular ion wave function

Since the assumed Hamiltonian is real, taking real and imaginary parts of the eigenvalue problem $H\psi = E\psi$ shows that the real and imaginary parts of ψ each separately are eigenfunctions with the same eigenvalue, and both are real. So you can take ψ to be real without losing anything.

The expectation value of the energy for $|\psi|$ is the same as that for ψ , assuming that an integration by parts has been done on the kinetic energy part to convert it into an integral of the square gradients of ψ . Therefore $|\psi|$ must be the same function as ψ within a constant, assuming that the ground state of lowest energy is non degenerate. That means that ψ cannot change sign and can be taken to be positive.

(Regrettably this argument stops working for more than two electrons due to the antisymmetrization requirement of section 4.6. It does keep working for bosons, like helium atoms in a box, [13, p. 321])

With a bit more sophistication, the above argument can be inverted to show that the ground state must indeed be unique. Assume that there would be two different (more precisely: not equal within a constant) ground state eigenfunctions, instead of just one. Then linear combinations of the two would exist that crossed zero. The absolute value of such a wave function would, again, have the same expectation energy as the wave function itself, the ground state energy. But the absolute value of such a wave function has kinks of finite slope at the zero crossings. (Just think of the graph of $|x|$.) If these kinks are locally slightly smoothed out, i.e. rounded off, the kinetic energy would decrease correspondingly, since kinetic energy is the integral of the square slope and the slope has been reduced nontrivially in the immediate vicinity of the zero crossings. However, there would not be a corresponding increase in potential energy, since the potential energy depends on the square of the wave function itself, not its slope, and the square of the wave function itself is vanishingly small in the immediate vicinity of a zero crossing. If the kinetic energy goes down, and the potential energy does not go up enough to compensate, the energy would be lowered. But that contradicts the fact that the ground state has the lowest possible energy. The contradiction implies that the original assumption of two different ground state eigenfunctions cannot be right; the ground state must be unique.

A.26 Molecular ion wave function symmetries

Let z be the horizontal coordinate measured from the symmetry plane towards the right nucleus. Let M be the “mirror operator” that changes the sign of z , in other words,

$$M\Psi(x, y, z) = \Psi(x, y, -z)$$

This operator commutes with the Hamiltonian H since the energy evaluates the same way at positive and negative z . This means that operators H and M must have a complete set of common eigenfunctions. That set must include the ground state of lowest energy: so the ground state must be an eigenfunction of M too. Now the eigenvalues of M , which can be seen to be a Hermitian operator, are either $+1$ or -1 : if M is applied twice, it gives back the same wave function, i.e. 1Ψ , so the square of the eigenvalue is 1, so that the eigenvalue itself can only be 1 and -1 . Eigenfunctions with eigenvalue 1 are called “symmetric”, eigenfunctions with eigenvalue -1 are called “antisymmetric”. Since the previous note found that the ground state must be everywhere positive, it can only be a symmetric eigenfunction of M .

Similarly, let R be the operator that rotates Ψ over a small angle ϕ around the axis of symmetry. The magnitude of the eigenvalues of R must be 1, since Ψ must stay normalized to 1 after the rotation. Complex numbers of magnitude 1 can be written as e^{ia} where a is a real number. Number a must be proportional to ϕ , since rotating Ψ twice is equivalent to rotating it once over twice the angle, so the eigenvalues are $e^{im\phi}$, where m is a constant independent of ϕ . (In addition, m must be integer since rotating over 360 degrees must give back the original wave function.) In any case, the only way that Ψ can be real and positive at all angular positions is if $m = 0$, and then the eigenvalue of R is 1, implying that the ground state Ψ does not change when rotated; it must be the same at all angles. That means that the wave function is axially symmetric.

For future reference, one other symmetry must be mentioned, for the ground state of the neutral hydrogen molecule that will be covered in the next chapter. The neutral molecule has two electrons, instead of one, with positions \vec{r}_1 and \vec{r}_2 . The Hamiltonian will commute with the operation of “exchanging the electrons,” i.e. swapping the values of \vec{r}_1 and \vec{r}_2 , because all electrons are identical. So, for the same reasons as for the mirror operator above, the spatial wave function will be symmetric, unchanged, under particle exchange.

(Regrettably this argument stops working for more than two electrons due to the antisymmetrization requirement of section 4.6. It does keep working for bosons, like helium atoms, [13, p. 321])

A.27 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 $a\psi_l\psi_r + b\psi_r\psi_l$.

First, for given a/b , the individual values of a and b can be computed from the normalization requirement

$$a^2 + b^2 + 2ab\langle\psi_l|\psi_r\rangle^2 = 1 \quad (\text{A.33})$$

where the value of the overlap integral $\langle\psi_l|\psi_r\rangle$ was given in note {A.23}.

The inner product

$$\langle a\psi_l\psi_r + b\psi_r\psi_l | H | a\psi_l\psi_r + b\psi_r\psi_l \rangle_6$$

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 note {A.23}. Cleaning up the inner product, and using the normalization condition, you can get:

$$\langle E \rangle = 2E_1 - \frac{e^2}{4\pi\epsilon_0} \left[A_1 + 2ab\langle\psi_l|\psi_r\rangle^2 A_2 \right]$$

using the abbreviations

$$A_1 = 2\langle\psi_l|r_r^{-1}\psi_l\rangle - \frac{1}{d} - \langle\psi_l\psi_r|r_{12}^{-1}\psi_l\psi_r\rangle$$

$$A_2 = \frac{2\langle\psi_l|r_1^{-1}\psi_r\rangle}{\langle\psi_l|\psi_r\rangle} - 2\langle\psi_l|r_r^{-1}\psi_l\rangle - \frac{\langle\psi_l\psi_r|r_{12}^{-1}\psi_r\psi_l\rangle}{\langle\psi_l|\psi_r\rangle^2} + \langle\psi_l\psi_r|r_{12}^{-1}\psi_l\psi_r\rangle$$

Values for several of the inner products in these expressions are given in note {A.23}. 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.

A.28 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 $\langle E \rangle = \langle \psi_{\text{gs}} | H \psi_{\text{gs}} \rangle$, characterizing the correct ground state, only occurs if all spatial components $\psi_{\pm\pm}$ of the ground state with spin,

$$\psi_{\text{gs}} = \psi_{++}\uparrow\uparrow + \psi_{+-}\uparrow\downarrow + \psi_{-+}\downarrow\uparrow + \psi_{--}\downarrow\downarrow,$$

are proportional to the no-spin spatial ground state $\psi_{\text{gs},0}$.

The reason is that the assumed Hamiltonian (4.3) does not involve spin at all, only spatial coordinates, so, for example,

$$(H\psi_{++}\uparrow\uparrow) \equiv H(\psi_{++}(\vec{r}_1, \vec{r}_2)\uparrow(S_{z1})\uparrow(S_{z2})) = (H\psi_{++})\uparrow\uparrow$$

and the same for the other three terms in $H\psi_{\text{gs}}$. So the expectation value of energy becomes

$$\begin{aligned} \langle E \rangle &= \langle \psi_{++}\uparrow\uparrow + \psi_{+-}\uparrow\downarrow + \psi_{-+}\downarrow\uparrow + \psi_{--}\downarrow\downarrow \\ &\quad | (H\psi_{++})\uparrow\uparrow + (H\psi_{+-})\uparrow\downarrow + (H\psi_{-+})\downarrow\uparrow + (H\psi_{--})\downarrow\downarrow \rangle \end{aligned}$$

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, $\langle \psi_{\text{gs}}|\psi_{\text{gs}} \rangle = 1$, 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 $\psi_{\text{gs},0}$ with the lowest energy E_{gs} , their individual contributions to the energy will be given by $\langle \psi_{\pm\pm}|H\psi_{\pm\pm} \rangle = E_{\text{gs}}\langle \psi_{\pm\pm}|\psi_{\pm\pm} \rangle$, the lowest possible. Then the total energy $\langle E \rangle$ will be E_{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 chapter 12.1.6. This note ignored that, as do most other derivations in this book.

A.29 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 A.4; 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.

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 $I + N - 1$



Figure A.4: Bosons in single-particle-state boxes.

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.

A.30 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 [17] 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 non-unit 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.

A.31 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 , (A.30). 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.

A.32 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 5.1. 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 5.11.

Now the octant of the sphere has a “volume” (in wave number space, not a physical volume)

$$\text{octant volume: } \frac{1}{8} \frac{4}{3} \pi k^3$$

Conversely, every wave number point is the top-left front corner of a little block of “volume”

$$\text{single state volume: } \Delta k_x \Delta k_y \Delta k_z$$

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.”

$$\text{number of spatial states inside: } \frac{\pi k^3}{6 \Delta k_x \Delta k_y \Delta k_z}$$

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, (5.3),

$$\Delta k_x = \frac{\pi}{\ell_x} \quad \Delta k_y = \frac{\pi}{\ell_y} \quad \Delta k_z = \frac{\pi}{\ell_z}$$

Plug this into the expression for the number of points in the octant to get:

$$\text{number of spatial states inside: } \frac{\mathcal{V}}{6\pi^2} k^3 \quad (\text{A.34})$$

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 dN in a small wave number range dk is then:

$$dN = \mathcal{V} \mathcal{D}_k dk \quad \mathcal{D}_k = \frac{2s + 1}{2\pi^2} k^2$$

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 $E^p = \hbar k^2 / 2m$. That gives:

$$dN = \mathcal{V} \mathcal{D} dE^p \quad \mathcal{D} = \frac{2s + 1}{4\pi^2} \left(\frac{2m}{\hbar^2} \right)^{3/2} \sqrt{E^p}$$

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 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 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 5.11. 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 5.12. In this case,

there are no states inside the sphere at all if k is less than Δk_y . Regardless of what (A.34) claims. In the range $\Delta k_y < k < 2\Delta k_y$, illustrated by the red sphere in figure 5.12, 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

$$\text{spatial states per plate: } \frac{A}{4\pi} \left[k^2 - \left(n_y \frac{\pi}{\ell_y} \right)^2 \right] \quad \text{if } [\dots] > 0 \quad (\text{A.35})$$

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 (A.35). For still larger values of k , just keep summing plates as long as the expression between the square brackets in (A.35) 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 5.13. 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

$$\text{spatial states per line: } \frac{\ell}{\pi} \left[k^2 - \left(n_y \frac{\pi}{\ell_y} \right)^2 - \left(n_z \frac{\pi}{\ell_z} \right)^2 \right]^{1/2} \quad \text{if } [\dots] > 0 \quad (\text{A.36})$$

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 5.14, the number of states simply becomes the number of points inside the sphere:

$$\text{spatial states per point: } 1 \quad \left[k^2 - \left(n_x \frac{\pi}{\ell_x} \right)^2 - \left(n_y \frac{\pi}{\ell_y} \right)^2 - \left(n_z \frac{\pi}{\ell_z} \right)^2 \right] > 0 \quad (\text{A.37})$$

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.

A.33 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 dt , a cylinder of photons of volume $Acdt$ 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 cdt . The fraction of photons with velocity directions within in a spherical element $\sin \theta d\theta d\phi$ will be $\sin \theta d\theta d\phi / 4\pi$. The amount of these photons that exits will be those in a skewed cylinder of volume $A c \cos \theta dt$. To get the energy involved multiply by ρ . So the energy leaving in this small range of velocity directions is

$$\rho Acdt \cos \theta \frac{\sin \theta d\theta d\phi}{4\pi}$$

Integrate over all ϕ and θ up to 90 degrees to get $\frac{1}{4}\rho Acdt$ 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.

A.34 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_g and emissivity e_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:

$$\dot{E} = e_g B + (1 - a_g)\dot{E} \implies \dot{E}/B = e_g/a_g$$

Similarly for the radiation going up from the material being examined:

$$\dot{E} = eB + (1 - a)\dot{E} \implies \dot{E}/B = e/a$$

By comparing the two results, $e/a = e_g/a_g$. 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, [25]. 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.

A.35 The thermionic emission equation

This note derives the thermionic emission equation for a typical metal following [29, 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 section 5.22.3 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

$$j = \rho ev_x$$

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

$$E_{\text{esc}}^p = \mu + e\varphi_w$$

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 dE^p above E_{esc}^p

can be found as

$$e^{-(e\varphi_w + E^p - E_{\text{esc}}^p)/k_B T} \frac{2}{4\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \sqrt{E^p} dE^p$$

That is because the initial exponential is a rewritten Maxwell-Boltzmann distribution (5.21) 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 (5.6).

Normally, the typical thermal energy $k_B T$ is very small compared to the minimum energy $e\varphi_w$ 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_{esc}^p the amount of electrons very quickly becomes much smaller still. Therefore only a very small range of energies above the minimum energy E_{esc}^p 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 $\sqrt{p^2 - p_{\text{esc}}^2}$, has enough x -momentum for the electron to escape, so

$$f_{\text{esc}} \approx \frac{\pi \sqrt{p^2 - p_{\text{esc}}^2}^2}{4\pi p^2} \approx \frac{E^p - E_{\text{esc}}^p}{4E^p}$$

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, $E^p \approx \frac{1}{2}m_e v_x^2$, which can be used to express v_x in terms of energy.

Putting it all together, the current density becomes

$$j = \int_{E^p=E_{\text{esc}}^p}^{\infty} e^{-(e\varphi_w + E^p - E_{\text{esc}}^p)/k_B T} \frac{2}{4\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \sqrt{E^p} \frac{E^p - E_{\text{esc}}^p}{4E^p} \left(\frac{2E^p}{m_e} \right)^{1/2} dE^p$$

Rewriting in terms of a new integration variable $u = (E^p - E_{\text{esc}}^p)/k_B T$ gives the thermionic emission equation.

If an external electric field E_{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:

$$V = -\frac{e^2}{16\pi\epsilon_0 d} - eE_{\text{ext}}d + \text{constant}$$

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 $\sqrt{e^3 E / 4\pi\epsilon_0}$.

A.36 Explanation of the band gaps

Section 5.21 explained the band gaps in spectra qualitatively as the remnants of the discrete energy states of the individual atoms. However, if you start from the free-electron gas point of view, it is much less clear why and when addition

of a bit of crystal potential would produce band gaps. This note explores that question based on the Kronig & Penney model.

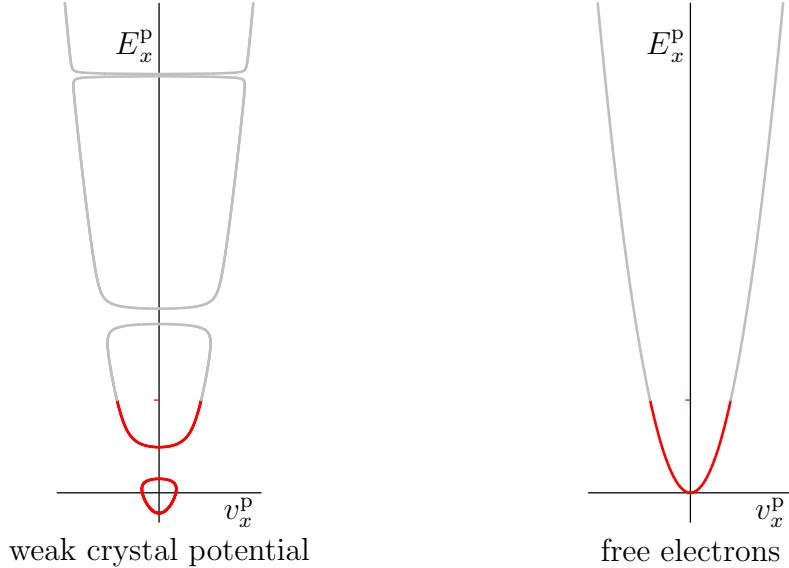


Figure A.5: Spectrum for a weak potential.

To allow an easier comparison with the free-electron gas solutions, the drops in potentials will be greatly reduced compared to figure 5.23. That results in the spectrum shown in figure A.5. The figure also shows the corresponding free-electron spectrum at the right.

You would expect that the relatively small potential drops would have little effect at high energies. Therefore, at high energies the two spectra should agree. And so they do, mostly. But periodically there is still a very narrow band gap, however high the energy. And at those gaps the electron velocity plunges sharply to zero.

To understand why this happens, it is necessary to revert from the Bloch waves to the more basic underlying real energy eigenfunctions. These more basic eigenfunctions can be obtained from the real and imaginary parts of the Bloch waves. In particular, for free electrons the energy eigenfunctions can be written as some unimportant constant times

$$e^{ik_x x} = \cos(k_x x) + i \sin(k_x x)$$

It are the cosine and sine energy eigenfunctions, not the exponentials, that explain the band gaps.

In the presence of the crystal potential, the eigenfunctions are no longer sines and cosines. Figure A.6 shows the first few energy eigenfunctions for a

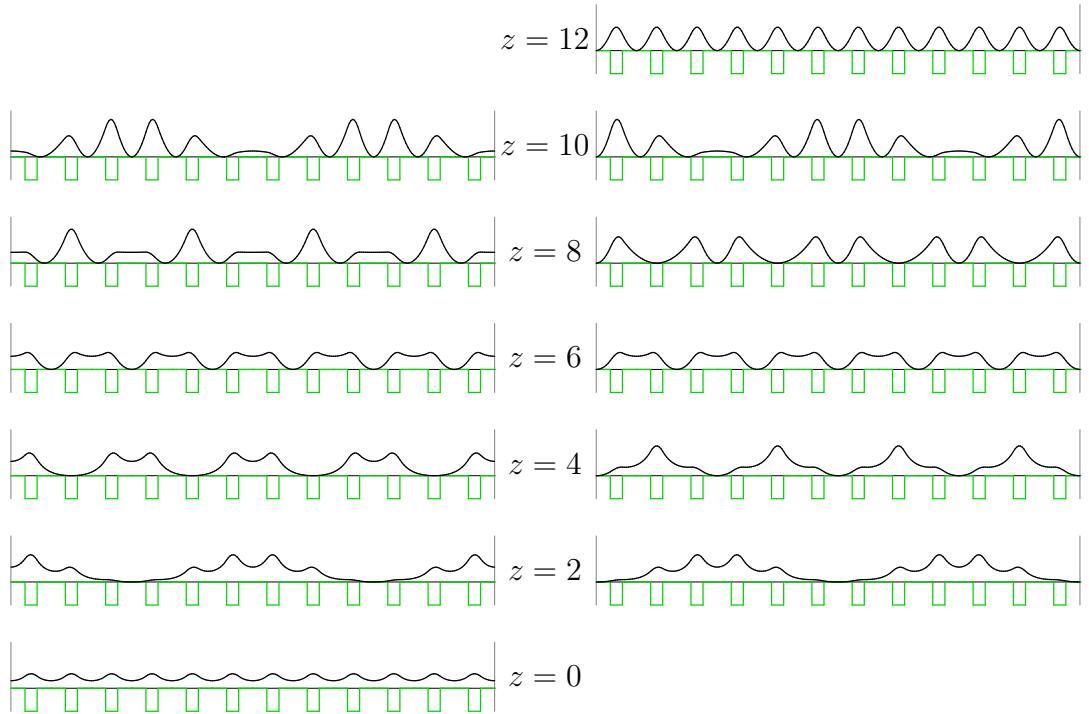
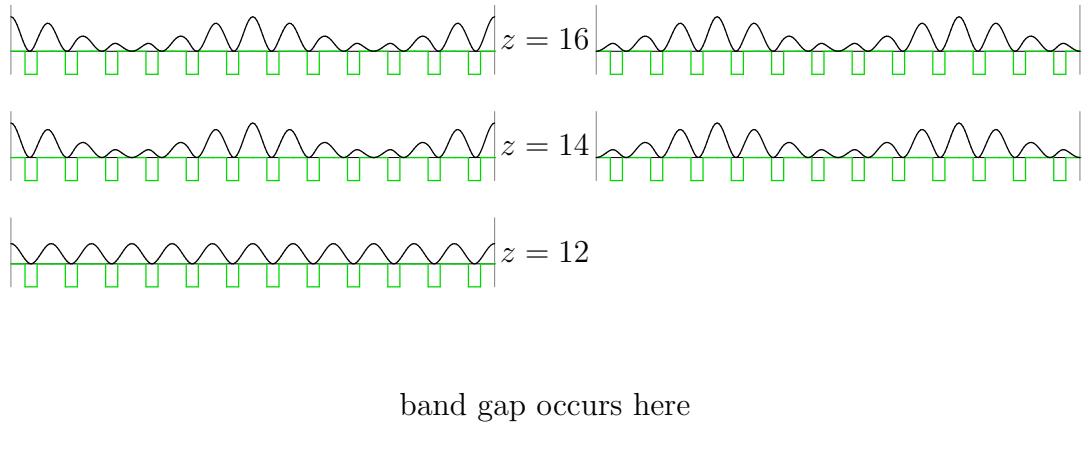


Figure A.6: The 17 real wave functions of lowest energy for a small one-dimensional periodic box with only 12 atomic cells. Black curves show the square wave function, which gives the relative probability of finding the electron at that location.

12 “atom” periodic crystal. The crystal potential is shown in green. The black curves show the energy eigenfunctions. More precisely, they show the square magnitudes of the energy eigenfunctions. The square magnitude is shown since it gives the probability of finding the electron at that location.

Consider first the eigenfunction with k_x zero. It is shown in the very bottom of figure A.6. The free-electron eigenfunction would be $\cos(0)$, which is a constant. In the presence of the crystal potential however, the eigenfunction is no longer a constant, as figure A.6 shows. It develops a ripple. The wave function now has peaks at the locations of the drops in potential. That makes it more likely for the electron to be at a drop in potential. The electron can lower its potential energy that way. If the ripple becomes too strong however, it would increase the kinetic energy more than the potential energy decreases.

Next consider the eigenfunctions for which $k_x = \pi/d_x$. For that wave number, the free-electron eigenfunctions $\cos(k_x x)$ and $\sin(k_x x)$ have a period that is twice the atomic period d_x . In particular, the energy eigenfunctions cross zero every atomic period. For a 12 atom crystal, that means that the eigenfunction has 12 zeros. In the presence of a crystal potential, the eigenfunctions are no longer a cosine and sine, but the period stays equal to $2d_x$. The number of zero crossings does not change.

It can be shown from the mathematical theory of equations like the one-dimensional Hamiltonian eigenvalue problem that the energy eigenfunctions remain arranged by zero crossings. The more crossings, the higher the energy.

The solutions in the presence of a crystal potential are marked as $z = 12$ in figure A.6. Note that what used to be the sine, the antisymmetric energy eigenfunction, has all its peaks at the drops in potential. That lowers the potential of this eigenfunction considerably. On the other hand, what used to be the cosine has all its zeros at the drops in potential. That means that for this eigenfunction, the electron has very little chance of being found at a location of low potential. Therefore the cosine solution has a much higher energy than the sine solution.

That produces the band gap. The energy of the sine-type solution gives the top energy of the lowest band in the spectrum figure A.5. The energy of the cosine-type solution is the bottom energy of the second band.

The next lower value of k_x corresponds to wave functions with 10 zeros instead of 12. That makes a big difference because there is no way to align 10 zeros with 12 atomic periods. The two eigenfunctions are now just shifted versions of each other and have the same energy. The latter is unavoidable. If you take any eigenfunction and shift it over an atomic period, it is still an eigenfunction, with the same energy. But it cannot be just a multiple of the unshifted eigenfunction because that would require the shifted and unshifted eigenfunction to have the same zeros. If you distribute 10 zeros over 12 atomic periods, some periods end up with no zeros and others with at least one. Shift it

over a period. and some periods must change their number of zeros. And if the shifted and unshifted eigenfunctions are different, then they are a complete set of eigenfunctions at that energy. Any other is just a combination of the two.

So no energy difference exists between the two $z = 10$ solutions and therefore no energy gap. But you might still wonder about the possibility of an energy gap between the two $z = 10$ solutions and the $z = 12$ solution of lowest energy. It does not happen. The $z = 10$ energy has to stay below the $z = 12$ one, but the eigenfunctions struggle to achieve that. By sitting right on top of every potential drop, the $z = 12$ eigenfunction is highly effective in lowering its potential energy. The $z = 10$ solutions cannot do the same because 10 zeros cannot properly center between 12 potential drops. The $z = 10$ solutions instead slowly modulate their amplitudes so that the amplitude is high where the peaks are on top of the potential drops, and low where they are not. That has the same effect of increasing the probability that the electron can be found at low energy. The modulation however forces the energy eigenfunctions to give up some of their kinetic energy advantage compared to the $z = 12$ solutions. So the energies become closer rather than further apart. Since the changes in energy are a measure of the propagation velocity, the velocity plunges to zero.

The second $z = 12$ solution is unusually effective in avoiding the regions of low potential energy, and the $z = 14$ solutions have to keep up with that.

If you actually want to show mathematically that the propagation velocity is indeed zero at the band gaps, you can do it using a linear algebra approach. 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 to find the two columns of G .

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

$$0 = \psi_1\psi''_2 - \psi_2\psi''_1 = (\psi_1\psi'_2 - \psi'_1\psi_2)'$$

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

$$e^{ik_x d_x} \quad \text{and} \quad e^{-ik_x d_x}$$

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 (1.5). That is less than 2 in magnitude. So the condition for valid periodic eigenfunctions becomes

$$\text{trace}(G) = 2 \cos(k_x d_x) \quad k_x d_x = \frac{n_x}{N_x} 2\pi$$

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

$$v_x^p = \frac{dE_x^p}{d\hbar k_x} = \frac{1}{\hbar} \frac{d2 \cos(k_x d_x)}{dk_x} \Bigg/ \frac{d\text{trace}(G)}{dE_x^p} = 0$$

because the cosine is at its ± 1 extrema at the gap energies.

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.

Since the eigenfunctions at the band gap have zero propagation velocity, the electrons in these states cannot move through the “crystal.” If you train an electron beam with such a wave number onto the crystal, the beam will be totally reflected. This can be verified using the so-called “Bragg” reflection

theory of wave mechanics. Indeed, the fact that the crystal spacing is a half-integer multiple of the wave lengths that are reflected is a standard Bragg result. It can be easily derived if you wave your hands a lot, chapter 8.7.2. It then provides an intuitive justification for some of the features of the band structure, in particular for the fact that the velocity is zero at the band gaps.

A.37 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 ν_e . Therefore the number of electrons in the conduction band per unit volume is given by

$$i_e = \int_{E_c^p}^{E_{\text{top}}^p} \mathcal{D} \nu_e dE^p$$

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 ν_e the Maxwell-Boltzmann expression (5.33) can be used, since the number of electrons per state is invariably small. And for the density of states the expression (5.6) 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 ν_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 ν_h from (5.34) and integrate over the valence band energies.

A.38 Thermoelectric effects

A.38.1 Peltier and Seebeck coefficient ballparks

The approximate expressions for the semiconductor Peltier coefficients come from [19]. Straub *et al* (App. Phys. Let. 95, 052107, 2009) note that to better

approximation, $\frac{3}{2}k_B T$ should be $(\frac{5}{2} + r)k_B T$ 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 $\overline{E^p v_x}$. Here v_x is the electron velocity in the x -direction, $-e$ the electron charge, E^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 section 5.20. The conduction occurred since the Fermi sphere got displaced a bit towards the right in the wave number space figure 5.17. Call the small amount of displacement k_d . Assume for simplicity that in a coordinate system $k_x k_y k_z$ 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_d along the k_x axis, the velocities and energies of the single-particle states are now given by

$$v_x = \frac{\hbar}{m}(k_x + k_d) \quad E^p = \frac{\hbar^2}{2m} [(k_x + k_d)^2 + k_y^2 + k_z^2] = \frac{\hbar^2}{2m} (k^2 + 2k_x k_d + k_d^2)$$

To simplify the notations, the above expressions will be abbreviated to

$$v_x = C_v(k_x + k_d) \quad E^p = C_E(k^2 + 2k_x k_d + k_d^2)$$

In this notation, the average charge and energy flows per electron become

$$\overline{-ev_x} = \overline{-eC_v(k_x + k_d)} \quad \overline{E^p v_x} = \overline{C_E(k^2 + 2k_x k_d + k_d^2) C_v(k_x + k_d)}$$

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_d , can be taken out of the averages. So the flows simplify to

$$\overline{-ev_x} = -eC_v k_d \quad \overline{E^p v_x} = C_E(2\overline{k_x^2} + \overline{k^2}) C_v k_d$$

where the term cubically small in k_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_{ave}^p in the absence of conduction. Also, by definition $C_v k_d$ is the drift velocity v_d that produces the current. Therefore:

$$\overline{-ev_x} = -ev_d \quad \overline{v_x E^p} = \frac{5}{3} E_{ave}^p v_d$$

Note that if you would simply have ballparked the average of $v_x E^P$ as the average of v_x times the average of E^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 $\overline{v_x E^P}$. The Peltier coefficient is the ratio of that heat flow to the charge flow:

$$\Pi = \frac{\overline{v_x(E^P - \mu)}}{-ev_x} = \frac{\frac{5}{3}E_{\text{ave}}^P - \mu}{-e}$$

If you plug in the expressions for the average energy per electron and the chemical potential found in note {A.85}, 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 note {A.85}:

$$-e d\varphi = -d \frac{\pi^2}{4} \frac{(k_B T)^2}{E_F^P}$$

Divide by $e dT$ to get the Seebeck coefficient.

A.38.2 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

$$M = \Pi \sqrt{\frac{\sigma}{\kappa T}}$$

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 Π 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, [6],

$$M_{\text{net}} = M_A \frac{\sqrt{\kappa_A/\sigma_A}}{\sqrt{\kappa_A/\sigma_A} + \sqrt{\kappa_B/\sigma_B}} - M_B \frac{\sqrt{\kappa_B/\sigma_B}}{\sqrt{\kappa_A/\sigma_A} + \sqrt{\kappa_B/\sigma_B}}$$

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 $-M_B$ 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.38.3 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 the fact 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 section 5.22.2.

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.38.4 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:

$$q = f_1 \left(\frac{dT}{dx}, \frac{d\varphi_\mu}{dx} \right) \quad j = f_2 \left(\frac{dT}{dx}, \frac{d\varphi_\mu}{dx} \right)$$

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 rarified 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:

$$q = A_{11} \frac{dT}{dx} + A_{12} \frac{d\varphi_\mu}{dx} \quad j = A_{21} \frac{dT}{dx} + A_{22} \frac{d\varphi_\mu}{dx}$$

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:

$$q = -(\kappa + \Pi \Sigma \sigma) \frac{dT}{dx} - \Pi \sigma \frac{d\varphi_\mu}{dx} \quad j = -\Sigma \sigma \frac{dT}{dx} - \sigma \frac{d\varphi_\mu}{dx} \quad (\text{A.38})$$

This *defines* the heat conductivity κ , the electrical conductivity σ , the Seebeck coefficient Σ and the Peltier coefficient Π 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

$$\boxed{\frac{d\varphi_\mu}{dx} = -\Sigma \frac{dT}{dx}} \quad (\text{A.39})$$

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

$$\varphi_{\mu,2} - \varphi_{\mu,1} = - \int_1^2 \Sigma dT \quad (\text{A.40})$$

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.38) produces

$$\boxed{q = -\kappa \frac{dT}{dx} + \Pi j} \quad (\text{A.41})$$

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 dx 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:

$$\dot{e} = -\frac{dj_E}{dx} = -\frac{dq}{dx} - j \frac{d\varphi_\mu}{dx}$$

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.41) for q and from the second of (A.38) for the gradient of the potential gives:

$$\dot{e} = \frac{d}{dx} \left(\kappa \frac{dT}{dx} \right) + \frac{j^2}{\sigma} - \mathcal{K} \frac{dT}{dx} \quad \mathcal{K} \equiv \frac{d\Pi}{dT} - \Sigma$$

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.38) 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.38.5 Charge locations in thermoelectrics

The statement that the charge density is neutral inside the material comes from [[7]].

A simplified macroscopic derivation can be given based on the thermoelectric equations (A.38). 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.38) 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 10.4. 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.38) 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 10.4. Any net charge must accumulate at the surfaces.

A.38.6 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 9, 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.

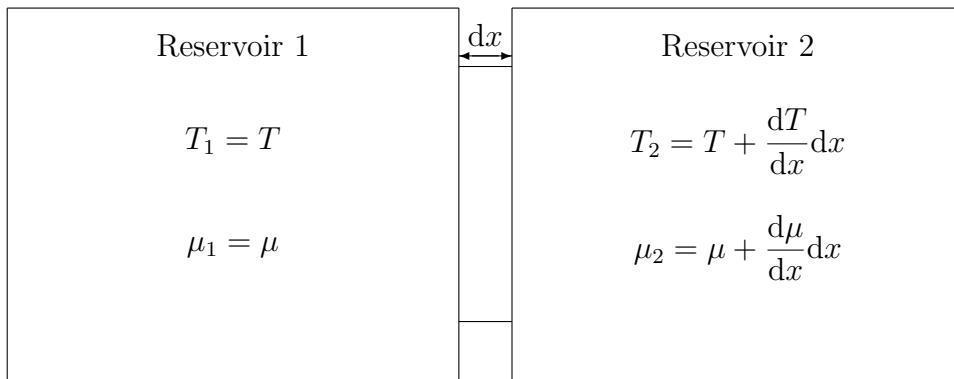


Figure A.7: Analysis of conduction.

The primary question is what is going on in a single short segment dx of such a bar. Here dx 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.7. 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:

$$j_I = -\frac{dI_1}{dt} = \frac{dI_2}{dt}$$

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:

$$j_E = -\frac{dE_1}{dt} = \frac{dE_2}{dt}$$

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:

$$q_2 = T_2 \frac{dS_2}{dt}$$

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 9.12 for more details, including generalization to more than one particle type. A combination of the first and second laws of thermodynamics produces

$$T d\bar{s} = d\bar{e} + P d\bar{v} \quad S = \bar{s}I \quad E = \bar{e}I \quad V = \bar{v}I$$

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:

$$T dS = T d(\bar{s}I) = T(d\bar{s})I + T\bar{s}(dI) = (d\bar{e} + P d\bar{v})I + T\bar{s}(dI)$$

That can be rewritten as

$$T dS = dE + PdV - (\bar{e} + P\bar{v} - T\bar{s})dI$$

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 9.13 explains that it is the same as the chemical potential μ in the distribution laws. Therefore:

$$T dS = dE + PdV - \mu dI$$

(Chapter 9.13 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 section 5.14.)

Using the above expression for the change in entropy in the definition of the heat flux gives, noting that the volume is constant,

$$q_2 = \frac{dE_2}{dt} - \mu_2 \frac{dI_2}{dt} = j_E - \mu_2 j_I$$

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.7 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,

$$\frac{dS_{\text{net}}}{dt} = \frac{dS_2}{dt} + \frac{dS_1}{dt}$$

By definition of the heat flux,

$$\frac{dS_{\text{net}}}{dt} = \frac{q_2}{T_2} - \frac{q_1}{T_1}$$

Substituting in the expression for the heat flux in terms of the energy and particle fluxes gives

$$\frac{dS_{\text{net}}}{dt} = \left(\frac{1}{T_2} j_E - \frac{\mu_2}{T_2} j_I \right) - \left(\frac{1}{T_1} j_E - \frac{\mu_1}{T_1} j_I \right)$$

Since the area of the bar is one, its volume is dx . Therefore, the entropy generation per unit volume is:

$$\frac{1}{dx} \frac{dS_{\text{net}}}{dt} = \frac{d1/T}{dx} j_E + \frac{d-\mu/T}{dx} j_I \quad (\text{A.42})$$

This used the fact that since dx is infinitesimal, any expression of the form $(f_2 - f_1)/dx$ 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:

$$j_E = L_{11} \frac{d1/T}{dx} + L_{12} \frac{d-\mu/T}{dx} \quad j_I = L_{21} \frac{d1/T}{dx} + L_{22} \frac{d-\mu/T}{dx} \quad (\text{A.43})$$

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.43) 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.38) given earlier, you find that the Onsager relation $L_{12} = L_{21}$ translates into the second Kelvin relation

$$\Pi = \Sigma T$$

That allows you to clean up the Kelvin coefficient to the first Kelvin relationship:

$$\kappa \equiv \frac{d\Pi}{dT} - \Sigma = T \frac{d\Sigma}{dT} = \frac{d\Sigma}{d \ln T}$$

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.39 Why energy eigenstates are stationary

The probability of measuring an eigenvalue a_i for any arbitrary physical quantity a is according to the orthodox interpretation the square magnitude of the coefficient of the corresponding eigenfunction α_i . This coefficient can be found as the inner product $\langle \alpha_i | \Psi \rangle$, which for a stationary state is $\langle \alpha_i | c_{\vec{n}}(0) e^{-iE_{\vec{n}}t/\hbar} \psi_{\vec{n}} \rangle$ and taking the square magnitude kills off the time-dependent exponential. So

the probability of measuring any value for any physical quantity remains exactly the same however long you wait.

It is of course assumed that the operator A does not explicitly depend on time. Otherwise its time variation would be automatic. (The eigenfunctions would depend on time.)

A.40 Better description of two-state systems

The given description of two state systems is a bit tricky, since the mentioned states of lowest and highest energy are only approximate energy eigenfunctions. 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 anti-symmetric 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. (It are the small differences in energy that drive the *nontrivial* part of the unsteady evolution.)

A.41 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 $H\Psi/i\hbar$ 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 $-i$.

A.42 The virial theorem

The virial theorem says that the expectation value of the kinetic energy of stationary states is given by

$$\boxed{\langle T \rangle = \frac{1}{2} \langle \vec{r} \cdot \nabla V \rangle} \quad (\text{A.44})$$

Note that according to the calculus rule for directional derivatives, $\vec{r} \cdot \nabla V = r \partial V / \partial r$.

For the $V = \frac{1}{2}c_x x^2 + \frac{1}{2}c_y y^2 + \frac{1}{2}c_z z^2$ potential of a harmonic oscillator, $x \partial V / \partial x + y \partial V / \partial y + z \partial V / \partial z$ produces $2V$. So for energy eigenstates of the

harmonic oscillator, the expectation value of kinetic energy equals the one of the potential energy. And since their sum is the total energy $E_{n_x n_y n_z}$, each must be $\frac{1}{2}E_{n_x n_y n_z}$.

For the $V = \text{constant}/r$ potential of the hydrogen atom, $r\partial V/\partial r$ produces $-V$. So the expectation value of kinetic energy equals minus one half the one of the potential energy. And since their sum is the total energy E_n , $\langle T \rangle = -E_n$ and $\langle V \rangle = 2E_n$. Note that E_n is negative, so that the kinetic energy is positive as it should be.

To prove the virial theorem, work out the commutator in

$$\frac{d\langle \vec{r} \cdot \vec{p} \rangle}{dt} = \frac{i}{\hbar} \langle [H, \vec{r} \cdot \vec{p}] \rangle$$

using the formulae in chapter 3.4.4,

$$\frac{d\langle \vec{r} \cdot \vec{p} \rangle}{dt} = 2\langle T \rangle - \langle \vec{r} \cdot \nabla V \rangle,$$

and then note that the left hand side above is zero for stationary states, (in other words, states with a definite total energy).

A.43 The energy-time uncertainty relationship

As mentioned in chapter 3.4.3, Heisenberg's formulae

$$\Delta p_x \Delta x \geq \frac{1}{2}\hbar$$

relating the typical uncertainties in momentum and position is often very convenient for qualitative descriptions of quantum mechanics, especially if you mis-read \geq as \approx .

So, taking a cue from relativity, people would like to write a similar expression for the uncertainty in the time coordinate,

$$\Delta E \Delta t \geq \frac{1}{2}\hbar$$

However, if you want to formally justify such an expression in general, 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 3.4.2 between the uncertainties in energy and A is:

$$\sigma_E \sigma_A \geq \frac{1}{2} |\langle [H, A] \rangle|.$$

But $|\langle [H, A] \rangle|$ is just $\hbar |d\langle A \rangle / dt|$.

So the Mandelstam-Tamm version of the energy-time uncertainty principle just *defines* the uncertainty in time to be

$$\sigma_t = \sigma_A \left/ \left| \frac{d\langle A \rangle}{dt} \right| \right..$$

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 it takes for A to change to a value sufficiently different that it will clearly show up in measurements.

A.44 The adiabatic theorem

This note derives the adiabatic theorem and then mentions some of its implications.

A.44.1 Derivation of the theorem

Consider the Schrödinger equation

$$i\hbar \frac{\partial \Psi}{\partial t} = H\Psi$$

If the Hamiltonian is independent of time, the solution can be written in terms of its eigenvalues $E_{\vec{n}}$ and eigenfunctions $\psi_{\vec{n}}$ as

$$\Psi = \sum_{\vec{n}} c_{\vec{n}} e^{i\theta_{\vec{n}}} \psi_{\vec{n}} \quad \theta_{\vec{n}} = -\frac{1}{\hbar} E_{\vec{n}} t$$

where \vec{n} stands for the quantum numbers of the eigenfunctions. But the Hamiltonian varies with time in 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

$$\Psi = \sum_{\vec{n}} c_{\vec{n}} e^{i\theta_{\vec{n}}} \psi_{\vec{n}} \quad \theta_{\vec{n}} = -\frac{1}{\hbar} \int E_{\vec{n}} dt \quad (\text{A.45})$$

However, the coefficients $c_{\vec{n}}$ can no longer be assumed to be constant. They may be different at different times.

To get an equation for their variation, plug the expression for Ψ in the Schrödinger equation:

$$i\hbar \sum_{\vec{n}} c'_{\vec{n}} e^{i\theta_{\vec{n}}} \psi_{\vec{n}} - i\hbar \sum_{\vec{n}} c_{\vec{n}} \frac{i}{\hbar} E_{\vec{n}} e^{i\theta_{\vec{n}}} \psi_{\vec{n}} + i\hbar \sum_{\vec{n}} c_{\vec{n}} e^{i\theta_{\vec{n}}} \psi'_{\vec{n}} = H \sum_{\vec{n}} c_{\vec{n}} e^{i\theta_{\vec{n}}} \psi_{\vec{n}}$$

where the primes indicate time derivatives. The middle sum in the left hand side and the right hand side cancel against each other since $\psi_{\vec{n}}$ is an eigenfunction of the Hamiltonian. For the remaining two sums, take an inner product with an arbitrary eigenfunction $\langle \psi_{\underline{\vec{n}}} \rangle$:

$$i\hbar c'_{\vec{n}} e^{i\theta_{\vec{n}}} + i\hbar \sum_{\vec{n}} c_{\vec{n}} e^{i\theta_{\vec{n}}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle = 0$$

where in the first sum only the term $\vec{n} = \underline{\vec{n}}$ survived because of the orthonormality of the eigenfunctions. Divide by $i\hbar e^{i\theta_{\vec{n}}}$ and rearrange to get

$$c'_{\underline{\vec{n}}} = - \sum_{\vec{n}} c_{\vec{n}} e^{i(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle \quad (\text{A.46})$$

This is still exact. However, the objective is now to show that in the adiabatic approximation of a slowly varying system, all terms in the sum above except the one with $\vec{n} = \underline{\vec{n}}$ can be ignored.

To do so, for simplicity assume that all eigenfunctions are fully nondegenerate, as is typical in one-dimensional problems. For example, think of a one-dimensional harmonic oscillator whose stiffness slowly changes. For such a system the eigenfunctions will change in a slow, regular way with the state of the system, making the time derivative of the eigenfunction in the sum above small. Then the complete equation implies that every arbitrary coefficient $c_{\vec{n}}$ also varies slowly with time. And now note that the change in coefficient $c_{\underline{\vec{n}}}$ can be found from the time integral of the sum. The time integral of the terms with $\vec{n} \neq \underline{\vec{n}}$ is almost exactly zero, since the exponential is periodic on a time scale that is not slow, and integrates to zero over each period. So, these terms nullify over each period of the exponential, and never succeed in making a significant contribution to the change in $c_{\underline{\vec{n}}}$.

To see this a bit more mathematically precisely, perform an integration by parts on such a term:

$$\begin{aligned} \int_{\text{period}} -\frac{i}{\hbar} (E_{\vec{n}} - E_{\underline{\vec{n}}}) e^{i(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} & \frac{\hbar c_{\vec{n}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle}{i(E_{\vec{n}} - E_{\underline{\vec{n}}})} dt = \\ & \frac{\hbar c_{\vec{n}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle}{i(E_{\vec{n}} - E_{\underline{\vec{n}}})} \Big|_{\text{start}}^{\text{end}} - \int_{\text{period}} e^{i(\theta_{\vec{n}} - \theta_{\underline{\vec{n}}})} \left(\frac{\hbar c_{\vec{n}} \langle \psi_{\underline{\vec{n}}} | \psi'_{\vec{n}} \rangle}{i(E_{\vec{n}} - E_{\underline{\vec{n}}})} \right)' dt \end{aligned}$$

If the evolution takes place over some large typical time T , both terms will be small of order $1/T^2$ over each period, due to the smallness and slow variation of the fraction. So the change in the coefficient $c_{\underline{\vec{n}}}$ that it causes during the order T evolution time is small of order $1/T$.

The remaining equation (A.46) for the coefficient $c_{\underline{\vec{n}}}$ is now readily integrated as

$$c_{\underline{\vec{n}}} = c_{\underline{\vec{n}},0} e^{i\gamma_{\underline{\vec{n}}}} \quad \gamma_{\underline{\vec{n}}} = i \int \langle \psi_{\underline{\vec{n}}} | \psi'_{\underline{\vec{n}}} \rangle dt$$

where $c_{\underline{n},0}$ is a constant that depends on the initial condition for Ψ , (and on the choice of integration constant for $\gamma_{\underline{n}}$, but usually you take $\gamma_{\underline{n}}$ zero at the initial time). This expression for the coefficients can be plugged in (A.45) to find the wave function Ψ .

Note that $\gamma_{\underline{n}}$ is real, because differentiating the normalization requirement produces

$$\langle \psi_{\vec{n}} | \psi_{\vec{n}} \rangle = 1 \implies \langle \psi'_{\vec{n}} | \psi_{\vec{n}} \rangle + \langle \psi_{\vec{n}} | \psi'_{\vec{n}} \rangle = 0$$

so the sum of the inner product plus its complex conjugate are zero. That makes it purely imaginary. It follows that the magnitude of the coefficients $c_{\vec{n}}$ does not change in time. In particular, if the system starts out in a single eigenfunction, then it stays in that eigenfunction for as long as its energy remains non-degenerate.

So what is all this fuss about the adiabatic theorem being so hard to prove? Well, it gets much messier if it cannot be assumed that all energy eigenfunctions are non-degenerate. Think of a three-dimensional harmonic oscillator, whose stiffnesses pass through a point where they are equal in all directions. There would be massive degeneracy at that point. If eigenfunctions are almost degenerate, the smallest perturbation may throw you from one onto another. Sure, when the adiabatic system is in an almost degenerate eigenstate, ambiguity in that state is to be expected. But suppose the system is in the non-degenerate ground state. Should not the adiabatic theorem still apply then, regardless of the degeneracy of states that the system is not in? Unfortunately the assumption above that the time derivatives of the wave functions are small and smoothly varying crashes.

Some handle on that problem may be obtained by getting rid of the time derivative of the eigenfunction, and that can be done by differentiating the eigenvalue problem for $\psi_{\vec{n}}$ with respect to time and then taking an inner product with $\langle \psi_{\underline{n}} |$:

$$(H - E_{\vec{n}})\psi_{\vec{n}} = 0 \implies \langle \psi_{\underline{n}} | (H - E_{\vec{n}})\psi'_{\vec{n}} \rangle + \langle \psi_{\underline{n}} | (H' - E'_{\vec{n}})\psi_{\vec{n}} \rangle = 0$$

In the first term, $H - E_{\vec{n}}$ can be taken to the other side of the inner product and the term then reduces to $E_{\underline{n}} - E_{\vec{n}}$ times the desired inner product; also, in the second term, $E'_{\vec{n}}$ can be dropped because of orthonormality. That gives

$$\langle \psi_{\underline{n}} | \psi'_{\vec{n}} \rangle = \frac{\langle \psi_{\underline{n}} | H' \psi_{\vec{n}} \rangle}{E_{\underline{n}} - E_{\vec{n}}}$$

The time derivative of the Hamiltonian does not have rapid variation in time if the system is changed slowly, even if there is degeneracy. Plugging the inner product into (A.46), that becomes

$$c'_{\vec{n}} = -c_{\underline{n}} \langle \psi_{\underline{n}} | \psi'_{\vec{n}} \rangle + \sum_{\vec{n} \neq \underline{n}} c_{\vec{n}} e^{i(\theta_{\vec{n}} - \theta_{\underline{n}})} \frac{\langle \psi_{\underline{n}} | H' \psi_{\vec{n}} \rangle}{E_{\underline{n}} - E_{\vec{n}}}$$

Note the need for the energy level $E_{\vec{n}}$ to be nondegenerate; otherwise you would be dividing by zero.

Some sources now claim the final sum can be ignored because H' is small in an adiabatic process. Unfortunately, while H' is indeed small, that is compensated for by the long evolution time. The correct reason is still the oscillating exponential factor. Maybe you are willing to believe that the bounded coefficient $c_{\vec{n}}$ and order $1/T$ inner product $\langle \psi_{\vec{n}} | H' \psi_{\vec{n}} \rangle$ do still essentially vary on the slow time scale T , so that the cancellation over periods remains valid. If not, there is a solid discussion in the original derivation by Born & Fock (Zeitschrift für Physik, Vol. 51, p. 165, 1928). More recent derivations allow the spectrum to be continuous, in which case the “energy-gap” $E_{\vec{n}} - E_{\vec{n}}$ can no longer be assumed to be larger than some nonzero amount. This note will assume it has already given much more detail than any engineer would care for.

A.44.2 Some implications

The derivation of the previous subsection gives the wave function of an adiabatic system as

$$\boxed{\Psi = \sum_{\vec{n}} c_{\vec{n},0} e^{i\gamma_{\vec{n}}} e^{i\theta_{\vec{n}}} \psi_{\vec{n}} \quad \gamma_{\vec{n}} = i \int \langle \psi_{\vec{n}} | \psi'_{\vec{n}} \rangle dt \quad \theta_{\vec{n}} = -\frac{1}{\hbar} \int E_{\vec{n}} dt} \quad (\text{A.47})$$

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.”

The geometric phase is 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 parameter λ that it depends on changes. Then the geometric phase can be written as

$$\gamma_{\vec{n}} = i \int \langle \psi_{\vec{n}} | \frac{\partial \psi_{\vec{n}}}{\partial \lambda} \rangle d\lambda \equiv \int f(\lambda) d\lambda$$

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 $\vec{\lambda} = (\lambda_1, \lambda_2, \dots)$ changes during the evolution. Then the geometric phase is

$$\gamma_{\vec{n}} = i \int \langle \psi_{\vec{n}} | \nabla_{\vec{\lambda}} \psi_{\vec{n}} \rangle \cdot d\vec{\lambda} \equiv \int f_1(\lambda_1, \lambda_2, \dots) d\lambda_1 + f_2(\lambda_1, \lambda_2, \dots) d\lambda_2 + \dots$$

and that is not necessarily zero when the system returns to the same state it started out with. 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 $\nabla_{\vec{x}} \times \vec{f} = 0$. 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 send 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.45 Symmetry eigenvalue conservation

Since a symmetry operator like R_φ commutes with the Hamiltonian H , the two have a common set of eigenfunctions. Hence, if ρ is an eigenfunction of R_φ with eigenvalue r , it can always be written as a linear combination of eigenfunctions ρ_1, ρ_2, \dots with the same eigenvalue that are also eigenfunctions of H . So the wave function is

$$c_1 e^{-iE_1 t/\hbar} \rho_1 + c_2 e^{-iE_2 t/\hbar} \rho_2 + \dots$$

which remains a linear combination of eigenfunctions with eigenvalue r , hence an eigenfunction with eigenvalue r .

(A mathematical condition for the property that two commuting operators have a complete set of common eigenvectors is that they are both diagonalizable. While R_φ is not Hermitian, it is still diagonalizable since it is unitary.)

A.46 The two-state approximation of radiation

An atom is really an infinite state system, not a two state system, and the wave function Ψ is a combination of all infinitely many eigenfunctions. But if it is assumed that the perturbation is small, and that only the coefficients a and b of ψ_L and ψ_H have non-negligible initial values, then you can ignore the effects of the other infinitely many coefficients as quadratically small: the small perturbation level insures that the other coefficients remain correspondingly small, and in addition their effect on a and b is much smaller still since the states hardly affect each other when the perturbation is small. (When the perturbation level is zero, they are energy eigenstates that evolve completely independently.)

While the other coefficients do not have a big effect on a and b if they are small, still if you start from the ground state $|a| = 1$, then b will remain

small and the other coefficients will probably be comparably small. Also, there is the likelihood that more than two coefficients have a significant magnitude. Typically, to find out what really happens to a complete system, you need to separately evaluate all possible transitions as two state systems, and then sum all the effects you get together.

A.47 Selection rules

This note derives the selection rules for electric dipole transitions between two hydrogen states $\psi_L = \psi_{n_L l_L m_L} \uparrow$ and $\psi_H = \psi_{n_H l_H m_H} \uparrow$. Some selection rules for forbidden transitions are also derived.

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,

$$\langle \psi_L | r_i | \psi_H \rangle \neq 0$$

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 \hat{L}_z , the following commutator is useful

$$\langle \psi_L | [r_i, \hat{L}_z] | \psi_H \rangle = \langle \psi_L | r_i \hat{L}_z - \hat{L}_z r_i | \psi_H \rangle$$

For the $r_i \hat{L}_z$ term, the operator \hat{L}_z acts on ψ_H and produces a factor $m_H \hbar$, while for the $\hat{L}_z r_i$ term, \hat{L}_z can be taken to the other side of the inner product and then acts on ψ_L , producing a factor $m_L \hbar$. So:

$$\langle \psi_L | [r_i, \hat{L}_z] | \psi_H \rangle = (m_H - m_L) \hbar \langle \psi_L | r_i | \psi_H \rangle \quad (\text{A.48})$$

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 3.4.4 $[z, \hat{L}_z]$ is zero. That means according to equation (A.48) above that the dipole moment $\langle \psi_L | z | \psi_H \rangle$ in the right hand side will have to be zero too, unless $m_H = m_L$. So the first conclusion is that the z component of the electric field does not do anything unless the magnetic quantum numbers are equal. One down, two to go.

For the x and y components, from chapter 3.4.4

$$[x, \hat{L}_z] = -i\hbar y \quad [y, \hat{L}_z] = i\hbar x$$

Plugging that into (A.48) produces

$$-i\hbar\langle\psi_L|y|\psi_H\rangle = (m_H - m_L)\hbar\langle\psi_L|x|\psi_H\rangle \quad i\hbar\langle\psi_L|x|\psi_H\rangle = (m_H - m_L)\hbar\langle\psi_L|y|\psi_H\rangle$$

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

$$i\hbar\langle\psi_L|x|\psi_H\rangle = \frac{(m_H - m_L)^2\hbar^2}{-i\hbar}\langle\psi_L|x|\psi_H\rangle$$

and if the x dipole moment is nonzero, that requires that $(m_H - m_L)^2$ is one, so $m_H = m_L \pm 1$. It follows that dipole transitions can only occur if $m_H = m_L$, through the z component of the electric field, or if $m_H = m_L \pm 1$, 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 use the commutator $[r_i, \hat{L}^2]$ since the quantum number l is produced by \hat{L}^2 . However, according to chapter 3.4.4, (3.50), this commutator will bring in the $\hat{\vec{r}} \times \hat{\vec{L}}$ operator, which cannot be handled. The commutator that works is the second of (3.55):

$$[[r_i, \hat{L}^2], \hat{L}^2] = 2\hbar^2(r_i\hat{L}^2 + \hat{L}^2r_i)$$

where by the definition of the commutator

$$[[r_i, \hat{L}^2], \hat{L}^2] = (r_i\hat{L}^2 - \hat{L}^2r_i)\hat{L}^2 - \hat{L}^2(r_i\hat{L}^2 - \hat{L}^2r_i) = r_i\hat{L}^2\hat{L}^2 - 2\hat{L}^2r_i\hat{L}^2 + \hat{L}^2\hat{L}^2r_i$$

Evaluating $\langle\psi_L|[[r_i, \hat{L}^2], \hat{L}^2]|\psi_H\rangle$ according to each expression and equating the two gives

$$2\hbar^2[l_H(l_H + 1) + l_L(l_L + 1)]\langle\psi_L|r_i|\psi_H\rangle = \hbar^2[l_H(l_H + 1) - l_L(l_L + 1)]^2\langle\psi_L|r_i|\psi_H\rangle$$

For $\langle\psi_L|r_i|\psi_H\rangle$ to be nonzero, the numerical factors in the left and right hand sides must be equal,

$$2[l_H(l_H + 1) + l_L(l_L + 1)] = [l_H(l_H + 1) - l_L(l_L + 1)]^2$$

The right hand side is obviously zero for $l_H = l_L$, so $l_H - l_L$ can be factored out of it as

$$[l_H(l_H + 1) - l_L(l_L + 1)]^2 = (l_H - l_L)^2(l_H + l_L + 1)^2$$

and the left hand side can be written in terms of these same factors as

$$2[l_H(l_H + 1) + l_L(l_L + 1)] = (l_H - l_L)^2 + (l_H + l_L + 1)^2 - 1$$

Combining the two results and simplifying gives

$$[(l_H - l_L)^2 - 1][(l_H + l_L + 1)^2 - 1] = 0$$

The second factor is only zero if $l_H = l_L = 0$, but then $\langle \psi_L | r_i | \psi_H \rangle$ 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 $l_H = l_L \pm 1$.

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.

Now consider the effect of the magnetic field on transitions. 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 10.6

$$H_1 = \frac{e}{2m_e} \vec{B} \cdot (\hat{\vec{L}} + 2\hat{\vec{S}})$$

Note that now electron spin must be included in the discussion.

According to this perturbation Hamiltonian, the perturbation coefficient H_{HL} for the z -component of the magnetic field is proportional to

$$\langle \psi_L | \hat{L}_z + 2\hat{S}_z | \psi_H \rangle$$

and that is zero because $\psi_H \downarrow$ is an eigenfunction of both operators and orthogonal to $\psi_L \downarrow$. 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

$$\langle \psi_L | \hat{L}_x | \psi_H \rangle + 2\langle \psi_L | \hat{S}_x | \psi_H \rangle$$

According to chapter 10.1.10, the effect of \hat{L}_x on a state with magnetic quantum number m_H is to turn it into a linear combination of two similar states with magnetic quantum numbers $m_H + 1$ and $m_H - 1$. Therefore, for the first inner product above to be nonzero, m_L will have to be either $m_H + 1$ or $m_H - 1$. Also the orbital azimuthal momentum numbers l will need to be the same, and so will the spin magnetic quantum numbers m_s . For the second inner product, it are the spin magnetic quantum numbers that have to be different by one unit, while the orbital magnetic quantum numbers must now be equal. So, all together

$$l_H = l_L \quad m_H = m_L \text{ or } m_L \pm 1 \quad m_{s,H} = m_{s,L} \text{ or } m_{s,L} \pm 1 \quad (\text{A.49})$$

and either the orbital or the spin magnetic quantum numbers must be unequal.

The logical way to proceed to electric quadrupole transitions would be to expand the electric field in a Taylor series in terms of y :

$$\vec{E} = \hat{k}E_0 \cos(\omega(t - y/c) - \phi) \approx \hat{k}E_0 \cos(\omega t - \phi) + \hat{k}\frac{\omega}{c}E_0 \sin(\omega t - \phi)y$$

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 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

$$\vec{A} = -\hat{k}\frac{1}{\omega}E_0 \sin(\omega(t - y/c) - \phi) \quad \vec{E} = -\frac{\partial \vec{A}}{\partial t} \quad \vec{B} = \nabla \times \vec{A}$$

In terms of the vector potential, the perturbation Hamiltonian is, chapter 10.3 and 10.6, and assuming a weak field,

$$H_1 = \frac{e}{2m_e}(\vec{A} \cdot \hat{\vec{p}} + \hat{\vec{p}} \cdot \vec{A}) + \frac{e}{m_e}\hat{\vec{S}} \cdot \vec{B}$$

Ignoring the spatial variation of \vec{A} , this expression produces an Hamiltonian perturbation coefficient

$$H_{\text{HL}} = -\frac{e}{m_e\omega}E_0 \sin(\omega t - \phi)\langle\psi_L|p_z|\psi_H\rangle$$

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, \hat{p}_z is equal to the commutator $i m_e[H_0, z]/\hbar$ 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

$$\frac{e}{m_e}\frac{1}{c}E_0 \cos(\omega t - \phi)y\hat{p}_z$$

The factor $y\hat{p}_z$ can be trivially rewritten to give

$$\frac{e}{2m_e}\frac{1}{c}E_0 \cos(\omega t - \phi)(y\hat{p}_z - z\hat{p}_y) + \frac{e}{2m_e}\frac{1}{c}E_0 \cos(\omega t - \phi)(y\hat{p}_z + z\hat{p}_y)$$

The first term has already been accounted for in the magnetic dipole transitions discussed above, because the factor within parentheses is \hat{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 as long as the atom is small compared to the wave length c/ω of the wave. Therefore they will not make a difference unless the electric dipole transition is forbidden.

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 $\hat{p}_i = im_e[H_0, r_i]/\hbar$ and H_0 does not change the eigenfunctions, just their coefficients. Therefore $y\hat{p}_z + z\hat{p}_y$ produces only eigenfunctions with azimuthal quantum number l either equal to l_H or to $l_H \pm 2$, depending on whether the two unit changes reinforce or cancel each other. Furthermore, it produces only eigenfunctions with m equal to $m_H \pm 1$. However, $x\hat{p}_y + y\hat{p}_x$, corresponding to a wave along another axis, will produce values of m equal to m_H or to $m_H \pm 2$. Therefore the selection rules become:

$$l_H = l_L \text{ or } l_L \pm 2 \quad m_H = m_L \text{ or } m_L \pm 1 \text{ or } m_L \pm 2 \quad m_{s,H} = m_{s,L} \quad (\text{A.50})$$

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 $l_H = 0$ to $l_L = 0$ are not possible because of spherical symmetry.

A.48 About spectral broadening

The fact that there is a frequency *range* that can be absorbed may seem to violate the postulate of quantum mechanics that only the eigenvalues are observable. But actually an atom perturbed by an electromagnetic field is a slightly different system than an unperturbed atom, and will have slightly different energy eigenvalues. Indeed, the frequency range ω_1 is proportional to the strength of the perturbation, and in the limit of the perturbation strength becoming zero, only the exact unperturbed frequency will be absorbed.

For some reason, this spectral line broadening due to the strength of the

transmitted light is not mentioned in the references the author has seen. Presumably it is included in what is called Stark broadening.

The “natural broadening” due to the always present ground state electromagnetic field perturbation is mentioned, but usually ascribed to the energy-time uncertainty $\Delta E \Delta t \geq \frac{1}{2}\hbar$ where ΔE is the uncertainty in energy and Δt some sort of uncertainty in time that in this case is claimed to be the typical life time of the excited state. And of course, a \geq sign is readily changed into an \approx sign; they are both mathematical symbols, not?

Anyway, considered as a dimensional argument rather than a law of physics, it does seem to work; if there was no ground state electromagnetic field perturbing the atom, then Schrödinger’s equation would have the excited state surviving forever; Δt would then be infinite, and the energy values would be the exact unperturbed ones. And transitions like the 21 cm line of astronomy that has a life time of 10 million years do indeed have a very small natural width.

Of course, broadening affects both the absorption spectra (frequencies removed from light that passes through the gas on its way towards us) and the emission spectra (spontaneously emitted radiation, like the “scattered” radiation re-emitted from absorbed light that passes through the gas not originally headed in our direction.)

An important other 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; 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 motion away from you. Line deviations can also be caused by surrounding atoms and other perturbations.

A.49 Derivation of the Einstein B coefficients

The purpose of this note is to derive the Einstein B coefficients that determine the transition probability between the energy states of atoms. It is assumed the atoms are subject to incoherent radiation and frequent elastic collisions with other atoms. It is also again assumed that there are just two atom energy eigenfunctions involved, a lower energy one ψ_L and an higher energy one ψ_H .

It is assumed that the elastic collisions do not change the average energy picture; that they do not affect the average probabilities $|a|^2$ and $|b|^2$ of the eigenfunctions ψ_L and ψ_H . However, they are assumed to leave the wave function of an atom immediately after a collision in some state $a_0\psi_L + b_0\psi_H$ in which a_0 and b_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 $|a|^2$ and $|b|^2$.

As noted in subsection 6.3.1, the Schrödinger equation simplifies if you switch to new variables \bar{a} and \bar{b} . These new variables have the same square magnitudes and initial conditions as a and b . Further, because the Schrödinger equation (6.18) is linear, the solution for the coefficients \bar{a} and \bar{b} can be written as a sum of two contributions, one proportional to the initial value a_0 and the other to b_0 :

$$\bar{a} = a_0 \bar{a}^L + b_0 \bar{a}^H \quad \bar{b} = a_0 \bar{b}^L + b_0 \bar{b}^H$$

Here (\bar{a}^L, \bar{b}^L) is the solution that starts out from the lower energy state $(\bar{a}^L, \bar{b}^L) = (1, 0)$ while (\bar{a}^H, \bar{b}^H) is the solution that starts out from the higher energy state $(\bar{a}^H, \bar{b}^H) = (0, 1)$.

Now consider what happens to the probability of an atom to be in the excited state in the time interval between collisions:

$$|\bar{b}|^2 - |b_0|^2 = (b_0 + a_0 \Delta \bar{b}^L + b_0 \Delta \bar{b}^H)^* (b_0 + a_0 \Delta \bar{b}^L + b_0 \Delta \bar{b}^H) - b_0^* b_0$$

Here $\Delta \bar{b}^L$ indicates the change in \bar{b}^L in the time interval between collisions; in particular $\Delta \bar{b}^L = \bar{b}^L$ since this solution starts from the ground state with $b^L = 0$. Similarly, the change $\Delta \bar{b}^H$ equals $\bar{b}^H - 1$ since this solution starts out from the excited state with $b^H = 1$. Like in section 6.3.8, it will again be assumed that the changes $\Delta \bar{b}^L$ and $\Delta \bar{b}^H$ are small; in view of the Schrödinger equation (6.18), that is true as long as the typical value of the Hamiltonian coefficient \bar{H}_{LH} times the time interval t between the collisions is small. Note also that $\Delta \bar{b}^H$ will be quadratically small, since the corresponding solution starts out from $a^H = 0$, so a^H is an additional small factor in the Schrödinger equation (6.18) for b^H .

Therefore, if the change in probability $|\bar{b}|^2$ 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):

$$|\bar{b}|^2 - |b_0|^2 = 2\Re(b_0^* a_0 \Delta \bar{b}^L) + |a_0|^2 |\Delta \bar{b}^L|^2 + |b_0|^2 2\Re(\Delta \bar{b}^H)$$

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 a_0 and b_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, $2\Re(\Delta \bar{b}^H)$ is the approximate change in $|\bar{b}^H|^2$ in the time interval, and that equals $-|\Delta \bar{a}^H|^2$ because of the normalization condition $|\bar{a}^H|^2 + |\bar{b}^H|^2 = 1$. So the relevant expression for the change in probability becomes

$$|\bar{b}|^2 - |b_0|^2 = |a_0|^2 |\Delta \bar{b}^L|^2 - |b_0|^2 |\Delta \bar{a}^H|^2$$

Summing the changes in the probabilities therefore means summing the changes in the square magnitudes of $\Delta\bar{b}^L$ and $\Delta\bar{a}^H$. According to the above, the Einstein coefficient $B_{L \rightarrow H}$ is the average change $|\Delta\bar{b}^L|^2$ per unit time.

Now for a single electromagnetic wave, subsection 6.3.8 found that the change $\Delta\bar{b}^L$ was given by

$$\Delta\bar{b}^L = \bar{b}^L = \omega_1 e^{i\phi} \frac{e^{-i(\omega-\omega_p)t} - 1}{2(\omega - \omega_p)} \quad \omega_1 = \frac{E_0 \langle \psi_L | ez | \psi_H \rangle}{\hbar} \quad (\text{A.51})$$

and $\Delta\bar{a}^H$ 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 , here $\Delta\bar{b}^L$ will consist of the sum of all their contributions:

$$\Delta\bar{b}^L = \sum_{\vec{k}, p} \Delta\bar{b}_{\vec{k}, p}^L$$

(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

$$|\Delta\bar{b}^L|^2 = \sum_{\vec{k}, p} \sum_{\vec{k}, p} \Delta\bar{b}_{\vec{k}, p}^{L,*} \Delta\bar{b}_{\vec{k}, p}^L = \sum_{\vec{k}, p} |\Delta\bar{b}_{\vec{k}, p}^L|^2$$

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 (A.51) above,

$$|\Delta\bar{b}_{\vec{k}, p}^L|^2 = \frac{1}{4} |\omega_1|^2 t^2 \left(\frac{\sin\left(\frac{1}{2}(\omega - \omega_p)t\right)}{\frac{1}{2}(\omega - \omega_p)t} \right)^2 \quad \omega_1 \equiv \frac{\langle \psi_L | ez | \psi_H \rangle}{\hbar} E_0$$

Now broadband radiation is described in terms of an electromagnetic energy density $\rho(\omega)$; in particular $\rho(\omega) d\omega$ gives the energy per unit volume due to the electromagnetic waves in an infinitesimal frequency range $d\omega$ around a frequency ω . For a single wave, this energy equals $\frac{1}{2}\epsilon_0 E_0^2$, chapter 12.2.3 (12.40). 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 $E_0^2 = 2\rho(\omega) d\omega / \epsilon_0$ and integrate:

$$|\Delta\bar{b}^L|^2 = \frac{|\langle \psi_L | ez | \psi_H \rangle|^2}{2\hbar^2 \epsilon_0} t^2 \int_{\omega=0}^{\infty} \rho(\omega) \left(\frac{\sin\left(\frac{1}{2}(\omega - \omega_p)t\right)}{\frac{1}{2}(\omega - \omega_p)t} \right)^2 d\omega$$

If a change of integration variable is made to $u = \frac{1}{2}(\omega - \omega_p)t$, the integral becomes

$$|\Delta\bar{b}^L|^2 = \frac{|\langle\psi_L|ez|\psi_H\rangle|^2}{\hbar^2\epsilon_0}t \int_{u=-\frac{1}{2}\omega_p t}^{\infty} \rho(\omega_p + 2(u/t)) \left(\frac{\sin u}{u}\right)^2 du$$

Recall that a starting assumption underlying these derivations was that $\omega_p t$ was large. So the lower limit of integration can be approximated as $-\infty$. Also, in the argument of the energy density ρ , the term $2u/t$ represents a negligible change in ω_p and can be ignored. Then $\rho(\omega_p)$ is a constant in the integration and can be taken out. The remaining integral is in table books, [28, 18.36], and the result is

$$|\Delta\bar{b}^L|^2 = \frac{\pi|\langle\psi_L|ez|\psi_H\rangle|^2}{\hbar^2\epsilon_0}\rho(\omega_p)t$$

This must still be averaged over all directions of wave propagation and polarization. That gives:

$$|\Delta\bar{b}^L|^2 = \frac{\pi|\langle\psi_L|e\vec{r}|\psi_H\rangle|^2}{3\hbar^2\epsilon_0}\rho(\omega_p)t$$

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 $B_{L \rightarrow H}$ is the average change per unit time, so the claimed (6.38) results from dividing by the time t between collisions. There is no need to do $B_{H \rightarrow L}$ separately from $\Delta\bar{a}^L$; it follows immediately from subsection 6.3.2 that it is the same.

A.50 Parseval and the Fourier inversion theorem

This note discusses the Parseval's relation and its relation to the Fourier inversion theorem. The analysis is in one dimension but extension to three dimensions is straightforward.

The Fourier inversion theorem is most simply derived from the formulae for Fourier series as discussed in note {A.5}. Rescale the Fourier series to an

arbitrary period, and then take the limit of the period going to infinity to get the Fourier integral expressions. The Fourier series itself becomes an integral in the limit.

The same way, you can verify that the inner product $\langle \Psi_1 | \Psi_2 \rangle$ of any two wave functions is the same as the inner product $\langle \Phi_1 | \Phi_2 \rangle$ of the corresponding momentum space wave functions. This very important result is known as “Parseval’s relation.”

In particular the norm of any wave function is the same as that of the corresponding momentum space wave function. That is especially relevant for quantum mechanics where both norms must be one: the particle must be found *somewhere*, and it must be found with *some* linear momentum.

To be sure, the mentioned derivations of these properties based on converting Fourier series into integrals only work 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. And since the differences between functions are the same for the corresponding momentum space wave functions because of Parseval, the momentum space wave functions converge to the momentum space wave function of the nasty wave function.

To show that rigorously is a messy exercise, to be sure, requiring the abstract Lebesgue variant of the theory of integration. The resulting “official version” of the inversion theorem is called “Plancherel’s theorem.” It applies as long as the nasty functions are square integrable (in the Lebesgue sense).

That is a very suitable version of the inversion theorem for quantum mechanics where the functions are normally square integrable because of the normalization requirement. But it is hardly the last word, 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.51 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 (6.56) for Ψ will be abbreviated as:

$$\Psi = \int_{k_1}^{k_2} f(k) e^{i\varphi t} dk \quad \varphi = k \frac{x}{t} - \omega \quad \varphi' = \frac{x}{t} - v_g \quad \varphi'' = -v'_g$$

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:

$$\Psi = \frac{f(k)}{i\varphi't} e^{i\varphi t} \Big|_{k_1}^{k_2} - \int_{k_1}^{k_2} \left(\frac{f(k)}{i\varphi't} \right)' e^{i\varphi t} dk$$

This is small of order $1/t$ for large times. And if $\bar{\Phi}_0(p)$ 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,

$$\Psi = e^{i\varphi_0 t} \int_{k_1}^{k_2} f(k) e^{i(\varphi - \varphi_0)t} dk$$

It is convenient to write the difference in φ in terms of a new variable \bar{k} :

$$\varphi - \varphi_0 = \frac{1}{2}\varphi_0''\bar{k}^2 \quad \bar{k} \sim k - k_0 \quad \text{for } k \rightarrow k_0$$

By Taylor series expansion it can be seen that \bar{k} is a well behaved monotonous function of k . The integral becomes in terms \bar{k} :

$$\Psi = e^{i\varphi_0 t} \int_{\bar{k}_1}^{\bar{k}_2} g(\bar{k}) e^{i\frac{1}{2}\varphi_0''\bar{k}^2 t} d\bar{k} \quad g(\bar{k}) = f(k) \frac{dk}{d\bar{k}}$$

Now split function g apart as in

$$g(\bar{k}) = g(0) + [g(\bar{k}) - g(0)]$$

The part within brackets produces an integral

$$e^{i\varphi_0 t} \int_{\bar{k}_1}^{\bar{k}_2} \frac{g(\bar{k}) - g(0)}{i\varphi_0'' \bar{k} t} i\varphi_0'' \bar{k} t e^{i\frac{1}{2}\varphi_0'' \bar{k}^2 t} d\bar{k}$$

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

$$\Psi = e^{i\varphi_0 t} f(k_0) \int_{\bar{k}_1}^{\bar{k}_2} e^{i\frac{1}{2}\varphi_0'' \bar{k}^2 t} d\bar{k}$$

Change to a new integration variable

$$u \equiv \sqrt{\frac{|\varphi_0''| t}{2}} \bar{k}$$

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

$$\Psi = e^{i\varphi_0 t} f(k_0) \sqrt{\frac{2}{|\varphi_0''| t}} \int_{u_1}^{u_2} e^{\pm i u^2} du$$

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

$$\Psi = e^{i\varphi_0 t} e^{\pm i\pi/4} f(k_0) \sqrt{\frac{2\pi}{|\varphi_0''| t}}$$

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.

A.52 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.

A.52.1 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 (6.56), simply add the periodic factor $\psi_{p,k}^P(x)$. 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 $d\omega/dk$. Since the energy is $E^P = \hbar\omega$ and the crystal momentum $p_{cm} = \hbar k$, the velocity of a localized electron can be written as

$$v = \frac{dE^P}{dp_{cm}}$$

In the absence of external forces, the electron will keep moving with the same velocity for all time. The large time wave function is

$$\Psi(x, t) \sim \frac{e^{\mp i\pi/4}}{\sqrt{|v'_{g0}|t}} \overline{\Phi}_0(k_0) \psi_{p,k_0}^P(x) e^{i(k_0 x - \omega_0 t)} \quad v_{g0} = \frac{x}{t}$$

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.

A.52.2 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 $V_{ext} = -F_{ext}x$ 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 $\langle T_d \rangle$ of the translation operator 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 $e^{ik_0 d}$. A wave packet must however include a small range of k -values. Then $\langle \mathcal{T}_d \rangle$ will be an *average* of e^{ikd} values over the k -values of the wave packet. Still, if the range of k values is small enough, you can write

$$\langle \mathcal{T}_d \rangle = Ae^{ik_0 d}$$

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^{ikd} 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^{ikd} would be exactly $e^{ik_0 d}$, 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 6.1.7:

$$\frac{d\langle \mathcal{T}_d \rangle}{dt} = \frac{i}{\hbar} \langle [H_0 + V_{\text{ext}}, \mathcal{T}_d] \rangle \quad (\text{A.52})$$

where H_0 is the Hamiltonian for the electron in the crystal, and V_{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,

$$[V_{\text{ext}}, \mathcal{T}_d]\Psi \equiv V_{\text{ext}}\mathcal{T}_d\Psi - \mathcal{T}_dV_{\text{ext}}\Psi$$

Writing this out with the arguments of the functions explicitly shown gives:

$$V_{\text{ext}}(x)\Psi(x+d) - V_{\text{ext}}(x+d)\Psi(x) = (V_{\text{ext}}(x) - V_{\text{ext}}(x+d))\mathcal{T}_d\Psi(x)$$

Now assume that the external force F_{ext} is constant over the extent of the wave packet. In that case the difference in the potentials is just $F_{\text{ext}}d$, and that is a constant that can be taken out of the expectation value of the commutator. So:

$$\frac{d\langle \mathcal{T}_d \rangle}{dt} = \frac{i}{\hbar} F_{\text{ext}} d \langle \mathcal{T}_d \rangle \quad (\text{A.53})$$

The solution to this equation is:

$$\langle \mathcal{T}_d \rangle = \langle \mathcal{T}_d \rangle_0 e^{iF_{\text{ext}}dt/\hbar}$$

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:

$$\frac{dk_0}{dt} = F_{\text{ext}}$$

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.

A.52.3 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 5.20. 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_{ext} . The time-dependent Schrödinger equation is

$$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} - F_{\text{ext}}x\Psi$$

Assume the initial condition to be

$$\Psi_0 = \sum_{k_0} c_{k_0} e^{ik_0 x}$$

in which a subscript 0 indicates the initial time.

The exact solution to this problem is

$$\Psi = \sum_{k_0} c(k_0, t) e^{ik(t)x} \quad \frac{dk}{dt} = F_{\text{ext}}$$

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 $e^{ik(t)x}$ 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 $F_{\text{ext}}t/\hbar$ 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 $e^{iF_{\text{ext}}\ell t/\hbar}$ 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.

A.53 Details of the animations

This note explains how the wave packet animations of sections 6.6 and 6.8 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 subsection 6.6.1. An energy eigenfunction with energy E in free space takes the general form

$$\psi_E = C_f e^{ipx/\hbar} + C_b e^{-ipx/\hbar} \quad p = \sqrt{2mE}$$

where p is the momentum of the particle and C_f and C_b are constants.

To study a single wave packet coming in from the far left, the coefficient C_b has to be set to zero. The reason was worked out in section 6.5: combinations of exponentials of the form $C_b e^{-ipx/\hbar}$ produce wave packets that propagate backwards in x , from right to left. Therefore, a nonzero value for C_b would add an unwanted second wave packet coming in from the far right.

With only the coefficient C_f of the forward moving part left, you may as well scale the eigenfunction so that $C_f = 1$, simplifying it to

$$\psi_E = e^{ipx/\hbar}$$

A typical example is shown in figure A.8. Plus and minus the magnitude of the

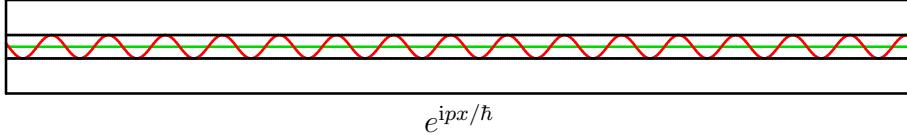


Figure A.8: Example energy eigenfunction for the particle in free space.

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

$$\Psi(x, t) = \int_{\text{all } p} c(p) e^{-iEt/\hbar} \psi_E(x) dp \quad (\text{A.54})$$

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.

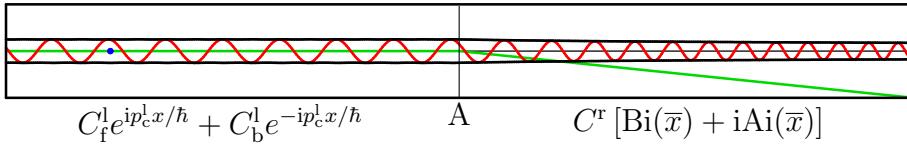


Figure A.9: Example energy eigenfunction for a particle entering a constant accelerating force field.

Next consider the animation of subsection 6.6.2, where the particle accelerates along a downward potential energy ramp starting from point A. A typical energy eigenfunction is shown in figure A.9. Since to the left of point A, the potential energy is still zero, in that region the energy eigenfunction is still of the form

$$\psi_E = C_f^l e^{ip_c^l x/\hbar} + C_b^l e^{-ip_c^l x/\hbar} \text{ for } x < x_A \quad p_c^l = \sqrt{2mE}$$

where p_c^l 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_b^l must be zero to avoid a packet entering from the far right. After all, the $C_b^l e^{-ip_c^l x/\hbar}$ 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 note {A.55}. 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

$$\psi_E = C_B \text{Bi}(\bar{x}) + C_A \text{Ai}(\bar{x}) \quad \bar{x} = \sqrt[3]{\frac{2mV'}{\hbar^2}} \frac{V - E}{V'} \quad V' \equiv \frac{dV}{dx}$$

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 \bar{x} is.

It may be deduced from the approximate analysis of section 6.7 that to prevent a second wave packet coming in from the far right, Ai and Bi must appear together in the combination $\text{Bi} + i\text{Ai}$ as shown in figure A.9. 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_f^l , C_b^l and C^r must still be determined. That goes as follows. For now, assume that C^r has the provisional value $c^r = 1$. Then provisional values c_f^l and c_b^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 $d\psi_E/dx$ at the point A in figure A.9 where they meet:

$$c_f^l e^{ip_c^l x_A/\hbar} + c_b^l e^{-ip_c^l x_A/\hbar} = c^r [\text{Bi}(\bar{x}_A) + i\text{Ai}(\bar{x}_A)]$$

$$c_f^l \frac{ip_c^l}{\hbar} e^{ip_c^l x_A/\hbar} - c_b^l \frac{ip_c^l}{\hbar} e^{-ip_c^l x_A/\hbar} = c^r [\text{Bi}'(\bar{x}_A) + i\text{Ai}'(\bar{x}_A)] \frac{d\bar{x}}{dx}$$

That is equivalent to two equations for the two constants c_f^l and c_b^l , since everything else can be evaluated, using the mentioned software. So c_f^l and c_b^l can be found from solving these two equations.

As the final step, it is desirable to normalize the eigenfunction ψ_E so that $C_f^l = 1$. To do so, the entire provisional eigenfunction can be divided by c_f^l , giving $C_b^l = c_b^l/c_f^l$ and $C^r = c^r/c_f^l$. The energy eigenfunction has now been found. And since $C_f^l = 1$, the $e^{ip_c^l x/\hbar}$ 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.54) with p replaced by p_c^l , the $e^{ip_c^l x/\hbar}$ terms produce the exact same wave packet coming in from the far left as in the free space case.

For larger times, the $C_b^l e^{-ip_c^l x/\hbar}$ terms produce a “reflected” wave packet that returns toward the far left. Note that $e^{-ip_c^l x/\hbar}$ is the complex conjugate of $e^{ip_c^l x/\hbar}$, 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_b^l turns out to be very small in this case, so there is little reflection.

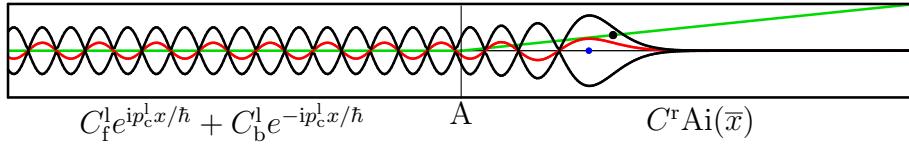


Figure A.10: Example energy eigenfunction for a particle entering a constant decelerating force field.

Next consider the animation of subsection 6.6.3, where the particle is turned back by an upward potential energy ramp. A typical energy eigenfunction for this case is shown in figure A.10. Unlike in the previous example, where the argument \bar{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 \bar{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 \bar{x} , as seen in figure A.10.

The further determination of the energy eigenfunctions proceeds along the same lines as in the previous example: give C^r a provisional value $c^r = 1$, then compute c_f^l and c_b^l from the requirements that the left and right regions produce the same values for ψ and $d\psi/dx$ at the point A where they meet. Finally divide the eigenfunction by c_f^l . The big difference is that now C_b^l is no longer small; C_b^l turns out to be of unit magnitude just like C_f^l . It means that the incoming wave packet is reflected back completely.

For the harmonic oscillator of subsection 6.6.4, the analysis is somewhat different. In particular, chapter 2.6.2 showed that the energy levels of the one-dimensional harmonic oscillator are discrete,

$$E_n = \frac{2n+1}{2}\hbar\omega \text{ for } n = 0, 1, 2, \dots$$

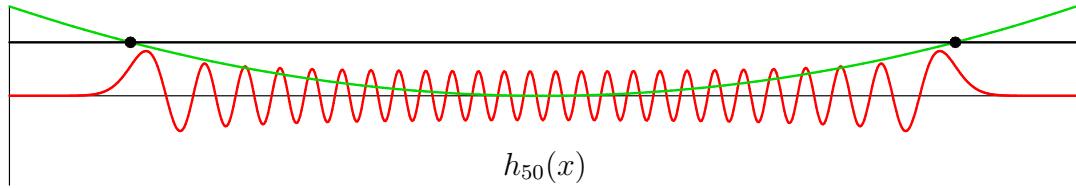


Figure A.11: Example energy eigenfunction for the harmonic oscillator.

so that unlike the motions just discussed, the solution of the Schrödinger equation is a sum, rather than the integral (A.54),

$$\Psi(x, t) = \sum_{n=0}^{\infty} c_n e^{-iE_n t/\hbar} h_n(x)$$

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.11, 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 (1.5). 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.

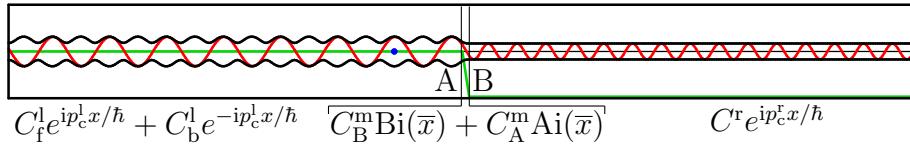


Figure A.12: Example energy eigenfunction for a particle encountering a brief accelerating force.

For the particle of section 6.8.1 that encounters a brief accelerating force, an example eigenfunction looks like figure A.12. 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 $e^{-ip_c^r x/\hbar}$ 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 $c^r = 1$ and from that, compute provisional values c_A^m and c_B^m by demanding that the Airy functions give the same values for ψ and $d\psi/dx$ as the far-right exponential at point B, where they meet. Next compute provisional values c_f^l and c_b^l by demanding that the far-left exponentials give the same values for ψ and $d\psi/dx$ as the Airy functions at point A, where they meet. Finally, divide all the constants by c_f^l to make $C_f^l = 1$.

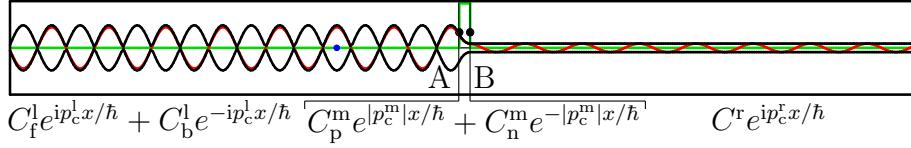


Figure A.13: Example energy eigenfunction for a particle tunneling through a barrier.

For the tunneling particle of section 6.8.2, an example eigenfunction is as shown in figure A.13. 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 $p_c^m = \sqrt{2m(E - V_m)}$ an imaginary number. Therefore the arguments of the exponentials become real when written in terms of the absolute value of the momentum $|p_c^m| = \sqrt{2m(V_m - E)}$. The rest of the analysis is similar to that of the previous example.

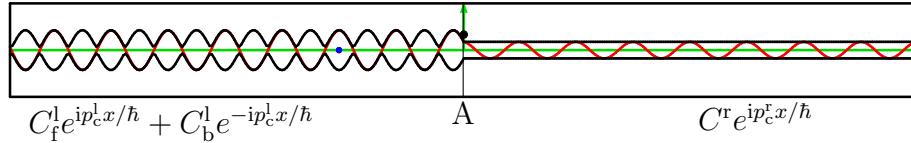


Figure A.14: Example energy eigenfunction for tunneling through a delta function barrier.

For the particle tunneling through the delta function potential in section 6.8.2, an example energy eigenfunction is shown in figure A.14. The potential energy in this case is $V = \nu\delta(x - x_A)$, where $\delta(x - x_A)$ 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 $\sqrt{2\hbar^2 E_{\text{nom}}/m}$ with E_{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.14 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 $d\psi/dx$ 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:

$$-\frac{\hbar^2}{2m} \int_{x=x_A-\varepsilon}^{x_A+\varepsilon} \frac{d^2\psi}{dx^2} dx + \nu \int_{x=x_A-\varepsilon}^{x_A+\varepsilon} \delta(x - x_A) \psi dx = \int_{x=x_A-\varepsilon}^{x_A+\varepsilon} E\psi dx$$

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

$$-\frac{\hbar^2}{2m} \frac{d\psi}{dx} \Big|_{x_A-\varepsilon}^{x_A+\varepsilon} + \nu\psi(x_A) = 0$$

For vanishingly small ε , $d\psi/dx$ at $x_A + \varepsilon$ becomes what the right hand part of the eigenfunction gives for $d\psi/dx$ at x_A , while $d\psi/dx$ at $x_A - \varepsilon$ becomes what the left hand part gives for it. As seen from the above equation, the difference is not zero, but $2m\nu\psi(x_A)/\hbar^2$.

So the correct equations for the provisional constants are in this case

$$\begin{aligned} c_f^l e^{ip_c^l x_A/\hbar} + c_b^l e^{-ip_c^l x_A/\hbar} &= c^r e^{ip_c^r x_A/\hbar} \\ \frac{ip_c^l}{\hbar} c_f^l e^{ip_c^l x_A/\hbar} - \frac{ip_c^l}{\hbar} c_b^l e^{-ip_c^l x_A/\hbar} &= \frac{ip_c^r}{\hbar} c^r e^{ip_c^r x_A/\hbar} - \frac{2m\nu}{\hbar^2} c^r e^{ip_c^r x_A/\hbar} \end{aligned}$$

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

$$-\frac{d^2\psi}{dx^2} + V\psi = E\psi$$

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.54) 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

$$c(p) = 1 \quad \text{if } (1 - r)p_{\text{nom}} < p < (1 + r)p_{\text{nom}} \quad c(p) = 0 \quad \text{otherwise}$$

where r is the relative deviation from the nominal momentum below which $c(p)$ is nonzero. However, it is known from Fourier analysis that the locations where $c(p)$ jumps from one to zero lead to lengthy wave packets when viewed in physical space. {A.51}. Functions $c(p)$ that do lead to nice compact wave packets are known to be of the form

$$c(p) = \exp\left(-\frac{(p - p_{\text{nom}})^2}{r^2 p_{\text{nom}}^2}\right)$$

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

$$c(p) = \exp\left(-\frac{(p - p_{\text{nom}})^2}{r^2[p_{\text{nom}}^2 - (p - p_{\text{nom}})^2]}\right) \quad \text{for } 0 < p < 2p_{\text{nom}} \quad c(p) = 0 \quad \text{otherwise}$$

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 \bar{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 note {A.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.54 Derivation of the WKB approximation

The purpose in this note is to derive an approximate solution to the Hamiltonian eigenvalue problem

$$\frac{d^2\psi}{dx^2} = -\frac{p_c^2}{\hbar^2}\psi$$

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,

$$\psi = e^{i\tilde{\theta}}$$

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 $\tilde{\theta}$

$$\frac{d\psi}{dx} = e^{i\tilde{\theta}} i \frac{d\tilde{\theta}}{dx} \quad \frac{d^2\psi}{dx^2} = -e^{i\tilde{\theta}} \left(\frac{d\tilde{\theta}}{dx} \right)^2 + e^{i\tilde{\theta}} i \frac{d^2\tilde{\theta}}{dx^2}$$

Then plugging ψ and its second derivative above into the Hamiltonian eigenvalue problem and cleaning up gives:

$$\left(\frac{d\tilde{\theta}}{dx}\right)^2 = \frac{p_c^2}{\hbar^2} + i\frac{d^2\tilde{\theta}}{dx^2} \quad (\text{A.55})$$

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 :

$$\tilde{\theta} = \frac{1}{\hbar} \left(f_0 + \hbar f_1 + \frac{1}{2}\hbar^2 f_2 + \dots \right)$$

You can think of this as writing $\hbar\theta$ as a Taylor series in \hbar . The coefficients f_0, f_1, f_2, \dots 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

$$\frac{1}{\hbar^2} f_0'^2 + \frac{1}{\hbar} 2f_0' f_1' + \dots = \frac{1}{\hbar^2} p_c^2 + \frac{1}{\hbar} i f_0'' + \dots$$

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, $f_0'^2 = p_c^2$, so there are two possible solutions

$$f_0' = \pm p_c$$

For the coefficients of $1/\hbar$ to be equal, $2f_0' f_1' = i f_0''$, or plugging in the solution for f_0' ,

$$f_1' = i \frac{p_c'}{2p_c}$$

It follows that the x -derivative of $\tilde{\theta}$ is given by

$$\tilde{\theta}' = \frac{1}{\hbar} \left(\pm p_c + \hbar i \frac{p_c'}{2p_c} + \dots \right)$$

and integrating gives $\tilde{\theta}$ as

$$\tilde{\theta} = \pm \frac{1}{\hbar} \int p_c dx + i \frac{1}{2} \ln p_c + \tilde{C} \dots$$

where \tilde{C} is an integration constant. Finally, $e^{i\tilde{\theta}}$ now gives the two terms in the WKB solution, one for each possible sign, with $e^{i\tilde{C}}$ equal to the constant C_f or C_b .

A.55 WKB solution near the turning points

Both the classical and tunneling WKB approximations of section 6.7 fail near so-called “turning points” where the classical kinetic energy $E - V$ becomes zero. This note explains how the problem can be fixed.

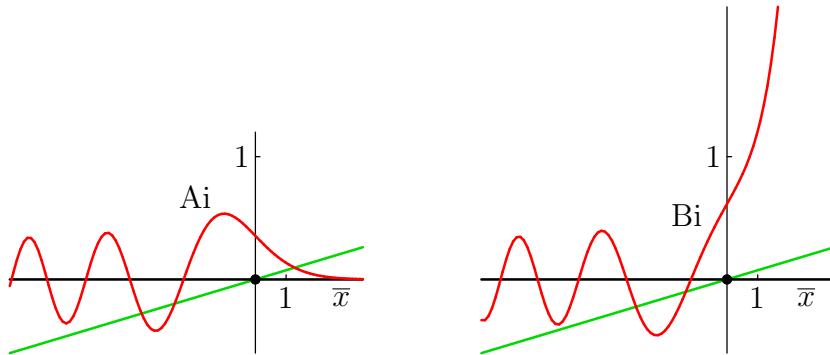


Figure A.15: The Airy Ai and Bi functions that solve the Hamiltonian eigenvalue problem for a linearly varying potential energy. Bi very quickly becomes too large to plot for positive values of its argument.

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.15. The general solution near a turning point is:

$$\psi = C_A \text{Ai}(\bar{x}) + C_B \text{Bi}(\bar{x}) \quad \bar{x} = \sqrt[3]{\frac{2mV'}{\hbar^2}} \frac{V - E}{V'} \quad V' \equiv \frac{dV}{dx}$$

Note that $(V - E)/V'$ is the x -position measured from the point where $V = E$, so that \bar{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 $|\bar{x}|$ 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 $\hbar^{1/3}$, $|\bar{x}|$ will be proportional to $1/\hbar^{1/3}$, and that is large. Such regions of large, but not too

large, $|\bar{x}|$ 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.16 and A.17. 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 \bar{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_p and C_n of the tunneling approximation to the coefficients of C_A and C_B of the Airy functions. At the other side, it relates the coefficients C_f and C_b (or C_c and C_s) of the classical WKB approximation to C_A and C_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.16 and A.17 are called the “connection formulae.”

You may have noted the appearance of an additional constant c in figures A.16 and A.17. This nasty constant is defined as

$$c = \frac{\sqrt{\pi}}{(2m|V'|\hbar)^{1/6}} \quad (\text{A.56})$$

and shows up uninvited when you approximate the Airy function solution for large $|\bar{x}|$. By cleverly absorbing it in a redefinition of the constants C_A and C_B , figures A.16 and A.17 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.16 apply. In the tunneling region towards the right, the term $C_p e^\gamma$ better be zero, because it blows up at large x , and that would put the particle at infinity for sure. So the constant C_p will have to be zero. Now the matching at the right side equates C_p to $C_B e^{-\gamma t}$ so C_B will have to be zero. That means that the solution in the vicinity of the turning 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 section 6.7 that the angular range of the classical WKB solution should be shortened by $\pi/4$ for each end

$$\left. \begin{aligned} & \frac{1}{\sqrt{p_c}} [C_f e^{i\theta} + C_b e^{-i\theta}] \\ & \frac{1}{\sqrt{p_c}} [C_c \cos \theta + C_s \sin \theta] \end{aligned} \right\} \quad C_B c \text{Bi}(\bar{x}) + C_A c \text{Ai}(\bar{x}) \quad \frac{1}{\sqrt{|p_c|}} [C_p e^\gamma + C_n e^{-\gamma}]$$

equate \Updownarrow

$$\frac{1}{\sqrt{p_c}} \left[C_B \cos\left(\theta - \theta_t - \frac{\pi}{4}\right) - C_A \sin\left(\theta - \theta_t - \frac{\pi}{4}\right) \right] \quad \frac{1}{\sqrt{|p_c|}} \left[C_B e^{\gamma - \gamma_t} + \frac{1}{2} C_A e^{\gamma_t - \gamma} \right]$$

\Updownarrow equate

Figure A.16: Connection formulae for a turning point from classical to tunneling.

$$\frac{1}{\sqrt{|p_c|}} [C_p e^\gamma + C_n e^{-\gamma}] \quad C_B c \text{Bi}(\bar{x}) + C_A c \text{Ai}(\bar{x})$$

equate \Updownarrow

$$\frac{1}{\sqrt{|p_c|}} \left[C_B e^{\gamma_t - \gamma} + \frac{1}{2} C_A e^{\gamma - \gamma_t} \right] \quad \frac{1}{\sqrt{p_c}} \left[C_B \cos\left(\theta - \theta_t + \frac{\pi}{4}\right) + C_A \sin\left(\theta - \theta_t + \frac{\pi}{4}\right) \right]$$

\Updownarrow equate

Figure A.17: Connection formulae for a turning point from tunneling to classical.

at which the particle is trapped by a gradually increasing potential instead of an impenetrable wall.

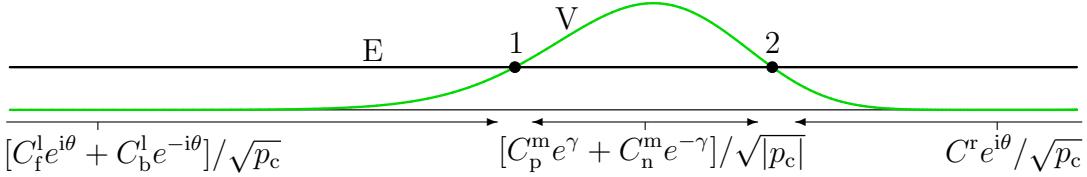


Figure A.18: WKB approximation of tunneling.

As another example, consider tunneling as discussed in sections 6.8 and 6.9. Figure A.18 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 $C^r e^{i\theta}$ with a forward wave speed. To simplify the analysis, the constant C^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.17 along with the Euler formula (1.5) show that the coefficients of the Airy functions at turning point 2 are $C_B = 1$ and $C_A = i$. Next, the integration constant in γ can be taken such that $\gamma = 0$ at turning point 2; then the connection formulae of figure A.17 imply that $C_p^m = \frac{1}{2}i$ and $C_n^m = 1$.

Next consider the connection formulae for turning point 1 in figure A.16. 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.16 that near turning point 1, $C_A = 2e^{\gamma_{12}}$ which is large while C_B is small and will be ignored. And that then implies, using the Euler formula to convert Ai's sine into exponentials, that $|C_f^l| = e^{\gamma_{12}}$. As discussed in section 6.9, the transmission coefficient is given by

$$T = \frac{p_c^r |C^r/\sqrt{p_c^r}|^2}{p_c^l |C_f^l/\sqrt{p_c^l}|^2}$$

and plugging in $C^r = 1$ and $|C_f^l| = e^{\gamma_{12}}$, the transmission coefficient is found to be $e^{-2\gamma_{12}}$.

A.56 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.

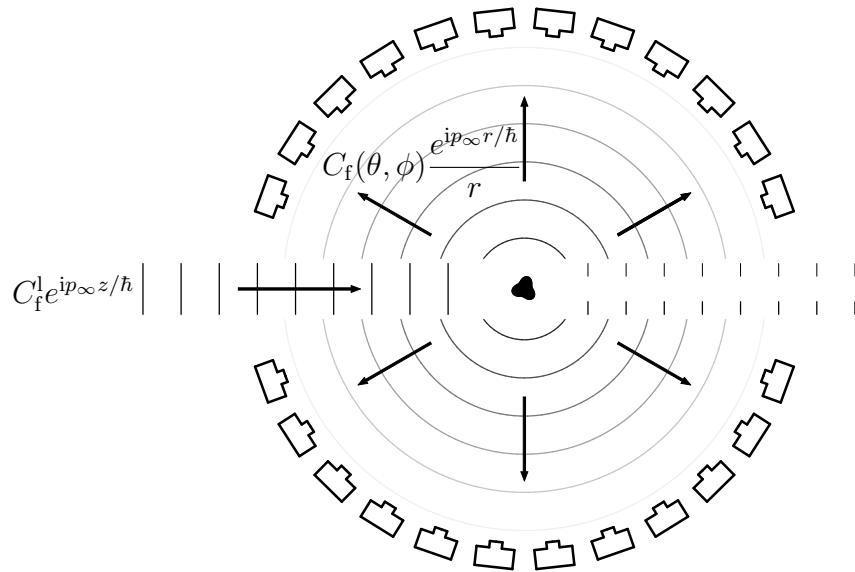


Figure A.19: Scattering of a beam off a target.

The basic idea is as sketched in figure A.19. A beam of particles is send 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 decays fairly quickly to zero away from the target.

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 rather than x as in the one-dimensional examples.

In the relevant (nominal) energy eigenfunction, the incoming beam can still be represented as a one-dimensional wave. However, unlike for the one-dimensional scattering of figure 6.19, 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 nominal energy eigenfunction is

$$\psi_E \sim C_f^l e^{ip_\infty z/\hbar} + C_f(\theta, \phi) \frac{e^{ip_\infty r/\hbar}}{r} \quad \text{for } r \rightarrow \infty \quad (\text{A.57})$$

where C_f is called the scattering amplitude.” The first term in the far field behavior allows the incoming wave packets to be described and the same packets going out again unperturbed. If some joker removes the target, that is all there is. The second term describes outgoing scattered waves. This term must be proportional to $e^{ip_\infty r/\hbar}$, without a $e^{-ip_\infty r/\hbar}$ term, since no wave packets should come in from infinity except those in the incoming beam. The magnitude of the second term decreases with r because the probability of finding a particle in a given detection area should decrease with distance. Indeed, the total detection area increases with its radius as $4\pi r^2$ and the total number of particles to detect per unit time is the same wherever the detectors are, so the probability of finding a particle per unit area should decrease as $1/r^2$. 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.

It follows that the number of particles in a small detection area dA is proportional to its angular extent, or “solid angle” $d\Omega = dA/r^2$. In spherical coordinates

$$d\Omega = \sin \theta \, d\theta \, d\phi$$

For a continuous beam of particles being directed at the target, the number of scattered particles per unit solid angle and per unit time will be proportional to the square magnitude of the scattered wave function:

$$\frac{d\dot{I}}{d\Omega} \propto |C_f(\theta, \phi)|^2$$

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 proportional to the square magnitude of the incoming one-dimensional wave function,

$$\frac{d\dot{I}}{dA_b} \propto |C_f^l|^2$$

Physicists like to take the ratio of the two in order that the rate at which the particles are sent in is scaled away, so they define

$$D(\theta, \phi) \equiv \frac{dA_b}{d\Omega} = \frac{|C_f(\theta, \phi)|^2}{|C_f^l|^2} \quad (\text{A.58})$$

It is the infinitesimal area dA_b of the incoming beam that is scattered into the infinitesimal solid angle $d\Omega$. So it is a scattered particle density expressed in suitable terms.

However, “scattered particle density” would be understandable, so physicists call it the “differential cross-section.” This is a particularly well chosen name, because it is not a differential, but a differential quotient. It will greatly confuse mathematically literate people. And “cross-section” is sufficiently vague that it can mean anything; it does not at all give the secret away that the thing is really a measure for the number of scattered particles.

The total area of the incoming beam that is scattered can be found by integrating over all deflection angles:

$$\sigma \equiv A_{b,\text{total}} = \int_{\text{all}} \frac{dA_b}{d\Omega} d\Omega = \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} D(\theta, \phi) \sin \theta d\theta d\phi \quad (\text{A.59})$$

Note that the integration should exclude the particles that pass through undeflected in figure A.19; in other words it should exclude $\theta = 0$. Physicists call σ the “total cross-section.” That is quite below their normal standards, 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 cross-section, and D instead of something like $dA_b/d\Omega$ or A'_b , or even σ' , for the differential cross-section, they 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 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 that the particles come in; it has no direct relation whatsoever to what sort of impact (if any) these particles end up experiencing.

A.56.1 Partial wave analysis

Jim Napolitano from RPI and Cornell writes “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 3.2. 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 nonzero 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 :

$$\psi_{Elm}(r, \theta, \phi) = R_{El}(r)Y_l^m(\theta, \phi) \quad (\text{A.60})$$

The reason is that any spherically symmetric potential commutes with both \hat{L}^2 and \hat{L}_z , so the energy eigenfunctions can be taken to be also eigenfunctions of \hat{L}^2 and \hat{L}_z . However, the functions R_{El} will not be the hydrogen ones R_{nl} .

In terms of the classical momentum

$$p_c \equiv \sqrt{2m(E - V)}$$

the Hamiltonian eigenvalue problem is

$$-\nabla^2\psi_{Elm} = \frac{p_c^2}{\hbar^2}\psi_{Elm}$$

Following chapter 3.2.2, for a spherically symmetric potential, this may be reduced to an equation for the R_{El} :

$$\frac{d}{dr} \left(r^2 \frac{dR_{El}}{dr} \right) + \left[\frac{p_c^2}{\hbar^2} r^2 - l(l+1) \right] R_{El} = 0$$

Note that the classical momentum p_c is constant wherever the potential energy is constant. That includes the detection region far from the scattering object, where the potential is zero and $p_c = p_\infty$. In such regions, the solution for R_{El} can be found in advanced mathematical handbooks, like for example [1]. Depending on whatever is easiest, the solution can be written in two ways. The first is as

$$R_{El} = c_s j_l(p_c r / \hbar) + c_n n_l(p_c r / \hbar)$$

where 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 be indicated by y_l or η_l . The other way to write the solution is as

$$R_{El} = c_f h_l^{(1)}(p_c r / \hbar) + c_b h_l^{(2)}(p_c r / \hbar)$$

where $h_l^{(1)}$ and $h_l^{(2)}$ are called the “spherical Hankel functions.”

The spherical Hankel functions can be found in advanced table books as

$$h_l^{(1)}(x) = -i(-x)^l \left(\frac{1}{x} \frac{d}{dx} \right)^l \frac{e^{ix}}{x} = j_l(x) + i n_l(x) \quad (\text{A.61})$$

$$h_l^{(2)}(x) = i(-x)^l \left(\frac{1}{x} \frac{d}{dx} \right)^l \frac{e^{-ix}}{x} = j_l(x) - i n_l(x) \quad (\text{A.62})$$

These are convenient for large r since the Hankel functions of the first kind represent the outgoing waves while those of the second kind are the incoming waves. Indeed for large x ,

$$\boxed{h_l^{(1)}(x) \sim (-i)^{l+1} \frac{e^{ix}}{x} \quad h_l^{(2)}(x) \sim i^{l+1} \frac{e^{-ix}}{x}} \quad (\text{A.63})$$

The spherical Bessel functions are

$$\boxed{j_l(x) = (-x)^l \left(\frac{1}{x} \frac{d}{dx} \right)^l \frac{\sin x}{x} = \frac{h_l^{(1)} + h_l^{(2)}}{2}} \quad (\text{A.64})$$

$$\boxed{n_l(x) = -(-x)^l \left(\frac{1}{x} \frac{d}{dx} \right)^l \frac{\cos x}{x} = \frac{h_l^{(1)} - h_l^{(2)}}{2i}} \quad (\text{A.65})$$

These are often convenient in a region of constant potential that includes the origin, because the Bessel function of the first kind j_l gives the solution that is finite at the origin. (Note that the Taylor series of $\sin x$ divided by x is a power series in x^2 , and that $x dx = \frac{1}{2} d x^2$.) Also, they are real for real x . However, in a region where the scattering potential is larger than the energy of the particles, the argument x of the Bessel or Hankel functions will be imaginary.

(In case you demand a derivation of the spherical Bessel and Hankel functions, well, OK. It is almost comically trivial compared to similar problems in quantum mechanics. Start with a change of dependent variable from f_l to $x^l g_l$ in the normalized ordinary differential equation to solve:

$$\frac{d}{dx} \left(x^2 \frac{df_l}{dx} \right) + [x^2 - l(l+1)] f_l = 0 \quad f_l \equiv x^l g_l \quad x \frac{d^2 g_l}{dx^2} + 2(l+1) \frac{dg_l}{dx} + x g_l = 0$$

Check, by simply plugging it in, that e^{ix}/x is a solution for $l = 0$. Now make a change in independent variable from x to $\xi = \frac{1}{2}x^2$ to give

$$2\xi \frac{d^2 g_l}{d\xi^2} + 2(l+1) \frac{dg_l}{d\xi} + g_l = 0$$

Note that the equation for $l = 1$ is obtained by differentiating the one for $l = 0$. 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.)

In case you are so inclined, it might be fun to create three-dimensional scattering wave packet animations. As long as you assume that the potential $V(r)$ is piecewise constant, for each piece the solution is given by spherical Bessel functions as above. You should then be able to tie these solutions together wherever different pieces meet much like in note {A.53}. But there is now an additional complication.

The problem is that the wave function that describes the incoming particle beam is a purely Cartesian expression, and the current analysis is in spherical coordinates. The Cartesian one-dimensional wave $e^{ip_\infty z/\hbar}$ will need to be converted to spherical coordinates to do the analysis.

Note that the one-dimensional wave is a solution to the Hamiltonian eigenvalue problem with zero potential. As a solution of that problem, it must be possible to write it as spherical harmonics times spherical Bessel functions, as discussed above. More specifically, the wave can be written as spherical harmonics Y_l^0 times the Bessel functions of the first kind j_l : there are no spherical harmonics for $m \neq 0$ since $z = r \cos \theta$ does not depend on ϕ , and there are no Bessel functions of the second kind because the one-dimensional wave is finite at the origin. Now the interaction of a one-dimensional wave with a three-dimensional object is not at all unique to quantum mechanics, of course, and Rayleigh worked out the correct multiples of these functions to use a very long time ago:

$$e^{ip_\infty z/\hbar} = \sum_{l=0}^{\infty} c_{w,l} j_l(p_\infty r/\hbar) Y_l^0(\theta) \quad c_{w,l} = i^l \sqrt{4\pi(2l+1)} \quad (\text{A.66})$$

This is where the term “partial waves” comes in. In spherical coordinates, the single Cartesian wave $e^{ip_\infty z/\hbar}$ falls apart in an infinite series of partial waves. The scattering problem for each needs to be solved separately and the results summed back together to get the total solution. It is the price to pay for training a Cartesian particle beam on a potential that needs to be solved in spherical coordinates.

(To derive the Rayleigh formula, set $x = p_\infty r/\hbar$ to make the one-dimensional wave $e^{ix \cos \theta}$. Expand in a Taylor series around the origin. The generic term $(ix \cos \theta)^l / l!$ must match $c_{w,l}$ times the term with the lowest power of x in $j_l(x)$ times the term with the highest power of $\cos \theta$ in $Y_l^0(\theta)$. Terms $l \neq l$ are not involved since they do not have a low enough power of x or a high enough power of $\cos \theta$. Deduce or look up the coefficients of the lowest power of x in $j_l(x)$ and highest power of $\cos \theta$ in $Y_l^0(\theta)$ to get $c_{w,l}$ as claimed.)

The solution procedure is now as follows: given an energy E , for every l , (up to some sufficiently large value), find the energy eigenfunction of the form

$\psi_{El} = R_{El}Y_l^0$ that for large r behaves like

$$\psi_{El} \sim [c_{w,l}j_l(p_\infty r/\hbar) + c_{f,l}h_l^{(1)}(p_\infty r/\hbar)] Y_l^0(\theta) \quad \text{for } r \rightarrow \infty \quad p_\infty = \sqrt{2mE} \quad (\text{A.67})$$

Note that at large r , the deviations from the one-dimensional wave must be outgoing reflected waves, so they must be Hankel functions of the first kind. Also, for simplicity the wave function has been scaled to make C_f^l one. As long as V is piecewise constant, it should be possible to solve for ψ_{El} using Bessel or Hankel functions as above and so find the $c_{f,l}$. Sum all these “partial waves” together to find the energy eigenfunction ψ_E for an incoming wave of that energy. (Or better, just sum the Hankel function terms together and add $e^{ip_\infty z/\hbar}$.) Sum such eigenfunctions together in a narrow range of energies to get a one-dimensional incoming wave packet being scattered.

In terms of the asymptotic behavior above, using (A.63), the differential cross section is

$$D(\theta) = \frac{\hbar^2}{p_\infty^2} \sum_{l=0}^{\infty} \sum_{l=0}^{\infty} i^{l-l} c_{f,l}^* c_{f,l} Y_l^0(\theta) Y_l^0(\theta) \quad (\text{A.68})$$

The Bessel functions form the incoming wave and do not contribute. For the total cross-section, note that the spherical harmonics are orthonormal, so

$$\sigma = \frac{\hbar^2}{p_\infty^2} \sum_{l=0}^{\infty} |c_{f,l}|^2$$

The magnitudes of the coefficients $c_{f,l}$ is not arbitrary, but constrained by conservation of probability. That allows these complex numbers to be written in terms of real phase shifts. But if you need that sort of detail, you will need to consult a book for physics students.

A.56.2 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 exactly zero. In that case, the Hamiltonian eigenvalue problem takes the form

$$\left[\nabla^2 + \frac{p_\infty^2}{\hbar^2} \right] \psi_E = 0 \quad (\text{A.69})$$

This equation is called the “Helmholtz equation.” The appropriate solution is here the unperturbed one-dimensional wave

$$\psi_E = e^{ip_\infty z/\hbar}$$

Now consider the case that the potential is not zero, but small. In that case the Hamiltonian eigenvalue problem becomes

$$\left[\nabla^2 + \frac{p_\infty^2}{\hbar^2} \right] \psi_E = f \quad f = \frac{2mV}{\hbar^2} \psi_E \quad (\text{A.70})$$

The trick is now that as long as the potential is very small, ψ_E will be almost the same as the one-dimensional wave. Substituting that into the expression for f , the approximate right hand side in the Helmholtz equation becomes a *known* function. The inhomogeneous Helmholtz equation may now be solved for ψ_E much like the Poisson equation was solved in chapter 10.5.4.

In particular, using symbol k for the constant p_∞/\hbar , the general solution to the Helmholtz equation can be written as

$$(\nabla^2 + k^2) \psi = f \implies \psi = \psi_h + \int_{\text{all } \vec{r}} G(\vec{r} - \vec{r}') f(\vec{r}') d^3 \vec{r}' \quad G(\vec{r}) = -\frac{e^{ik|\vec{r}|}}{4\pi|\vec{r}|}$$

(A.71)

where ψ_h describes the effects of any waves that come in from infinity, and can be any solution of the homogeneous Helmholtz equation. To see why this is the solution of the Helmholtz equation, first consider the solution G for the special case that f is a delta function at the origin:

$$(\nabla^2 + k^2) G = \delta^3(\vec{r})$$

The solution G to this problem is called the “Green’s function of the Helmholtz equation. Since the delta function in the right hands side is zero everywhere except at the origin, everywhere except at the origin G is a solution of the homogeneous Helmholtz equation. According to the previous section, that means it must be Bessel or Hankel functions times spherical harmonics. The solution of interest is the one where no waves are propagating in from infinity, because ψ_h takes care of that, so Hankel functions of the first kind only. Further, since there is no angular preference in this problem at all, the solution must be spherically symmetric; that means the spherical harmonic must be Y_0^0 and the Hankel function is then $h_0^{(1)}$. Also, the appropriate normalization factor multiplying this solution is essentially the same as the one for the Green’s function of the Laplace equation: it is the Laplacian ∇^2 of G that produces the delta function; the integral of $k^2 G$ is vanishingly small over the immediate vicinity of the origin. That gives the Green’s function as stated above. Next, to solve the Helmholtz equation for an *arbitrary* right hand side f , just think of that right hand side as made up of infinitely many “spikes” $f(\vec{r}') d\vec{r}'$. Each of these spikes produces a solution given by the Green’s function shifted to location \vec{r}' and scaled. That gives the general solution as stated.

If the solution of the Helmholtz equation is applied to the Schrödinger equation in the form (A.70), it produces the so-called “integral form of the Schrödinger equation”,

$$\psi_E(\vec{r}) = \psi_{E,0}(\vec{r}) - \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}} \frac{e^{ip_\infty |\vec{r} - \underline{\vec{r}}|/\hbar}}{|\vec{r} - \underline{\vec{r}}|} V(\underline{\vec{r}}) \psi_E(\underline{\vec{r}}) d^3 \underline{\vec{r}} \quad p_\infty = \sqrt{2mE} \quad (\text{A.72})$$

where $\psi_{E,0}$ is any free-space wave function of energy E . Whereas the normal Schrödinger equation is called a partial differential equation, this is called an integral equation, because the unknown wave function ψ_E appears in an integral rather than in partial derivatives.

In the Born approximation, $\psi_{E,0}$ is the incoming one-dimensional wave and so is the approximation for ψ_E inside the integral, so

$$\psi_E(\vec{r}) \approx e^{ip_\infty z/\hbar} - \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}} \frac{e^{ip_\infty |\vec{r} - \underline{\vec{r}}|/\hbar}}{|\vec{r} - \underline{\vec{r}}|} V(\underline{\vec{r}}) e^{ip_\infty z/\hbar} d^3 \underline{\vec{r}}$$

This can be cleaned up a bit by noting that the interest is really only in $\psi_E(\vec{r})$ at large distances from the scattering region. In that case $\underline{\vec{r}}$ can be ignored compared to \vec{r} in the denominator, while in the argument of the exponential

$$|\vec{r} - \underline{\vec{r}}| \sim r - \frac{\vec{r} \cdot \underline{\vec{r}}}{r}$$

Also, the *vector* momenta of the incoming and scattered waves may be defined as, respectively,

$$\vec{p}_\infty = p_\infty \hat{k} \quad \vec{p}'_\infty = p_\infty \frac{\vec{r}}{r} \quad p_\infty = \sqrt{2mE}$$

(A.73)

Then

$$\psi_E(\vec{r}) \sim e^{ip_\infty z/\hbar} - \frac{e^{ip_\infty r/\hbar}}{r} \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}} e^{i(\vec{p}_\infty - \vec{p}'_\infty) \cdot \underline{\vec{r}}/\hbar} V(\underline{\vec{r}}) d^3 \underline{\vec{r}}$$

(A.74)

The differential cross section is therefore

$$D(\theta, \phi) \approx \left| \frac{m}{2\pi\hbar^2} \int_{\text{all } \vec{r}} e^{i(\vec{p}_\infty - \vec{p}'_\infty) \cdot \underline{\vec{r}}/\hbar} V(\underline{\vec{r}}) d^3 \underline{\vec{r}} \right|^2 \quad (\text{A.75})$$

Note that \vec{p}'_∞ is a function of ϕ and θ because of its definition above. In the further approximation that the energy of the incoming beam is small, the exponential can be approximated by one.

A.56.3 The Born series

Following Griffiths [17], this section takes a more philosophical view of the Born approximation.

The basic idea of the Born approximation can very schematically be represented as

$$\psi_E \approx \psi_{E,0} + \int gV\psi_{E,0}$$

where g represents the Green's function, absorbing the various constants, and $d^3\vec{r}$ was left away for brevity. The error in this expression is because ψ_E in the integral has been replaced by the unperturbed incoming wave $\psi_{E,0}$. To reduce the error, you could plug the above *two-term* approximation of the wave function into the Green's function integral instead, to give

$$\psi_E \approx \psi_{E,0} + \int gV\psi_{E,0} + \iint gVgV\psi_{E,0}$$

Then you could go one better still by plugging the so obtained three-term approximation into the integral, and so on:

$$\psi_E \approx \psi_{E,0} + \int gV\psi_{E,0} + \iint gVgV\psi_{E,0} + \iiint gVgVgV\psi_{E,0} + \dots$$

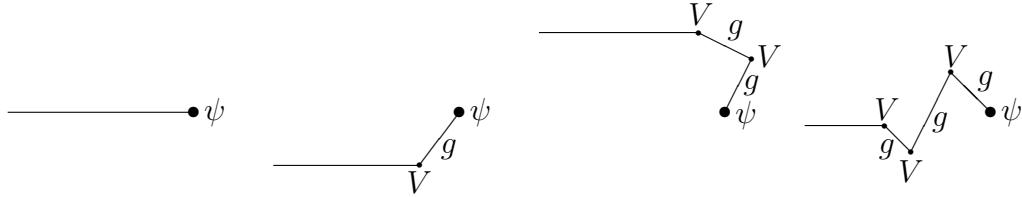


Figure A.20: Graphical interpretation of the Born series.

Graphically, these contributions to ψ_E can be represented as in figure A.20. The first term in the series simply takes ψ at some location \vec{r} from the unmodified incoming wave. The second contribution takes the incoming wave at an intermediate location \vec{r} , multiplies it by a “vertex factor” V , (the potential), and adds it to the wave function at \vec{r} multiplied by some factor g , the “propagator.” This must then be integrated over all possible locations of the intermediate point. The third term takes the wave function at some location, multiplies it by the vertex factor, propagates it to another location, multiplies it again by a vertex factor, and then propagates that to the wave function at \vec{r} . This is to be integrated over all possible locations of the two intermediate points. And so on.

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, [12], there does not seem much need to go into the details here.

A.57 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:

$$\int_{x=-\infty}^{\infty} |\Psi|^2 dx = 1 \text{ at all times}$$

Fortunately, the Schrödinger equation

$$i\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V\Psi$$

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 $\Psi^*/i\hbar$ and add the complex conjugate of the Schrödinger equation,

$$-i\hbar \frac{\partial \Psi^*}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi^*}{\partial x^2} + V\Psi^*,$$

times $-\Psi/i\hbar$. The potential energy terms drop out, and what is left is

$$\frac{\partial |\Psi|^2}{\partial t} = \frac{i\hbar}{2m} \left(\Psi^* \frac{\partial^2 \Psi}{\partial x^2} - \Psi \frac{\partial^2 \Psi^*}{\partial x^2} \right).$$

Now it can be verified by differentiating out that the right hand side can be rewritten as a derivative:

$$\frac{\partial |\Psi|^2}{\partial t} = -\frac{\partial J}{\partial x} \quad \text{where } J = \frac{i\hbar}{2m} \left(\Psi \frac{\partial \Psi^*}{\partial x} - \Psi^* \frac{\partial \Psi}{\partial x} \right) \quad (\text{A.76})$$

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.76) is integrated over all x , the desired result is obtained:

$$\frac{d}{dt} \int_{x=-\infty}^{\infty} |\Psi|^2 dx = -J \Big|_{x=-\infty}^{\infty} = 0.$$

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 $a \leq x \leq b$. That probability is given by

$$\int_{x=a}^b |\Psi|^2 dx$$

and it can change with time. A wave packet might enter or leave the region. In particular, integration of (A.76) gives

$$\frac{d}{dt} \int_{x=a}^b |\Psi|^2 dx = J_a - J_b$$

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:

$$\vec{J} = \frac{i\hbar}{2m} (\Psi \nabla \Psi^* - \Psi^* \nabla \Psi) \quad (\text{A.77})$$

and the net probability flowing out of a region is given by

$$\int \vec{J} \cdot \vec{n} dA \quad (\text{A.78})$$

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,

$$-\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V\psi = E\psi$$

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:

$$\begin{aligned} -\frac{\hbar^2}{2m} \frac{d^2\psi_1}{dx^2} + V\psi_1 &= E\psi_1 \\ -\frac{\hbar^2}{2m} \frac{d^2\psi_2}{dx^2} + V\psi_2 &= E\psi_2 \end{aligned}$$

If you multiply the first equation above by ψ_2 , the second by ψ_1 and then subtract the two, you get

$$\frac{\hbar^2}{2m} \left(\psi_1 \frac{d^2\psi_2}{dx^2} - \psi_2 \frac{d^2\psi_1}{dx^2} \right) = 0$$

The constant $\hbar^2/2m$ can be divided out, and by differentiation it can be verified that the remainder can be written as

$$\frac{dW}{dx} = 0 \quad \text{where } W = \psi_1 \frac{d\psi_2}{dx} - \psi_2 \frac{d\psi_1}{dx}$$

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.10 in note {A.53} 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

$$\psi_E = C_f^l e^{ip_c^l x/\hbar} + C_b^l e^{-ip_c^l x/\hbar} \text{ for } x < x_A \quad \text{where } p_c^l = \sqrt{2m(E - V_l)}$$

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

$$W = \frac{2ip_c^l}{\hbar} (|C_b^l|^2 - |C_f^l|^2).$$

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 $C_f^l e^{ip_c^l x/\hbar}$ 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 $C_b^l e^{-ip_c^l x/\hbar}$ 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 6.19. To the far left, the eigenfunction is again of the form

$$\psi_E = C_f^l e^{ip_c^l x/\hbar} + C_b^l e^{-ip_c^l x/\hbar} \text{ for } x \ll 0 \quad \text{where } p_c^l = \sqrt{2m(E - V_l)}$$

while at the far right it is now of the form

$$\psi_E = C^r e^{ip_c^r x/\hbar} \text{ for } x \gg 0 \quad \text{where } p_c^r = \sqrt{2m(E - V_r)}$$

The Wronskian can be found the same way as before:

$$W = \frac{2ip_c^l}{\hbar} (|C_b^l|^2 - |C_f^l|^2) = -\frac{2ip_c^r}{\hbar} |C^r|^2$$

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:

$$R = \frac{|C_b^l|^2}{|C_f^l|^2}$$

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

$$T = \frac{p_c^r |C^r|^2}{p_c^l |C_f^l|^2} \quad p_c^l = \sqrt{2m(E - V_l)} \quad p_c^r = \sqrt{2m(E - V_r)}$$

If the potential energy is the same at the far right and far left, the two classical momenta are the same, $p_c^r = p_c^l$. 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.58 A basic description of 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:

$$\begin{aligned} \text{stationarize: } & f(x_1, x_2, x_3, x_4) \\ \text{subject to: } & g(x_1, x_2, x_3, x_4) = 0, \quad h(x_1, x_2, x_3, x_4) = 0 \end{aligned}$$

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 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:

$$df = \frac{\partial \tilde{f}}{\partial g} dg + \frac{\partial \tilde{f}}{\partial h} dh + \frac{\partial \tilde{f}}{\partial x_3} dx_3 + \frac{\partial \tilde{f}}{\partial x_4} dx_4.$$

At the desired stationary point, acceptable changes in variables are those that keep g and h constant at zero; they have $dg = 0$ and $dh = 0$. 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 dx_3 and dx_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:

$$df - \frac{\partial \tilde{f}}{\partial g} dg - \frac{\partial \tilde{f}}{\partial h} dh = 0 \text{ always.}$$

In other words, if we “penalize” the change in f for going out of bounds by amounts dg and dh 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

$$df - \epsilon_1 dg - \epsilon_2 dh = 0$$

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 .

A.59 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

$$\psi = \varepsilon_1 \psi_1 + \varepsilon_2 \psi_2 + \dots + \varepsilon_{n-1} \psi_{n-1} + (1 + \varepsilon_n) \psi_n + \varepsilon_{n+1} \psi_{n+1} + \dots$$

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 $\varepsilon_1, \varepsilon_2, \dots$ are small.

The normalization condition $\langle \psi | \psi \rangle = 1$ is, using orthonormality:

$$1 = \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_{n-1}^2 + (1 + \varepsilon_n)^2 + \varepsilon_{n+1}^2 + \dots$$

The expectation energy is

$$\langle E \rangle = \varepsilon_1^2 E_1 + \varepsilon_2^2 E_2 + \dots + \varepsilon_{n-1}^2 E_{n-1} + (1 + \varepsilon_n)^2 E_n + \varepsilon_{n+1}^2 E_{n+1} + \dots$$

or plugging in the normalization condition to eliminate $(1 + \varepsilon_n)^2$

$$\langle E \rangle = \varepsilon_1^2 (E_1 - E_n) + \varepsilon_2^2 (E_2 - E_n) + \dots + \varepsilon_{n-1}^2 (E_{n-1} - E_n) + E_n + \varepsilon_{n+1}^2 (E_{n+1} - E_n) + \dots$$

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 $\varepsilon_1, \varepsilon_2, \dots$ will slightly change, by small amounts that will be indicated by $\delta\varepsilon_1, \delta\varepsilon_2, \dots$, and you get

$$\begin{aligned} \delta\langle E \rangle &= 2\varepsilon_1(E_1 - E_n)\delta\varepsilon_1 + 2\varepsilon_2(E_2 - E_n)\delta\varepsilon_2 + \dots \\ &\quad + 2\varepsilon_{n-1}(E_{n-1} - E_n)\delta\varepsilon_{n-1} + 2\varepsilon_{n+1}(E_{n+1} - E_n)\delta\varepsilon_{n+1} + \dots \end{aligned}$$

This is zero when $\varepsilon_1 = \varepsilon_2 = \dots = 0$, so when ψ is the exact eigenfunction ψ_n . And it is nonzero as soon as any of $\varepsilon_1, \varepsilon_2, \dots$ 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

$$0 = 2\varepsilon_1\delta\varepsilon_1 + 2\varepsilon_2\delta\varepsilon_2 + \dots + 2\varepsilon_{n-1}\delta\varepsilon_{n-1} + 2(1 + \varepsilon_n)\delta\varepsilon_n + 2\varepsilon_{n+1}\delta\varepsilon_{n+1} + \dots$$

At $\varepsilon_1 = 1$ and $\varepsilon_2 = \dots = \varepsilon_{n-1} = 1 + \varepsilon_n = \varepsilon_{n+1} = \dots = 0$, 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.

A.60 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 ψ^E with $H\psi^E = E^E\psi^E$ can be split into separate spatial functions for the distinct combinations of electron spin values as

$$\begin{aligned}\psi^E = & \psi_{+++}^E \uparrow\uparrow\uparrow + \psi_{+--}^E \uparrow\downarrow\downarrow + \psi_{-+-}^E \downarrow\uparrow\downarrow + \psi_{--+}^E \downarrow\downarrow\uparrow + \\ & \psi_{---}^E \downarrow\downarrow\downarrow + \psi_{-++}^E \downarrow\uparrow\uparrow + \psi_{+-+}^E \uparrow\downarrow\uparrow + \psi_{++-}^E \uparrow\uparrow\downarrow.\end{aligned}$$

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^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_{nn} are zero because the spins are orthogonal.

A.61 Derivation of the approximation

This note gives a derivation of the Born-Oppenheimer Hamiltonian eigenvalue problems (7.13) 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 $\psi_1^E, \psi_2^E, \dots$ 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

$$\psi = \sum_{\underline{n}} c_{\underline{n}} \psi_{\underline{n}}^E$$

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:

$$\begin{aligned} \psi(\vec{r}_1, S_{z1}, \dots, \vec{r}_I, S_{zI}, \vec{r}_1^n, S_{z1}^n, \dots, \vec{r}_J^n, S_{zJ}^n) = \\ \sum_{\underline{n}} c_{\underline{n}} (\vec{r}_1^n, S_{z1}^n, \dots, \vec{r}_J^n, S_{zJ}^n) \psi_{\underline{n}}^E(\vec{r}_1, S_{z1}, \dots, \vec{r}_I, S_{zI}; \vec{r}_1^n, S_{z1}^n, \dots, \vec{r}_J^n, S_{zJ}^n) \end{aligned}$$

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 $|\psi_{\underline{n}}^E|^2$ 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 $\psi_{\underline{n}}^E$, 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 $\psi_{\underline{n}}^N$ for when the electrons are in energy eigenstate $\psi_{\underline{n}}^E$. So from now on, just call it $\psi_{\underline{n}}^N$ instead of $c_{\underline{n}}$. The full wave function is then

$$\boxed{\psi = \sum \psi_{\underline{n}}^N \psi_{\underline{n}}^E} \quad (\text{A.79})$$

In the unsteady case, the $c_{\underline{n}}$, hence the $\psi_{\underline{n}}^N$, will also be functions of time. The $\psi_{\underline{n}}^E$ 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 $i\hbar\partial/\partial t$ replacing E .

So far, no approximations have been made; the only thing that has been done is to define the nuclear wave functions $\psi_{\underline{n}}^N$. But the objective is still to derive the claimed equation (7.13) for them. To do so plug the expression $\psi = \sum \psi_{\underline{n}}^N \psi_{\underline{n}}^E$ into the exact Hamiltonian eigenvalue problem:

$$[\hat{T}^N + \hat{T}^E + V^{NE} + V^{EE} + V^{NN}] \sum_{\underline{n}} \psi_{\underline{n}}^N \psi_{\underline{n}}^E = E \sum_{\underline{n}} \psi_{\underline{n}}^N \psi_{\underline{n}}^E$$

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 ψ_n^E . More precisely, multiply the entire equation by ψ_n^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 $\langle \psi_n^E | \psi_n^E \rangle_e$ 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 $E\psi_n^N$. Check, $E\psi_n^N$ is the correct right hand side in the nuclear-wave-function Hamiltonian eigenvalue problem (7.13).

Turning to the latter four terms in the left-hand side, remember that by definition the electron eigenfunctions satisfy

$$[\hat{T}^E + V^{NE} + V^{EE} + V^{NN}] \psi_{\underline{n}}^E = (E_{\underline{n}}^E + V^{NN}) \psi_{\underline{n}}^E$$

and if you then take an inner product of $\sum \psi_{\underline{n}}^N (E_{\underline{n}}^E + V^{NN}) \psi_{\underline{n}}^E$ with ψ_n^E , it is just like the earlier term, and you get $(E_n^E + V^{NN}) \psi_n^E$. Check, that are two of the terms in the left-hand side of (7.13) that you need.

That leaves only the nuclear kinetic term, and that one is a bit tricky. Recalling the definition (7.4) of the kinetic energy operator \hat{T}^N in terms of the nuclear coordinate Laplacians, you have

$$-\sum_{j=1}^J \sum_{\alpha=1}^3 \sum_{\underline{n}} \frac{\hbar^2}{2m_j^n} \frac{\partial^2}{\partial r_{\alpha j}^n} \psi_{\underline{n}}^N \psi_{\underline{n}}^E$$

Remember that not just the nuclear wave functions, but also the electron wave functions depend on the nuclear coordinates. So, if you differentiate the product, you get

$$-\sum_{j=1}^J \sum_{\alpha=1}^3 \sum_{\underline{n}} \frac{\hbar^2}{2m_j^n} \frac{\partial^2 \psi_{\underline{n}}^N}{\partial r_{\alpha j}^n} \psi_{\underline{n}}^E - \sum_{j=1}^J \sum_{\alpha=1}^3 \sum_{\underline{n}} \frac{\hbar^2}{m_j^n} \frac{\partial \psi_{\underline{n}}^N}{\partial r_{\alpha j}^n} \frac{\partial \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n} - \sum_{j=1}^J \sum_{\alpha=1}^3 \sum_{\underline{n}} \frac{\hbar^2}{2m_j^n} \psi_{\underline{n}}^N \frac{\partial^2 \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n}$$

Now if you take the inner product with electron eigenfunction ψ_n^E , the first term 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,

$$[\hat{T}^N + V^{NN} + E_n^E] \psi_n^N = E\psi_n^N + \sum_{\underline{n}} a_{n\underline{n}} \psi_{\underline{n}}^N$$

(A.80)

where

$$\hat{T}^N = - \sum_{j=1}^J \sum_{\alpha=1}^3 \frac{\hbar^2}{2m_j^n} \frac{\partial^2}{\partial r_{\alpha j}^n} \quad (\text{A.81})$$

$$a_{nn\underline{n}} = \sum_{j=1}^J \sum_{\alpha=1}^3 \frac{\hbar^2}{2m_j^n} \left(2 \left\langle \psi_n^E \left| \frac{\partial \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n} \right. \right\rangle \frac{\partial}{\partial r_{\alpha j}^n} + \left\langle \psi_n^E \left| \frac{\partial^2 \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n} \right. \right\rangle \right) \quad (\text{A.82})$$

The first thing to note is the final sum in (A.80). Unless you can talk away this sum as negligible, (7.13) is not valid. The “off-diagonal” coefficients, the $a_{nn\underline{n}}$ 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 (A.81) and (A.82) to see how big the various terms really are. At first appearance it might seem that both the nuclear kinetic energy \hat{T}^N and the coefficients $a_{nn\underline{n}}$ 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^E . But do not go too quick here. First ballpark the typical derivative, $\partial/\partial r_{\alpha j}^n$ when applied to the nuclear wave function. You can estimate such a derivative as $1/\ell^N$, where ℓ^N is the typical length over which there are significant changes in a nuclear wave function ψ_n^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 ℓ^E . It means that the distance ℓ^N is small. So the relative importance of the nuclear kinetic energy increases by a factor $(\ell^E/\ell^N)^2$ 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_{nn} ? Well, *normally* the electron eigenfunctions only change appreciable when you vary the nuclear positions over a length comparable to the electron blob scale ℓ^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 $\partial/\partial r_{\alpha j}^n$ applied to the electron wave functions are normally not by far as large as those applied to the nuclear wave functions, hence the a_{nn} terms are relatively small compared to the nuclear kinetic energy, and ignoring them is usually justified. So the final conclusion is that equation (7.13) 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_{nn} terms and (7.13) becomes invalid.

You can be a bit more precise about that claim with a few tricks. Consider the factors

$$\left\langle \psi_n^E \left| \frac{\partial \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n} \right. \right\rangle$$

appearing in the a_{nn} , (A.82). First of all, these factors are zero when $\underline{n} = n$. The reason is that because of orthonormality, $\langle \psi_n^E | \psi_n^E \rangle = 1$, and taking the $\partial/\partial r_{\alpha j}^n$ 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:

$$\begin{aligned} \left\langle \psi_n^E \left| \frac{\partial}{\partial r_{\alpha j}^n} H^E - H^E \frac{\partial}{\partial r_{\alpha j}^n} \right| \psi_{\underline{n}}^E \right\rangle &= (E_{\underline{n}}^E - E_n^E) \left\langle \psi_n^E \left| \frac{\partial \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n} \right. \right\rangle \\ &= \frac{Z_j e^2}{4\pi\epsilon_0} \sum_{i=1}^I \left\langle \psi_n^E \left| \frac{r_{\alpha j}^n - r_{\alpha i}}{r_{ij}^3} \right| \psi_{\underline{n}}^E \right\rangle \end{aligned}$$

The first equality is just a matter of the definition of the electron eigenfunctions and taking the second H^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 subsection 7.2.1 and then working out the commutator in the leftmost inner product. (V^{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_{nn} , (A.82), blows up like $1/(E_n^E - E_{\underline{n}}^E)$ when those energy levels become equal.

As far as the final term in a_{nn} is concerned, like the second term, you would expect it to become important when the scale of non-trivial 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,

$$\left\langle \frac{\partial \psi_n^E}{\partial r_{\alpha j}^n} \left| \frac{\partial \psi_{\underline{n}}^E}{\partial r_{\alpha j}^n} \right. \right\rangle + \left\langle \psi_n^E \left| \frac{\partial^2 \psi_{\underline{n}}^E}{\partial r_{\alpha j}^{n/2}} \right. \right\rangle = \frac{\partial}{\partial r_{\alpha j}^n} \frac{1}{E_{\underline{n}}^E - E_n^E} \frac{Z_j e^2}{4\pi\epsilon_0} \sum_{i=1}^I \left\langle \psi_n^E \left| \frac{r_{\alpha j}^n - r_{\alpha i}}{r_{ij}^3} \right| \psi_{\underline{n}}^E \right\rangle$$

The first term should not be large: while the left hand side of the inner product has a large component along $\psi_{\underline{n}}^E$, the other side has zero component and vice-versa. The final term should be of order $1/(E_{\underline{n}}^E - E_n^E)^2$, 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 $a_{\underline{n}\underline{n}}$ is of order $1/(E_{\underline{n}}^{\text{E}} - E_n^{\text{E}})^2$. 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_{nn} , 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 ψ_n^{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 (A.82) for the a_{nn} 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 ψ_n^{N} , the diagonal term a_{nn} multiplies the predominant nuclear wave function ψ_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.

A.62 Why 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

$$F(x, y) = \frac{a}{\sqrt{2}} (f_1(x)f_2(y) - f_2(x)f_1(y))$$

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

$$F(x, y_1) = c_{11}f_1(x) + c_{12}f_2(x) \quad F(x, y_2) = c_{21}f_1(x) + c_{22}f_2(x)$$

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 $f_1^{(2)}, f_2^{(2)}, f_1^{(3)}, f_2^{(3)}, \dots$ 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.

A.63 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

$$\Psi = \frac{1}{\sqrt{I!}} \left| \det(\psi_1^s \uparrow_1, \psi_2^s \uparrow_2, \psi_3^s \uparrow_3, \dots) \right\rangle$$

you are going to get terms, or Hartree products if you want, of the form

$$\frac{\pm}{\sqrt{I!}} \psi_{n_1}^s(\vec{r}_1) \uparrow_{n_1}(S_{z1}) \psi_{n_2}^s(\vec{r}_2) \uparrow_{n_2}(S_{z2}) \psi_{n_3}^s(\vec{r}_3) \uparrow_{n_3}(S_{z3}) \dots$$

where the numbers n_1, n_2, n_3, \dots 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, ..., 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 $I(I - 1)(I - 2) \dots 2 \cdot 1 = I!$ 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,

$$\begin{aligned} \sqrt{3!} |\det(\psi_1^s \uparrow_1, \psi_2^s \uparrow_2, \psi_3^s \uparrow_3)\rangle = \\ +\psi_1^s(\vec{r}_1) \downarrow_1(S_{z1}) \psi_2^s(\vec{r}_2) \downarrow_2(S_{z2}) \psi_3^s(\vec{r}_3) \downarrow_3(S_{z3}) \\ -\psi_1^s(\vec{r}_1) \downarrow_1(S_{z1}) \psi_3^s(\vec{r}_2) \downarrow_3(S_{z2}) \psi_2^s(\vec{r}_3) \downarrow_2(S_{z3}) \\ -\psi_2^s(\vec{r}_1) \downarrow_2(S_{z1}) \psi_1^s(\vec{r}_2) \downarrow_1(S_{z2}) \psi_3^s(\vec{r}_3) \downarrow_3(S_{z3}) \\ +\psi_2^s(\vec{r}_1) \downarrow_2(S_{z1}) \psi_3^s(\vec{r}_2) \downarrow_3(S_{z2}) \psi_1^s(\vec{r}_3) \downarrow_1(S_{z3}) \\ +\psi_3^s(\vec{r}_1) \downarrow_3(S_{z1}) \psi_1^s(\vec{r}_2) \downarrow_1(S_{z2}) \psi_2^s(\vec{r}_3) \downarrow_2(S_{z3}) \\ -\psi_3^s(\vec{r}_1) \downarrow_3(S_{z1}) \psi_2^s(\vec{r}_2) \downarrow_2(S_{z2}) \psi_1^s(\vec{r}_3) \downarrow_1(S_{z3}) \end{aligned}$$

The first row in the expansion covers the possibility that $n_1 = 1$, with the first term the possibility that $n_2 = 2$ and the second term the possibility that $n_2 = 3$; note that there are then no choices left for n_3 . The second row covers the possibilities in which $n_1 = 2$, and the third $n_1 = 3$. You see that there are $3! = 6$ Hartree product terms total.

Next, recall that the Hamiltonian consists of single-electron Hamiltonians h_i^e and electron-pair repulsion potentials v_{ii}^{ee} . The expectation value of a single electron Hamiltonian h_i^e will be done first. In forming the inner product $\langle \Psi | h_i^e | \Psi \rangle$, 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

$$\left\langle \frac{\pm}{\sqrt{I!}} \psi_{n_1}^s(\vec{r}_1) \downarrow_{n_1}(S_{z1}) \psi_{n_2}^s(\vec{r}_2) \downarrow_{n_2}(S_{z2}) \dots \psi_{n_I}^s(\vec{r}_I) \downarrow_{n_I}(S_{zI}) \right| h_i^e \left| \frac{\pm}{\sqrt{I!}} \psi_{\bar{n}_1}^s(\vec{r}_1) \downarrow_{\bar{n}_1}(S_{z1}) \psi_{\bar{n}_2}^s(\vec{r}_2) \downarrow_{\bar{n}_2}(S_{z2}) \dots \psi_{\bar{n}_I}^s(\vec{r}_I) \downarrow_{\bar{n}_I}(S_{zI}) \right\rangle$$

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

$$\begin{aligned} & \frac{\pm}{\sqrt{I!}} \frac{\pm}{\sqrt{I!}} \left\langle \psi_{n_1}^s(\vec{r}_1) \uparrow_{n_1}(S_{z1}) \middle| \psi_{\bar{n}_1}^s(\vec{r}_1) \uparrow_{\bar{n}_1}(S_{z1}) \right\rangle \\ & \quad \times \left\langle \psi_{n_2}^s(\vec{r}_2) \uparrow_{n_2}(S_{z2}) \middle| \psi_{\bar{n}_2}^s(\vec{r}_2) \uparrow_{\bar{n}_2}(S_{z2}) \right\rangle \\ & \quad \times \dots \\ & \quad \times \left\langle \psi_{n_i}^s(\vec{r}_i) \uparrow_{n_i}(S_{zi}) \middle| h_i^e \middle| \psi_{\bar{n}_i}^s(\vec{r}_i) \uparrow_{\bar{n}_i}(S_{zi}) \right\rangle \\ & \quad \times \dots \\ & \quad \times \left\langle \psi_{n_I}^s(\vec{r}_I) \uparrow_{n_I}(S_{zI}) \middle| \psi_{\bar{n}_I}^s(\vec{r}_I) \uparrow_{\bar{n}_I}(S_{zI}) \right\rangle \end{aligned}$$

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:

$$n_1 = \bar{n}_1, n_2 = \bar{n}_2, \dots, n_{i-1} = \bar{n}_{i-1}, n_{i+1} = \bar{n}_{i+1}, \dots, n_I = \bar{n}_I$$

Note that $\langle \psi_{n_i}^s(\vec{r}_i) \uparrow_{n_i}(S_{zi}) | h_i^e | \psi_{\bar{n}_i}^s(\vec{r}_i) \uparrow_{\bar{n}_i}(S_{zi}) \rangle$ does not require $n_i = \bar{n}_i$ 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, \dots in an individual term are all different. So the numbers $n_1, n_2, \dots, n_{i-1}, n_{i+1}, \dots$ include all the numbers that are *not* equal to n_i . Then so do $\bar{n}_1, \bar{n}_2, \dots, \bar{n}_{i-1}, \bar{n}_{i+1}, \dots$, because they are the same. And since \bar{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 \bar{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:

$$\frac{\pm}{\sqrt{I!}} \frac{\pm}{\sqrt{I!}} \left\langle \psi_{n_i}^s(\vec{r}_i) \uparrow_{n_i}(S_{zi}) \middle| h_i^e \middle| \psi_{n_i}^s(\vec{r}_i) \uparrow_{n_i}(S_{zi}) \right\rangle$$

Also, the two signs are equal, because with all the \bar{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:

$$\frac{1}{I!} \left\langle \psi_{n_i}^s(\vec{r}_i) \middle| h_i^e \middle| \psi_{n_i}^s(\vec{r}_i) \right\rangle \left\langle \uparrow_{n_i}(S_{zi}) \middle| \uparrow_{n_i}(S_{zi}) \right\rangle = \frac{1}{I!} \left\langle \psi_{n_i}^s(\vec{r}_i) \middle| h_i^e \middle| \psi_{n_i}^s(\vec{r}_i) \right\rangle \equiv \frac{1}{I!} E_n^e$$

where for brevity the remaining inner product was called E_n^e . Normally you would call it $E_{n,i}^e$, 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, ... 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 $(I - 1)(I - 2) \dots 1 = (I - 1)!$ 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_n^e/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

$$\sum_{n=1}^I E_n^e \quad E_n^e = \langle \psi_n^s(\vec{r}) | h^e | \psi_n^s(\vec{r}) \rangle$$

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_{ii}^{ee} between an electron numbered i and another numbered \underline{i} makes a contribution to the expectation energy equal to $\langle \Psi | v_{ii}^{ee} | \Psi \rangle$, and if you multiply out Ψ , you get terms of the general form:

$$\frac{1}{I!} \left\langle \psi_{n_1}^s(\vec{r}_1) \downarrow_{n_1} (S_{z1}) \psi_{n_2}^s(\vec{r}_2) \downarrow_{n_2} (S_{z2}) \dots \psi_{n_i}^s(\vec{r}_i) \downarrow_{n_i} (S_{zi}) \dots \psi_{n_{\underline{i}}}^s(\vec{r}_{\underline{i}}) \downarrow_{n_{\underline{i}}} (S_{z\underline{i}}) \dots \right| \\ v_{ii}^{ee} \left| \psi_{\bar{n}_1}^s(\vec{r}_1) \downarrow_{\bar{n}_1} (S_{z1}) \psi_{\bar{n}_2}^s(\vec{r}_2) \downarrow_{\bar{n}_2} (S_{z2}) \dots \psi_{\bar{n}_i}^s(\vec{r}_i) \downarrow_{\bar{n}_i} (S_{zi}) \dots \psi_{\bar{n}_{\underline{i}}}^s(\vec{r}_{\underline{i}}) \downarrow_{\bar{n}_{\underline{i}}} (S_{z\underline{i}}) \dots \right\rangle$$

You can again split into a product of individual inner products, except that you cannot split between electrons i and \underline{i} since v_{ii}^{ee} involves both electrons in a nontrivial way. Still, you get again that all the other n -values must be the same as the corresponding \bar{n} -values, eliminating those inner products from the expression:

$$\frac{1}{I!} \left\langle \psi_{n_i}^s(\vec{r}_i) \downarrow_{n_i} (S_{zi}) \psi_{n_{\underline{i}}}^s(\vec{r}_{\underline{i}}) \downarrow_{n_{\underline{i}}} (S_{z\underline{i}}) \right| v_{ii}^{ee} \left| \psi_{\bar{n}_i}^s(\vec{r}_i) \downarrow_{\bar{n}_i} (S_{zi}) \psi_{\bar{n}_{\underline{i}}}^s(\vec{r}_{\underline{i}}) \downarrow_{\bar{n}_{\underline{i}}} (S_{z\underline{i}}) \right\rangle$$

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 $n = \bar{n}$ -values of the other $I - 2$ electrons.

Next, \bar{n}_i and $\bar{n}_{\underline{i}}$ 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:

$$\bar{n}_i = n_i, \bar{n}_{\underline{i}} = n_{\underline{i}} \quad \text{or} \quad \bar{n}_i = n_{\underline{i}}, \bar{n}_{\underline{i}} = n_i.$$

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 \underline{n}_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 \underline{n}_i , so for each pair of electrons i and $\underline{i} \neq i$, and each pair of states $n = n_i$ and $\underline{n} = \underline{n}_i$, you get one term

$$\frac{1}{I(I-1)} J_{nn}$$

with

$$J_{nn} \equiv \langle \psi_n^s(\vec{r}) \psi_{\underline{n}}^s(\vec{r}) | v^{ee} | \psi_n^s(\vec{r}) \psi_{\underline{n}}^s(\vec{r}) \rangle.$$

Again, the J_{nn} are the same regardless of what i and \underline{i} are; they depend only on what $n = n_i$ and $\underline{n} = \underline{n}_i$ are. So the subscripts i and \underline{i} were left out, after setting $\vec{r} = \vec{r}_i$ and $\underline{\vec{r}} = \underline{\vec{r}}_i$.

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

$$\frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I J_{nn}$$

The factor $\frac{1}{2}$ was added since for every electron pair, you are summing both v_{ii}^{ee} and $v_{\underline{i}\underline{i}}^{ee}$, 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 \bar{n}_i and $\bar{n}_{\underline{i}}$ reversed, producing a change of sign, and that the inner product of the spins is not trivial. Define

$$K_{nn} \equiv \langle \psi_n^s(\vec{r}) \psi_{\underline{n}}^s(\vec{r}) | v^{ee} | \psi_{\underline{n}}^s(\vec{r}) \psi_n^s(\vec{r}) \rangle.$$

and then the total contribution is

$$-\frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I K_{nn} \langle \uparrow_n | \uparrow_{\underline{n}} \rangle^2$$

Finally, you can leave the constraint $\underline{n} \neq n$ on the sums away since $K_{nn} = J_{nn}$, so they cancel each other.

A.64 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

$$\begin{aligned} J_{nn} &= \langle \psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}}) | v_{ii}^{ee} | \psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}}) \rangle \\ &= \int_{\text{all } \vec{r}_i} \int_{\text{all } \vec{r}_{\underline{i}}} |\psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}})|^2 \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}} d^3\vec{r}_i d^3\vec{r}_{\underline{i}}. \end{aligned}$$

is real and positive is self-evident, since it is an integral of a real and positive function.

The fact that

$$\begin{aligned} K_{nn} &= \langle \psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}}) | v_{ii}^{ee} | \psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}}) \rangle \\ &= \int_{\text{all } \vec{r}_i} \int_{\text{all } \vec{r}_{\underline{i}}} \psi_n^s(\vec{r}_i)^* \psi_{\underline{n}}^s(\vec{r}_{\underline{i}})^* \frac{e^2}{4\pi\epsilon_0} \frac{1}{r_{ii}} \psi_{\underline{n}}^s(\vec{r}_i) \psi_n^s(\vec{r}_{\underline{i}}) d^3\vec{r}_i d^3\vec{r}_{\underline{i}} \end{aligned}$$

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_{nn} and K_{nn} are symmetric matrices; $J_{nn} = J_{\underline{n}n}$ and $K_{nn} = K_{\underline{n}n}$.

That K_{nn} is positive is a bit trickier; write it as

$$\int_{\text{all } \vec{r}_i} -ef^*(\vec{r}_i) \left(\int_{\text{all } \vec{r}_{\underline{i}}} \frac{-ef(\vec{r}_{\underline{i}})}{4\pi\epsilon_0} \frac{1}{r_{ii}} d^3\vec{r}_{\underline{i}} \right) d^3\vec{r}_i$$

with $f = \psi_{\underline{n}}^s \psi_n^s$. 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, $\vec{E} = -\nabla V$, and according to Maxwell's equation, the divergence of the electric field is the charge density divided by ϵ_0 : $\text{div} \vec{E} = -\nabla^2 V = -ef/\epsilon_0$. So $-ef^* = -\epsilon_0 \nabla^2 V^*$ and the integral is

$$-\epsilon_0 \int_{\text{all } \vec{r}_i} V \nabla^2 V^* d^3\vec{r}_i$$

and integration by parts shows it is positive. Or zero, if $\psi_{\underline{n}}^s$ is zero wherever ψ_n^s is not, and vice versa.

To show that $J_{nn} \geq K_{nn}$, note that

$$\langle \psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}}) - \psi_{\underline{n}}^s(\vec{r}_i) \psi_n^s(\vec{r}_{\underline{i}}) | v_{ii}^{ee} | \psi_n^s(\vec{r}_i) \psi_{\underline{n}}^s(\vec{r}_{\underline{i}}) - \psi_{\underline{n}}^s(\vec{r}_i) \psi_n^s(\vec{r}_{\underline{i}}) \rangle$$

is nonnegative, for the same reasons as J_{nn} but with $\psi_n^s \psi_{\underline{n}}^s - \psi_{\underline{n}}^s \psi_n^s$ replacing $\psi_n^s \psi_{\underline{n}}^s$. If you multiply out the inner product, you get that $2J_{nn} - 2K_{nn}$ is nonnegative, so $J_{nn} \geq K_{nn}$.

A.65 Generalized orbitals

This note has a brief look at generalized orbitals of the form

$$\psi_n^p(\vec{r}) = \psi_{n+}^s(\vec{r})\uparrow(S_z) + \psi_{n-}^s(\vec{r})\downarrow(S_z).$$

For such orbitals, the expectation energy can be worked out in exactly the same way as in {A.63}, except without simplifying the spin terms. The energy is

$$\langle E \rangle = \sum_{n=1}^I \left\langle \psi_n^p \middle| h^e \right| \psi_n^p \rangle + \frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I \left\langle \psi_n^p \psi_{\underline{n}}^p \middle| v^{ee} \right| \psi_n^p \psi_{\underline{n}}^p \rangle - \frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I \left\langle \psi_n^p \psi_{\underline{n}}^p \middle| v^{ee} \right| \psi_{\underline{n}}^p \psi_n^p \rangle$$

To multiply out to the individual spin terms, it is convenient to normalize the spatial functions, and write

$$\psi_n^p = c_{n+} \psi_{n+,0}^s \uparrow + c_{n-} \psi_{n-,0}^s \downarrow,$$

$$\langle \psi_{n+,0}^s | \psi_{n+,0}^s \rangle = \langle \psi_{n-,0}^s | \psi_{n-,0}^s \rangle = 1, \quad |c_{n+}|^2 + |c_{n-}|^2 = 1$$

In that case, the expectation energy multiplies out to

$$\begin{aligned} \langle E \rangle &= \sum_{n=1}^I \left\langle \psi_{n+,0}^s \middle| h^e \right| \psi_{n+,0}^s \rangle |c_{n+}|^2 + \sum_{n=1}^I \left\langle \psi_{n-,0}^s \middle| h^e \right| \psi_{n-,0}^s \rangle |c_{n-}|^2 \\ &\quad + \frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I \left(\left\langle \psi_{n+,0}^s \psi_{\underline{n}+,0}^s \middle| v^{ee} \right| \psi_{n+,0}^s \psi_{\underline{n}+,0}^s \right. \\ &\quad \left. - \left\langle \psi_{n+,0}^s \psi_{\underline{n}+,0}^s \middle| v^{ee} \right| \psi_{\underline{n}+,0}^s \psi_{n+,0}^s \right) |c_{n+}|^2 |c_{\underline{n}+}|^2 \\ &\quad + \frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I 2 \left\langle \psi_{n+,0}^s \psi_{\underline{n}-,0}^s \middle| v^{ee} \right| \psi_{n+,0}^s \psi_{\underline{n}-,0}^s \rangle |c_{n+}|^2 |c_{\underline{n}-}|^2 \\ &\quad + \frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I \left(\left\langle \psi_{n-,0}^s \psi_{\underline{n}-,0}^s \middle| v^{ee} \right| \psi_{n-,0}^s \psi_{\underline{n}-,0}^s \right. \\ &\quad \left. - \left\langle \psi_{n-,0}^s \psi_{\underline{n}-,0}^s \middle| v^{ee} \right| \psi_{\underline{n}-,0}^s \psi_{n-,0}^s \right) |c_{n-}|^2 |c_{\underline{n}-}|^2 \\ &\quad - \frac{1}{2} \sum_{n=1}^I \sum_{\substack{\underline{n}=1 \\ \underline{n} \neq n}}^I 2 \Re \left(\left\langle \psi_{n+,0}^s \psi_{\underline{n}-,0}^s \middle| v^{ee} \right| \psi_{\underline{n}+,0}^s \psi_{n-,0}^s \right) c_{n+}^* c_{n-} c_{\underline{n}-}^* c_{\underline{n}+} \end{aligned}$$

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 $\psi_m^s \uparrow \equiv \psi_{m+,0}^s \uparrow$ 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 maximum 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.

A.66 Derivation of the Hartree-Fock equations

This note derives the canonical Hartree-Fock equations. The derivation below will be performed under the normally stated rules of engagement that the orbitals are of the form $\psi_n^s \uparrow$ or $\psi_n^s \downarrow$, so that only the spatial orbitals ψ_n^s are continuously variable. The derivations allow for the fact that some spatial spin states may be constrained to be equal.

First, you can make things a lot less messy by a priori specifying the ordering of the orbitals. The ordering makes no difference physically, and it simplifies the mathematics. In particular, in restricted Hartree-Fock some spatial orbitals appear in pairs, but you can only count each spatial orbital as one unknown function. The easiest way to handle that is to push the spin-down versions of the duplicated spatial orbitals to the end of the Slater determinant. Then the start of the determinant comprises a list of unique spatial orbitals.

So, it will be assumed that the orbitals are ordered as follows:

1. the paired spatial states in their spin-up version; assume there are $N_p \geq 0$ of them;
2. unpaired spin-up states; assume there are N_u of them;
3. unpaired spin-down states; assume there are N_d of them;
4. and finally, the paired spatial states in their spin-down version.

That means that the Slater-determinant wave function looks like:

$$\frac{1}{\sqrt{I!}} \left| \det(\psi_1^s \uparrow, \dots, \psi_{N_p}^s \uparrow, \psi_{N_p+1}^s \uparrow, \dots, \psi_{N_1}^s \uparrow, \psi_{N_1+1}^s \downarrow, \dots, \psi_N^s \downarrow, \psi_{N+1}^s \downarrow, \dots, \psi_I^s \downarrow) \right\rangle$$

$$N_1 = N_p + N_u \quad N = N_p + N_u + N_d \\ \psi_{N+1}^s = \psi_1^s, \quad \psi_{N+2}^s = \psi_2^s, \quad \dots, \quad \psi_I^s = \psi_{N_p}^s$$

The total number of unknown spatial orbitals is N , and you need a corresponding N equations for them.

The variational method discussed in section 7.1 says that the expectation energy must be unchanged under small changes in the orbitals, provided that penalty terms are added for changes that violate the orthonormality requirements on the orbitals.

The expectation value of energy was in subsection 7.3.3 found to be:

$$\begin{aligned} \langle E \rangle &= \sum_{n=1}^I \langle \psi_n^s | h^e | \psi_n^s \rangle \\ &+ \frac{1}{2} \sum_{n=1}^I \sum_{\underline{n}=1}^I \langle \psi_n^s \psi_{\underline{n}}^s | v^{ee} | \psi_n^s \psi_{\underline{n}}^s \rangle - \frac{1}{2} \sum_{n=1}^I \sum_{\underline{n}=1}^I \langle \psi_n^s \psi_{\underline{n}}^s | v^{ee} | \psi_{\underline{n}}^s \psi_n^s \rangle \langle \uparrow_n | \uparrow_{\underline{n}} \rangle^2 \end{aligned}$$

(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 $\underline{\vec{r}}$)

The penalty terms require penalty factors called Lagrangian variables. The penalty factor for violations of the normalization requirement

$$\langle \psi_n^s | \psi_n^s \rangle = 1$$

will be called ϵ_{nn} for reasons evident in a second. Orthogonality between any two spatial orbitals ψ_n^s and $\psi_{\underline{n}}^s$ requires

$$\frac{1}{2} (\langle \psi_n^s | \psi_{\underline{n}}^s \rangle + \langle \psi_{\underline{n}}^s | \psi_n^s \rangle) = 0, \quad \frac{1}{2} i (\langle \psi_n^s | \psi_{\underline{n}}^s \rangle - \langle \psi_{\underline{n}}^s | \psi_n^s \rangle) = 0.$$

where the first constraint says that the real part of $\langle \psi_n^s | \psi_{\underline{n}}^s \rangle$ must be zero and the second that its imaginary part must be zero too. (Remember that if you switch sides in an inner product, you turn it into its complex conjugate.) To avoid including the same orthogonality condition twice, the constraint will be written only for $\underline{n} > n$. The penalty factor for the first constraint will be called $2\epsilon_{n\underline{n},r}$, and the one for the second constraint $2\epsilon_{n\underline{n},i}$.

In those terms, the penalized change in expectation energy becomes, in the restricted case that all unique spatial orbitals are mutually orthogonal,

$$\delta\langle E \rangle - \sum_{n=1}^N \sum_{\underline{n}=1}^N \epsilon_{n\underline{n}} \delta\langle \psi_n^s | \psi_{\underline{n}}^s \rangle = 0$$

where $\epsilon_{n\underline{n}}$ is an Hermitian matrix, with $\epsilon_{n\underline{n}} = \epsilon_{n\underline{n},r} + i\epsilon_{n\underline{n},i}$. The notations for the Lagrangian variables were chosen above to achieve this final result.

But for unrestricted Hartree-Fock, spatial orbitals are not required to be orthogonal if they have opposite spin, because the spins will take care of orthogonality. You can remove the erroneously added constraints by simply specifying that the corresponding Lagrangian variables are zero:

$$\epsilon_{n\underline{n}} = 0 \text{ if unrestricted Hartree-Fock and } \langle \uparrow_n | \uparrow_{\underline{n}} \rangle = 0$$

or equivalently, if $n \leq N_u$, $\underline{n} > N_u$ or $n > N_u$, $\underline{n} \leq N_u$.

Now work out the penalized change in expectation energy due to a change in the values of a selected spatial orbital ψ_m^s with $m \leq N$. It is

$$\begin{aligned} & \sum_{\psi_n^s = \psi_m^s} \left(\langle \delta\psi_m^s | h^e | \psi_m^s \rangle + \langle \psi_m^s | h^e | \delta\psi_m^s \rangle \right) \\ & + \frac{1}{2} \sum_{\psi_n^s = \psi_m^s} \sum_{\underline{n}=1}^I \left(\langle \delta\psi_m^s \psi_{\underline{n}}^s | v^{ee} | \psi_m^s \psi_{\underline{n}}^s \rangle + \langle \psi_m^s \psi_{\underline{n}}^s | v^{ee} | \delta\psi_m^s \psi_{\underline{n}}^s \rangle \right) \\ & + \frac{1}{2} \sum_{n=1}^I \sum_{\psi_n^s = \psi_m^s} \left(\langle \psi_n^s \delta\psi_m^s | v^{ee} | \psi_n^s \psi_m^s \rangle + \langle \psi_n^s \psi_m^s | v^{ee} | \psi_n^s \delta\psi_m^s \rangle \right) \\ & - \frac{1}{2} \sum_{\psi_n^s = \psi_m^s} \sum_{\underline{n}=1}^I \left(\langle \delta\psi_m^s \psi_{\underline{n}}^s | v^{ee} | \psi_{\underline{n}}^s \psi_m^s \rangle + \langle \psi_m^s \psi_{\underline{n}}^s | v^{ee} | \psi_{\underline{n}}^s \delta\psi_m^s \rangle \right) \langle \uparrow_n | \uparrow_{\underline{n}} \rangle^2 \\ & - \frac{1}{2} \sum_{n=1}^I \sum_{\psi_n^s = \psi_m^s} \left(\langle \psi_n^s \delta\psi_m^s | v^{ee} | \psi_m^s \psi_n^s \rangle + \langle \psi_n^s \psi_m^s | v^{ee} | \delta\psi_m^s \psi_n^s \rangle \right) \langle \uparrow_n | \uparrow_{\underline{n}} \rangle^2 \\ & - \sum_{n=1}^N \epsilon_{m\underline{n}} \langle \delta\psi_m^s | \psi_{\underline{n}}^s \rangle - \sum_{n=1}^N \epsilon_{nm} \langle \psi_n^s | \delta\psi_m^s \rangle = 0 \end{aligned}$$

OK, OK it is a mess. Sums like $\psi_n^s = \psi_m^s$ are for the restricted Hartree-Fock case, in which spatial orbital ψ_m^s may appear twice in the Slater determinant. From now on, just write them as [2], meaning, put in a factor 2 if orbital ψ_m^s appears twice. The exception is for the exchange integrals which produce exactly one nonzero spin product; write that as $[\langle \downarrow_n | \uparrow_m \rangle^2]$, meaning take out that product if the orbital appears twice.

Next, note that the second term in each row is just the complex conjugate of the first. Considering $i\delta\psi_m^s$ as a second possible change in orbital, as was done in the example in section 7.1, it is seen that the first terms by themselves must be zero, so you can just ignore the second term in each row. And the integrals with the factors $\frac{1}{2}$ are pairwise the same; the difference is just a name swap of the first and second summation and integration variables. So all that you really have left is

$$\begin{aligned} & [2]\langle \delta\psi_m^s | h^e | \psi_m^s \rangle + \\ & \sum_{n=1}^I \left\{ [2]\langle \delta\psi_m^s \psi_n^s | v^{ee} | \psi_m^s \psi_n^s \rangle - \langle \delta\psi_m^s \psi_n^s | v^{ee} | \psi_n^s \psi_m^s \rangle [\langle \uparrow_n | \uparrow_m \rangle^2] \right\} - \\ & \sum_{n=1}^N \epsilon_{mn} \langle \delta\psi_m^s | \psi_n^s \rangle \end{aligned}$$

Now note that if you write out the inner product over the first position coordinate, you will get an integral of the general form

$$\begin{aligned} & \int_{\text{all } \vec{r}} \delta\psi_m^s * \left([2]h^e \psi_m^s \right. \\ & \left. + \sum_{n=1}^I \left\{ [2]\langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s [\langle \uparrow_n | \uparrow_m \rangle^2] \right\} - \sum_{n=1}^N \epsilon_{mn} \psi_n^s \right) d^3\vec{r} \end{aligned}$$

If this integral is to be zero for *whatever* you take $\delta\psi_m^s$, then the terms within the parentheses must be zero. (Just take $\delta\psi_m^s$ proportional to the parenthetical expression; you would get the integral of an absolute square, only zero if the square is.) Unavoidably, you must have that

$$[2]h^e \psi_m^s + \sum_{n=1}^I [2]\langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \sum_{n=1}^I [\langle \uparrow_n | \uparrow_m \rangle^2] \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s = \sum_{n=1}^N \epsilon_{mn} \psi_n^s$$

You can divide by [2]:

$$h^e \psi_m^s + \sum_{n=1}^I \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \sum_{n=1}^I \left\{ \frac{\langle \uparrow_n | \uparrow_m \rangle^2}{\frac{1}{2}} \right\} \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s = \sum_{n=1}^N \left\{ \frac{1}{\frac{1}{2}} \right\} \epsilon_{mn} \psi_n^s$$

(A.83)

where you use the lower of the choices between the braces in the case that spatial orbital ψ_m^s appears twice in the Slater determinant, or equivalently, if $m \leq N_p$. There you have the Hartree-Fock equations, one for each $m \leq N$. Recall that they apply assuming the ordering of the Slater determinant given in the beginning, and that for unrestricted Hartree-Fock, ϵ_{mn} is zero if $\langle \downarrow_m | \uparrow_n \rangle = 0$ is.

How about those ϵ_{mn} , you say? Shouldn't the right hand side just be $\epsilon_m \psi_m^s$? Ah, you want the *canonical* Hartree-Fock equations, not just the plain vanilla version.

OK, let's do the restricted closed-shell Hartree-Fock case first, then, since it is the easiest one. Every state is paired, so the lower choice in the curly brackets always applies, and the number of unique unknown spatial states is $N = I/2$. Also, you can reduce the summation upper limits $n = I$ to $n = N = I/2$ if you add a factor 2, since the second half of the spatial orbitals are the same as the first half. So you get

$$h^e \psi_m^s + 2 \sum_{n=1}^N \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \sum_{n=1}^N \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s = \sum_{n=1}^N \frac{1}{2} \epsilon_{mn} \psi_n^s$$

Now, suppose that you define a *new* set of orbitals, each a linear combination of the current ones:

$$\bar{\psi}_\nu^s \equiv \sum_{n=1}^N u_{\nu n} \psi_n^s \quad \text{for } \nu = 1, 2, \dots, N$$

where the $u_{\nu n}$ are the multiples of the original orbitals. Will the new ones still be an orthonormal set? Well, they will be if

$$\langle \bar{\psi}_\mu^s | \bar{\psi}_\nu^s \rangle = \delta_{\mu\nu}$$

where $\delta_{\mu\nu}$ is the Kronecker delta, one if $\mu = \nu$, zero otherwise. Substituting in the definition of the new orbitals, making sure not to use the same name for two different indices,

$$\sum_{m=1}^N \sum_{n=1}^N u_{\mu m}^* u_{\nu n} \langle \psi_m^s | \psi_n^s \rangle = \delta_{\mu\nu}.$$

Now note that the ψ_n^s are orthonormal, so to get a nonzero value, m must be n , and you get

$$\sum_{n=1}^N u_{\mu n}^* u_{\nu n} = \delta_{\mu\nu}.$$

Consider n to be the component index of a vector. Then this really says that vectors \vec{u}_μ and \vec{u}_ν must be orthonormal. So the “matrix of coefficients” 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.

Call the complete matrix U . Then, according to the rules of matrix multiplication and Hermitian adjoint, the orthonormality condition above is equivalent to $UU^H = I$ where I is the unit matrix. That means U^H is the inverse matrix to U , $U^H = U^{-1}$ and then you also have $U^H U = I$:

$$\sum_{\nu=1}^N u_{\nu m}^* u_{\nu n} = \delta_{mn}.$$

Now premultiply the definition of the new orbitals above by U^H ; you get

$$\sum_{\nu=1}^N u_{\nu m}^* \bar{\psi}_{\nu}^s = \sum_{\nu=1}^N \sum_{n=1}^N u_{\nu m}^* u_{\nu n} \psi_n^s$$

but the sum over ν in the right hand side is just δ_{mn} and you get

$$\sum_{\nu=1}^N u_{\nu m}^* \bar{\psi}_{\nu}^s = \sum_{n=1}^N \delta_{mn} \psi_n^s = \psi_m^s.$$

That gives you an expression for the original orbitals in terms of the new ones. For aesthetic reasons, you might just as well renotate ν to μ , the Greek equivalent of m , to get

$$\psi_m^s = \sum_{\mu=1}^N u_{\mu m}^* \bar{\psi}_{\mu}^s.$$

Now plug that into the noncanonical restricted closed-shell Hartree-Fock equations, with equivalent expressions for ψ_n^s using whatever summation variable is still available,

$$\psi_n^s = \sum_{\mu=1}^N u_{\mu n}^* \bar{\psi}_{\mu}^s \quad \psi_n^s = \sum_{\kappa=1}^N u_{\kappa n}^* \bar{\psi}_{\kappa}^s \quad \psi_n^s = \sum_{\lambda=1}^N u_{\lambda n}^* \bar{\psi}_{\lambda}^s$$

and use the reduction formula $UU^H = I$,

$$\sum_{m=1}^N u_{\nu m} u_{\mu m}^* = \delta_{\mu\nu} \quad \sum_{n=1}^N u_{\kappa n} u_{\lambda n}^* = \delta_{\kappa\lambda}$$

premultiplying all by U , i.e. put $\sum_{m=1}^N u_{\nu m}$ before each term. You get

$$h^e \bar{\psi}_{\nu}^s + 2 \sum_{\lambda=1}^N \langle \bar{\psi}_{\lambda}^s | v^{ee} | \bar{\psi}_{\lambda}^s \rangle \bar{\psi}_{\nu}^s - \sum_{\lambda=1}^N \langle \bar{\psi}_{\lambda}^s | v^{ee} | \bar{\psi}_{\nu}^s \rangle \bar{\psi}_{\lambda}^s = \sum_{m=1}^N \sum_{n=1}^N \sum_{\mu=1}^N \frac{1}{2} u_{\nu m} \epsilon_{mn} u_{\mu n}^* \bar{\psi}_{\mu}^s$$

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 value of μ , take $u_{\mu n}^*$ as the μ -th orthonormal eigenvector of Hermitian matrix ϵ_{mn} , calling the eigenvalue $2\epsilon_\mu$. Then the right hand side becomes

$$\sum_{m=1}^I \sum_{\mu=1}^N u_{\nu m} \epsilon_\mu u_{\mu m}^* \bar{\psi}_\mu^s = \sum_{\mu=1}^N \delta_{\mu\nu} \epsilon_\mu \bar{\psi}_\mu^s = \epsilon_\nu \bar{\psi}_\nu^s$$

So, in terms of the new orbitals defined by the requirement that $u_{\mu n}^*$ gives the eigenvectors of ϵ_{mn} , the right hand side simplifies to the canonical one.

Since you no longer care about the old orbitals, you can drop the overlines on the new ones, and revert to sensible roman indices n and \underline{n} instead of the Greek ones ν and λ . You then have the canonical restricted closed-shell Hartree-Fock equations

$$h^e \psi_n^s + 2 \sum_{\underline{n}=1}^{I/2} \langle \psi_{\underline{n}}^s | v^{ee} | \psi_{\underline{n}}^s \rangle \psi_n^s - \sum_{\underline{n}=1}^{I/2} \langle \psi_{\underline{n}}^s | v^{ee} | \psi_n^s \rangle \psi_{\underline{n}}^s = \epsilon_n \psi_n^s$$

(A.84)

that, as noted, assume that the Slater determinant is ordered so that the $I/2$ spin-up orbitals are at the start of it. Note that the left hand side directly provides a Hermitian Fock operator if you identify it as $\mathcal{F}\psi_n^s$; there is no need to involve spin here.

In the unrestricted case, the noncanonical equations are

$$h^e \psi_m^s + \sum_{n=1}^I \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \sum_{n=1}^I \langle \uparrow_n | \uparrow_m \rangle^2 \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s = \sum_{n=1}^I \epsilon_{mn} \psi_n^s$$

In this case the spin-up and spin-down spatial states are not mutually orthonormal, and you want to redefine the group of spin up states and the group of spin down states separately.

The term in linear algebra is that you want to partition your U matrix. What that means is simply that you separate the orbital numbers into two sets. The set of numbers $n \leq N_u$ of spin-up orbitals will be indicated as U, and the set of values $n > N_u$ of spin-down ones by D. So you can partition (separate) the noncanonical equations above into equations for $m \in U$ (meaning m is one of the values in set U):

$$h^e \psi_m^s + \sum_{n \in U} \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s + \sum_{n \in D} \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \sum_{n \in U} \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s = \sum_{n \in U} \epsilon_{mn} \psi_n^s,$$

and equations for $m \in D$

$$h^e \psi_m^s + \sum_{n \in U} \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s + \sum_{n \in D} \langle \psi_n^s | v^{ee} | \psi_n^s \rangle \psi_m^s - \sum_{n \in D} \langle \psi_n^s | v^{ee} | \psi_m^s \rangle \psi_n^s = \sum_{n \in D} \epsilon_{mn} \psi_n^s.$$

In these two 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 sets are allowed to be mutually nonorthogonal.

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 equivalent of the ϵ_{mn} for $m \neq n$, leaving only ϵ_{mm} -equivalent values. Then combine the spin-up and spin-down equations again into a single expression. You get, in terms of revised names,

$$h^e \psi_n^s + \sum_{\underline{n}=1}^I \langle \psi_{\underline{n}}^s | v^{ee} | \psi_{\underline{n}}^s \rangle \psi_n^s - \sum_{\underline{n}=1}^I \langle \uparrow_{\underline{n}} | \downarrow_{\underline{n}} \rangle^2 \langle \psi_{\underline{n}}^s | v^{ee} | \psi_{\underline{n}}^s \rangle \psi_{\underline{n}}^s = \epsilon_n \psi_n^s \quad (\text{A.85})$$

In the restricted open-shell Hartree-Fock method, the partitioning also needs to include the set P of orbitals $n \leq N_p$ whose spatial orbitals appear both with spin-up and spin-down in the Slater determinant. In that case, the procedure above to eliminate the ϵ_{mn} values for $m \neq n$ no longer works, since there are coefficients relating the sets. This (even more) elaborate case will be left to the references that you can find in [32].

Woof.

A.67 Why the Fock operator is Hermitian

To verify that the Fock operator is Hermitian, first note that h^e is Hermitian since it is an Hamiltonian. Next if you form the inner product $\langle \overline{\psi^e \uparrow} | v^{\text{HF}} \psi^s \uparrow \rangle$, the first term in v^{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,

$$- \sum_{\underline{n}=1}^I \left\langle \overline{\psi^e(\vec{r}) \uparrow(S_z)} \middle| \langle \psi_{\underline{n}}^s(\vec{r}) \uparrow_{\underline{n}}(S_{z1}) | v^{ee} | \psi_{\underline{n}}^s(\vec{r}) \uparrow(S_{z1}) \rangle \psi_{\underline{n}}^s(\vec{r}) \uparrow_{\underline{n}}(S_z) \right\rangle$$

and if you take the operator to the other side, you get

$$- \sum_{\underline{n}=1}^I \left\langle \langle \psi_{\underline{n}}^s(\vec{r}) \uparrow_{\underline{n}}(S_z) | v^{ee} | \overline{\psi^e(\vec{r}) \uparrow(S_z)} \rangle \psi_{\underline{n}}^s(\vec{r}) \uparrow_{\underline{n}}(S_{z1}) \middle| \psi^s(\vec{r}) \uparrow(S_{z1}) \right\rangle$$

and writing out these inner products as six-dimensional spatial integrals and sums over spin, you see that they are the same.

A.68 “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., [32, 20, 22] 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 [20] 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 [32, 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*. They then explain: that’s OK; the exchange requirements *do not allow uncorrelated positions*. This really helps an engineer trying to figure out why the “motion” of the two electrons is uncorrelated!

The unrestricted Hartree-Fock solution of the dissociated hydrogen molecule is of this type. Since if one electron is around the left proton, the other is around the right one, and vice versa, many people would call the positions of the electrons strongly correlated. But now we engineers understand that this “does not count,” because an uncorrelated state in which electron 1 is around the left proton for sure and electron 2 is around the right one for sure is not allowed.

Having done so well, the authors offer us no further guidance how we are supposed to figure out whether or not electrons 1 and 2 are of opposite spin if there are more than two electrons. It is true that if the wave function

$$\Psi(\vec{r}_1, \frac{1}{2}\hbar, \vec{r}_2, -\frac{1}{2}\hbar, \vec{r}_3, S_{z3}, \dots)$$

is represented by a single small determinant, (like for helium or lithium, say), it leads to uncorrelated spatial probability distributions for electrons 1 and 2. However, that stops being true as soon as there are at least two spin-up states and two spin-down states. And of course it is again just a symptom of the single-determinant disease, not the disease itself. Not a sliver of evidence is given that the supposed lack of correlation is an important source of the error in Hartree-Fock, let alone the only error.

Koch and Holthausen, [20, 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

$$\frac{1}{2}|\psi_1(\vec{r}_1)|^2|\psi_{rmr}(\vec{r}_2)|^2 + \frac{1}{2}|\psi_{rmr}(\vec{r}_1)|^2|\psi_1(\vec{r}_2)|^2$$

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 ψ_1 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, $|\psi_1(\vec{r}_1)|^2|\psi_{rmr}(\vec{r}_2)|^2$. 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, be antisymmetric with respect to exchange of any two electrons. And the wave function is antisymmetric even if the two terms above are not the same.

Wikipedia, [[18]], Hartree-Fock entry June 2007, lists electron correlation, (defined here vaguely as “effects” arising from the mean-field approximation, i.e. using the same v^{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 than the single-determinant approximation! And again, even if the potential energy *was* computed exactly, instead of using the v^{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 correlation 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.

A.69 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.21 (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.21 (c); it is really the same thing seen in the mirror. In cases like figure A.21 (a), where the electron clouds move towards 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.”

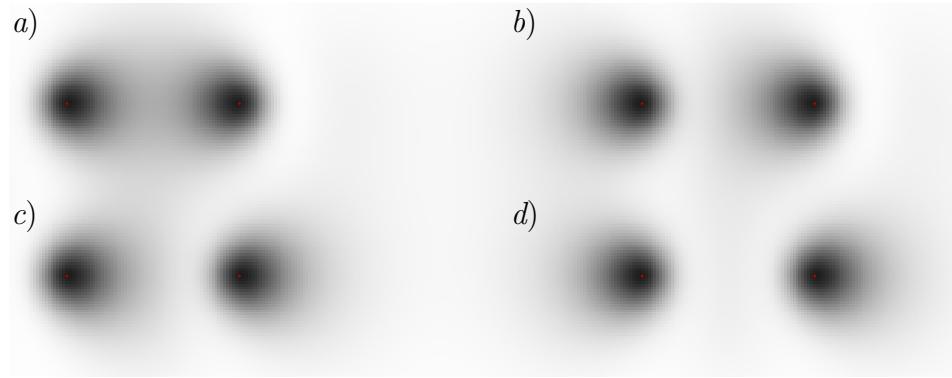


Figure A.21: Possible polarizations of a pair of hydrogen atoms.

Before examining what is wrong with this explanation, first consider what is right. It is perfectly right that figure A.21 (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:

$$V_{lr} = \frac{e^2}{4\pi\epsilon_0} \left(\frac{1}{d} - \frac{1}{r_l} - \frac{1}{r_r} + \frac{1}{r_{lr}} \right)$$

where $e = 1.6 \cdot 10^{-19}$ C is the magnitude of the charges of the protons and electrons, $\epsilon_0 = 8.85 \cdot 10^{-12}$ C²/J m is the permittivity of space, d is the distance between the nuclei, r_l is the distance between the left electron and the right nucleus, r_r the one between the right electron and the left nucleus, and r_{lr} is the distance between the two electrons. If the electrons charges are distributed over space according to densities $n_l(\vec{r}_l)$ and $n_r(\vec{r}_r)$, the classical potential energy is

$$V_{lr} = \frac{e^2}{4\pi\epsilon_0} \int_{\text{all } \vec{r}_l} \int_{\text{all } \vec{r}_r} \left(\frac{1}{d} - \frac{1}{r_l} - \frac{1}{r_r} + \frac{1}{r_{lr}} \right) n_l(\vec{r}_l) n_r(\vec{r}_r) d^3\vec{r}_l d^3\vec{r}_r$$

(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:

$$V_{lr} = \frac{e^2}{4\pi\epsilon_0} \int_{\text{all } \vec{r}_l} \int_{\text{all } \vec{r}_r} \frac{x_l x_r + y_l y_r - 2z_l z_r}{d^3} n_l(\vec{r}_l) n_r(\vec{r}_r) d^3\vec{r}_l d^3\vec{r}_r$$

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.21, the $x_l x_r$ and $y_l y_r$ terms integrate to zero because of odd symmetry. However, for a distribution like in figure A.21 (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

$$\psi(\vec{r}_l, \vec{r}_r) = \psi_{100}(\vec{r}_l)\psi_{100}(\vec{r}_r)$$

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} “2p _{z} ” states, as follows:

$$\psi(\vec{r}_l, \vec{r}_r) = \sqrt{1 - \varepsilon^2}\psi_{100}(\vec{r}_l)\psi_{100}(\vec{r}_r) + \varepsilon\psi_{210}(\vec{r}_l)\psi_{210}(\vec{r}_r).$$

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 $\psi_{210}(\vec{r}_l)\psi_{210}(\vec{r}_r)$ 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, $\psi_{210}(\vec{r}_l)\psi_{210}(\vec{r}_r)$ is negative, and the wave function is reduced.

Now write the expectation value of the energy:

$$\langle E \rangle = \langle \sqrt{1 - \varepsilon^2}\psi_{100}\psi_{100} + \varepsilon\psi_{210}\psi_{210} | H_l + H_r + V_{lr} | \sqrt{1 - \varepsilon^2}\psi_{100}\psi_{100} + \varepsilon\psi_{210}\psi_{210} \rangle$$

where H_l and H_r are the Hamiltonians of the individual electrons and

$$V_{lr} = \frac{e^2}{4\pi\epsilon_0} \frac{x_l x_r + y_l y_r - 2z_l z_r}{d^3}$$

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

$$\langle E \rangle = 2E_1 + 2\varepsilon^2(E_2 - E_1) - 4\varepsilon\sqrt{1 - \varepsilon^2} \frac{e^2}{4\pi\epsilon_0} \frac{1}{d^3} \langle \psi_{100}\psi_{100} | z_l z_r | \psi_{210}\psi_{210} \rangle.$$

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

$$\frac{2}{E_2 - E_1} \left(\frac{e^2}{4\pi\epsilon_0} \langle \psi_{100} | z | \psi_{210} \rangle \right)^2 \frac{1}{d^6}.$$

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 $2p_x$ and $2p_y$ 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:

$$\langle \psi_{100}\psi_{100} | V_{lr} | \psi_{210}\psi_{210} \rangle.$$

That is the energy if both electrons are in the spherically symmetric ψ_{100} ground state if both electrons are in the antisymmetric $2p_z$ state. The savior term is a twilight term, like the ones discussed earlier in chapter 4.3 for chemical bonds. It reflects nature's habit of doing business in terms of an unobservable wave function instead of observable probabilities.

A.70 Ambiguities in the definition of 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 Britannica 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.

A.71 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.

A.72 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 $k_B T$ 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 ^3He 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 $\frac{1}{2}$ neutron in ^3He atoms makes them fermions. (The eventual superfluid transition of ^3He 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 [13, 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, {A.78}.

Tilley and Tilley [33, 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 ^4He .” 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 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, [33, 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. In that case, if the concept of Bose-Einstein condensation is still meaningful for liquid helium, it would presumably imply that normal helium I would condensate at a temperature below the lambda point, and helium II at one above the lambda point.

However, Tilley and Tilley [33, 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.

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 [26]:

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.

An undisputed 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).

A.73 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 4.9 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 4.9 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 4.9 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 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 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:

$$\langle \psi_1\psi_2 | V_{\alpha\beta} | \psi_2\psi_1 \rangle = \int_{\text{all } \vec{r}_1} \int_{\text{all } \vec{r}_2} V_{\alpha\beta} f(\vec{r}_1, \vec{r}_2) f^*(\vec{r}_2, \vec{r}_1) d^3\vec{r}_1 d^3\vec{r}_2$$

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 4.2 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.74 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 [29, p. 282] in 1956 describes ferromagnetism as an interaction between electrons localized at neighboring atoms, Feynman [15, 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 Lett 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)

A.75 Number of system eigenfunctions

This note derives the number of energy eigenfunctions $Q_{\vec{I}}$ for a given set $\vec{I} = (I_1, I_2, I_3, \dots)$ 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 9.1 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

$$I \times (I - 1) \times (I - 2) \times \dots \times (I - I_1 + 1)$$

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

$$I_1! = I_1 \times (I_1 - 1) \times (I_1 - 2) \times \dots \times 3 \times 2 \times 1;$$

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

$$\frac{I \times (I - 1) \times (I - 2) \times \dots \times (I - I_1 + 1)}{I_1 \times (I_1 - 1) \times (I_1 - 2) \times \dots \times 3 \times 2 \times 1}.$$

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:

$$N_1^{I_1} \frac{I \times (I - 1) \times (I - 2) \times \dots \times (I - I_1 + 1)}{I_1 \times (I_1 - 1) \times (I_1 - 2) \times \dots \times 3 \times 2 \times 1}.$$

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 choose 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

$$N_1^{I_1} \frac{I!}{I_1!(I - I_1)!} \times N_2^{I_2} \frac{(I - I_1)!}{I_2!(I - I_1 - I_2)!} \times N_3^{I_3} \frac{(I - I_1 - I_2)!}{I_3!(I - I_1 - I_2 - I_3)!} \times \dots$$

This terminates at the shelf number S beyond which there are no more particles left, when $I - I_1 - I_2 - I_3 - \dots - I_B = 0$. 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

$$Q_I^d = I! \frac{N_1^{I_1}}{I_1!} \times \frac{N_2^{I_2}}{I_2!} \times \frac{N_3^{I_3}}{I_3!} \times \dots$$

Mathematicians like to symbolically write a product of indexed factors like this using the product symbol \prod :

$$Q_I^d = I! \prod_{\text{all } s} \frac{N_s^{I_s}}{I_s!}.$$

It means exactly the same as the written-out product.

Next the eigenfunction count for fermions. Refer now to figure 9.3. 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

$$N_s \times (N_s - 1) \times (N_s - 2) \times \dots \times (N_s - I_s + 1) = \frac{N_s!}{(N_s - I_s)!}$$

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

$$Q_I^f = \frac{N_1!}{(N_1 - I_1)!I_1!} \times \frac{N_2!}{(N_2 - I_2)!I_2!} \times \frac{N_3!}{(N_3 - I_3)!I_3!} \times \dots = \prod_{\text{all } s} \frac{N_s!}{(N_s - I_s)!I_s!}.$$

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 A.22: Schematic of an example boson distribution on a shelf.

Last but not least, the eigenfunction count for bosons. Refer now to figure 9.2. This one is tricky, but a trick solves it. To illustrate the idea, take shelf 2 in figure 9.2 as an example. It is reproduced in condensed form in figure A.22. 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 A.22? 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

$$Q_I^b = \frac{(I_1 + N_1 - 1)!}{I_1!(N_1 - 1)!} \times \frac{(I_2 + N_2 - 1)!}{I_2!(N_2 - 1)!} \times \frac{(I_3 + N_3 - 1)!}{I_3!(N_3 - 1)!} \times \dots = \prod_{\text{all } s} \frac{(I_s + N_s - 1)!}{I_s!(N_s - 1)!}.$$

A.76 The fundamental assumption of quantum 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], [13, p. 1], [17, p. 230], [34, 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 [13, 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

$$\left(\frac{d|c_A|^2}{dt} \right)_{\text{duetoB}} = \frac{1}{\hbar} \Im(c_A^* H_{AB} c_B) \quad \left(\frac{d|c_B|^2}{dt} \right)_{\text{duetoA}} = -\frac{1}{\hbar} \Im(c_A^* H_{AB} c_B)$$

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 [13, p. 8] gives an argument based on time-dependent perturbation theory, subsection 9.10. 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.

A.77 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 subsection 9.3 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 subsection 9.3 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 9.5 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 $2.5I$ 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 $2.4999\dots I$ instead of $2.5I$ exactly. Here typical textbooks will pontificate about the accuracy of your system-energy measurement device. However, this book shudders to con-

template 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 9.6, 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 . . . 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.

A.78 Derivation of 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 $\vec{I} = (I_1, I_2, I_3, \dots)$ 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 note {A.75} into sums: mathematics says that $\ln(ab)$ equals $\ln(a)$ plus $\ln(b)$ for any (positive) a and b .

It will be assumed, following note {A.77}, 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 :

$$\sum_s I_s = I \quad \sum_s I_s E_s^p = E.$$

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, {A.58}, and the correct function whose derivatives must be zero is

$$F = \ln(Q_{\vec{I}}) - \epsilon_1 \left(\sum_s I_s - I \right) - \epsilon_2 \left(\sum_s I_s E_s^p - E \right)$$

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

$$0 = \frac{\partial F}{\partial I_s} \approx \frac{\Delta F}{\Delta I_s} \equiv \frac{F(I_1, I_2, \dots, I_{s-1}, I_s + 1, I_{s+1}, \dots) - F(I_1, I_2, \dots, I_{s-1}, I_s, I_{s+1}, \dots)}{I_s + 1 - I_s}.$$

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, $I_s + \frac{1}{2} \equiv I_{s,\text{best}}$ will be no more than 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 $Q_{\vec{I}}^d$ as found in note {A.75} produces

$$F = \ln(I!) + \sum_s [I_s \ln(N_s) - \ln(I_s!)] - \epsilon_1 \left(\sum_s I_s - I \right) - \epsilon_2 \left(\sum_s I_s E_s^p - E \right)$$

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:

$$\frac{\Delta F}{\Delta I_s} = \ln(N_s) - \ln(I_s + 1) - \epsilon_1 - \epsilon_2 E_s^p = 0$$

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

$$I_{s,\text{best}} = I_s + \frac{1}{2} = \frac{N_s}{e^{\epsilon_2 E_s^p + \epsilon_1}} - \frac{1}{2}$$

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 $\iota^d \equiv I_{s,\text{best}}/N_s$ 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

$$F = \sum_s [\ln(N_s!) - \ln(I_s!) - \ln((N_s - I_s)!)] - \epsilon_1 \left(\sum_s I_s - I \right) - \epsilon_2 \left(\sum_s I_s E_s^p - E \right)$$

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

$$\frac{\Delta F}{\Delta I_s} = -\ln(I_s + 1) + \ln(N_s - I_s) - \epsilon_1 - \epsilon_2 E_s^p = 0$$

which can be solved to give

$$I_{s,\text{best}} = I_s + \frac{1}{2} = \frac{N_s}{e^{\epsilon_2 E_s^p + \epsilon_1} + 1} + \frac{1}{2} \frac{1 - e^{\epsilon_2 E_s^p + \epsilon_1}}{1 + e^{\epsilon_2 E_s^p + \epsilon_1}}$$

The final term, less than half a particle, is again best left away, to ensure that $0 \leq I_{s,\text{best}} \leq N_s$ 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

$$\begin{aligned} F = & \sum_s [\ln((I_s + N_s - 1)!) - \ln(I_s!) - \ln((N_s - 1)!)] \\ & - \epsilon_1 \left(\sum_s I_s - I \right) - \epsilon_2 \left(\sum_s I_s E_s^p - E \right) \end{aligned}$$

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 $-(\epsilon_1 + \epsilon_2 E_s^p)I_s$. This tends to minus infinity if $\epsilon_1 + \epsilon_2 E_s^p$ is positive, so for now assume it is. Then the difference quotient approximation produces

$$\frac{\Delta F}{\Delta I_s} = \ln(I_s + N_s) - \ln(I_s + 1) - \epsilon_1 - \epsilon_2 E_s^p = 0$$

which can be solved to give

$$I_{s,\text{best}} = I_s + \frac{1}{2} = \frac{N_s - 1}{e^{\epsilon_2 E_s^p + \epsilon_1} - 1} - \frac{1}{2}.$$

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 $\epsilon_1 + \epsilon_2 E_s^p$ 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

$$\begin{aligned} F = & \ln(Q_{\vec{I}_A} Q_{\vec{I}_B}) \\ & - \epsilon_{1,A} \left(\sum_{s_A} I_{s_A} - I_A \right) - \epsilon_{1,B} \left(\sum_{s_B} I_{s_B} - I_B \right) \\ & - \epsilon_2 \left(\sum_{s_A} I_{s_A} E_{s_A}^p + \sum_{s_B} I_{s_B} E_{s_B}^p - E \right) \end{aligned}$$

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 $T = 1/\epsilon_2 k_B$, so that $\epsilon_2 = 1/k_B T$, with k_B the Boltzmann constant. The same way, for now the chemical potential μ will simply be defined to be the constant $-\epsilon_1/\epsilon_2$. Subsection 9.14.4 will eventually establish that the temperature defined here is the ideal gas temperature, while note {A.84} 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 $(E_s^p - \mu)/k_B T$ 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

$$I_s = \frac{N_s - 1}{e^{(E_s^p - \mu)/k_B T} - 1}$$

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

$$\begin{aligned} I_1 &= \frac{(N_1 - 1)k_B T}{E_1^p - \mu} \\ I_2 &= \frac{(N_2 - 1)k_B T}{E_1^p - \mu + (E_2^p - E_1^p)} \\ I_3 &= \frac{(N_3 - 1)k_B T}{E_1^p - \mu + (E_2^p - E_1^p) + (E_3^p - E_2^p)} \\ &\vdots \end{aligned}$$

For I_2 to also be a finite fraction of the total number of particles, $E_2^p - E_1^p$ must be similarly small as $E_1^p - \mu$. 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 $\sum_{s=2}^{\infty} \text{const}/s\Delta E$, (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_1^p than E_1^p is to the next energy level E_2^p , 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_1^p . (Specific data for particles in a box is given in section 9.14.1. The entire story may of course need to be modified in the presence of confinement, compare chapter 5.12.)

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:

$$\frac{\Delta F}{\Delta I_1} = \ln(I_1 + 1) - \ln(I_1 + 1) - \epsilon_1 - \epsilon_2 E_1^p = 0$$

implies that $\epsilon_1 - \epsilon_2 E_1^p = 0$. In other words, μ is equal to the ground state energy E_1^p 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 {A.84}.

A.79 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, \dots with a given total energy E . Call the combined system the collective. Following the same reasoning as in note {A.78} for two systems, the thermodynamically stable equilibrium state has shelf occupation numbers of the subsystems satisfying

$$\begin{aligned}\frac{\partial \ln Q_{\vec{I}_A}}{\partial I_{s_A}} - \epsilon_{1,A} - \epsilon_2 E_{s_A}^p &= 0 \\ \frac{\partial \ln Q_{\vec{I}_B}}{\partial I_{s_B}} - \epsilon_{1,B} - \epsilon_2 E_{s_B}^p &= 0 \\ &\dots\end{aligned}$$

where ϵ_2 is a shorthand for $1/k_B T$.

An individual system, take A as the example, no longer has an individual energy that is 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 $\psi_{q_A}^S$ with somewhat different energy $E_{q_A}^S$ can no longer be excluded with certainty. However, still assume, following the fundamental assumption of quantum statistics, {A.76}, 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 $\psi_{q_A}^S$ will be assumed to depend only on its energy $E_{q_A}^S$:

$$P_{q_A} = P(E_{q_A}^S).$$

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 $\psi_{q_A}^S$ for a given set of shelf numbers \vec{I}_A have the same system energy $E_{\vec{I}_A}^S = \sum_{s_A} I_{s_A} E_{s_A}^P$. 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:

$$P_{\vec{I}_A} = Q_{\vec{I}_A} P(E_{\vec{I}_A}^S).$$

Mathematically, the function whose partial derivatives must be zero to find the most probable shelf numbers is

$$F = \ln(P_{\vec{I}_A}) - \epsilon_{1,A} \left(\sum_{s_A} I_{s_A} - I_A \right).$$

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 $P_{\vec{I}_A} = Q_{\vec{I}_A} P(E_{\vec{I}_A}^S)$, taking apart the logarithm, and differentiating, produces

$$\frac{\partial \ln Q_{\vec{I}_A}}{\partial I_{s_A}} + \frac{d \ln(P)}{d E_{\vec{I}_A}^S} E_{s_A}^P - \epsilon_{1,A} = 0$$

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. $-1/k_B T$. 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 $d\ln P/dE_{I_A}^S = -1/k_B T$ at least in the vicinity of the most probable energy $E_{I_A}^S$. Hence in the vicinity of that energy

$$P(E_A^S) = \frac{1}{Z_A} e^{-E_A^S/k_B T}$$

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

$$P_{AB} = \frac{1}{Z_A Z_B} e^{-(E_A^S + E_B^S)/k_B T}.$$

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

$$Z_A Z_B = \sum_{q_A} \sum_{q_B} e^{-(E_{q_A}^S + E_{q_B}^S)/k_B T} = Z_{AB}.$$

A.80 Analysis of the ideal gas Carnot cycle

Refer to figure A.23 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. Section 9.14 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 molecular mass.

The differential version of the first law, energy conservation, (9.11), says that

$$dE = \delta Q - P dV$$

or getting rid of internal energy and pressure using the given expressions,

$$mIC_v dT = \delta Q - mIRT \frac{dV}{V}.$$

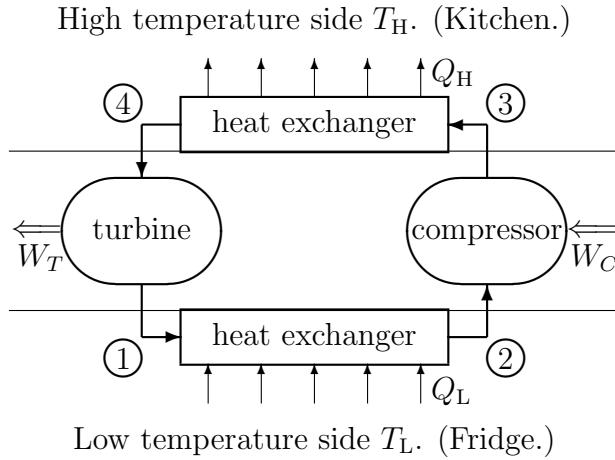


Figure A.23: Schematic of the Carnot refrigeration cycle.

Now for the transitions through the heat exchangers, from 1 to 2 or from 3 to 4 in figure A.23, the temperature is approximated to be constant. The first law above can then be integrated to give the heat added to the substance as:

$$Q_L = mIRT_L (\ln V_2 - \ln V_1) \quad Q_H = -mIRT_H (\ln V_4 - \ln V_3).$$

Remember that unlike Q_L , Q_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

$$mIC_v (\ln T_H - \ln T_L) = -mIR (\ln V_3 - \ln V_2)$$

$$mIC_v (\ln T_L - \ln T_H) = -mIR (\ln V_1 - \ln V_4)$$

Adding these two expressions shows that

$$\ln V_3 - \ln V_2 + \ln V_1 - \ln V_4 = 0 \implies \ln V_3 - \ln V_4 = \ln V_2 - \ln V_1$$

and plugging that into the expressions for the exchanged heats shows that $Q_H/T_H = Q_L/T_L$.

A.81 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 non-messy 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 non-messy 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, (9.59). 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!

A.82 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

$$\int \frac{C}{T} dT$$

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,

$$S = -k_B \sum P_q \ln(P_q)$$

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 4.7. That produces an entropy of only about $-k_B \ln(2/I^2)$. 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

$$s_{\text{h.s.}}(I, T, P) = \frac{IT}{1 + IT}$$

This system is consistent with the expression for the entropy 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.83 Checks on the expression for entropy

According to the microscopic definition, the differential of the entropy S should be

$$dS = -k_B d \left[\sum_q P_q \ln P_q \right]$$

where the sum is over all system energy eigenfunctions ψ_q^S and P_q is their probability. The differential can be simplified to

$$dS = -k_B \sum_q [\ln P_q + 1] dP_q = -k_B \sum_q \ln P_q dP_q,$$

the latter equality since the sum of the probabilities is always one, so $\sum_q dP_q = 0$.

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 $e^{-E_q^S/k_B T}/Z$ where E_q^S is the energy of system eigenfunction ψ_q^S . It simplifies the microscopic differential of the entropy to

$$dS = -k_B \sum_q \left[-\frac{E_q^S}{k_B T} - \ln Z \right] dP_q = -k_B \sum_q \left[-\frac{E_q^S}{k_B T} \right] dP_q = \frac{1}{T} \sum_q E_q^S dP_q, \quad (\text{A.86})$$

the second inequality since Z is a constant in the summation and $\sum_q dP_q = 0$.

The macroscopic expression for the differential of entropy is given by (9.18),

$$dS = \frac{\delta Q}{T}.$$

Substituting in the differential first law (9.11),

$$dS = \frac{1}{T} dE + \frac{1}{T} P dV$$

and plugging into that the definitions of E and P ,

$$dS = \frac{1}{T} d \left[\sum_q P_q E_q^S \right] - \frac{1}{T} \left[\sum_q P_q \frac{dE_q^S}{dV} \right] dV$$

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 (A.86). 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

$$F = -k_B \sum_q P_q \ln P_q - \epsilon \left(\sum_q P_q - 1 \right)$$

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, {A.58}. Taking derivatives produces

$$-k_B \ln(P_q) - k_B - \epsilon = 0$$

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

$$F = -k_B \sum_q P_q \ln P_q - \epsilon_1 \left(\sum_q P_q - 1 \right) - \epsilon_2 \left(\sum_q E_q^S P_q - E \right)$$

and the derivatives become

$$-k_B \ln(P_q) - k_B - \epsilon_1 - \epsilon_2 E_q^S = 0$$

which can be solved to show that

$$P_q = C_1 e^{-E_q^S/C_2}$$

with C_1 and C_2 constants. The requirement to conform with the given definition of temperature identifies C_2 as $k_B T$ 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

$$\begin{aligned} F &= -k_B \sum_{q_A} \sum_{q_B} P_{q_A} P_{q_B} \ln(P_{q_A} P_{q_B}) \\ &\quad - \epsilon_{1,A} (\sum_{q_A} P_{q_A} - 1) - \epsilon_{1,B} (\sum_{q_B} P_{q_B} - 1) \\ &\quad - \epsilon_2 (\sum_{q_A} P_{q_A} E_{q_A}^S + \sum_{q_B} P_{q_B} E_{q_B}^S - E) \end{aligned}$$

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

$$\begin{aligned} F &= -k_B \sum_{q_A} P_{q_A} \ln(P_{q_A}) - k_B \sum_{q_B} P_{q_B} \ln(P_{q_B}) \\ &\quad - \epsilon_{1,A} (\sum_{q_A} P_{q_A} - 1) - \epsilon_{1,B} (\sum_{q_B} P_{q_B} - 1) \\ &\quad - \epsilon_2 (\sum_{q_A} P_{q_A} E_{q_A}^S + \sum_{q_B} P_{q_B} E_{q_B}^S - E) \end{aligned}$$

Differentiation now produces

$$\begin{aligned} -k_B \ln(P_{q_A}) - k_B - \epsilon_{1,A} - \epsilon_2 E_{q_A}^S &= 0 \\ -k_B \ln(P_{q_B}) - k_B - \epsilon_{1,B} - \epsilon_2 E_{q_B}^S &= 0 \end{aligned}$$

which produces $P_{q_A} = C_{1,A} e^{-E_{q_A}^S/C_2}$ and $P_{q_B} = C_{1,B} e^{-E_{q_B}^S/C_2}$ and the common constant C_2 then implies that the two systems have the same temperature.

A.84 Chemical potential and distribution functions

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 non engineering author should be allowed one mistake.

The derivations of the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein distributions given previously, {A.78} and {A.79}, 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 ψ_n^p . 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 ψ_q^S has in single-particle state ψ_n^p , and P_q the probability of the eigenfunction. Since

thermal equilibrium is assumed, the canonical probability value $e^{-E_q^S/k_B T}/Z$ can be substituted for P_q . Then, if the energy E_q^S is written as the sum of the ones of the single particle states times the number of particles in that state, it gives:

$$\iota_n = \frac{1}{Z} \sum_q i_n e^{-(i_1 E_1^P + i_2 E_2^P + \dots + i_{n-1} E_{n-1}^P + i_n E_n^P + i_{n+1} E_{n+1}^P + \dots)/k_B T}.$$

Note that i_n is the occupation number of single-particle state ψ_n^P , 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 $\vec{i} = |i_1, i_2, \dots\rangle$, there is exactly *one* system energy eigenfunction. Compare figures 9.2 and 9.3: 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 9.1, because for them the numbers on the particles can still vary for given occupation numbers, but as noted in subsection 9.11, 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:

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1} \sum_{i_2} \dots \sum_{i_{n-1}} \sum_{i_n} \sum_{i_{n+1}} \dots}_{i_1 + i_2 + \dots + i_{n-1} + i_n + i_{n+1} + \dots = I} i_n e^{-(i_1 E_1^P + i_2 E_2^P + \dots + i_{n-1} E_{n-1}^P + i_n E_n^P + i_{n+1} E_{n+1}^P + \dots)/k_B T}$$

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 :

$$\iota_n = \frac{1}{Z} \underbrace{\sum_{i_1=0}^I \sum_{i_2=0}^I \dots \sum_{i_{n-1}=0}^I \sum_{i_n=0}^I \sum_{i_{n+1}=0}^I \dots}_{i_1 + i_2 + \dots + i_{n-1} + i_n + i_{n+1} + \dots = I} i_n e^{-(i_1 E_1^P + i_2 E_2^P + \dots + i_{n-1} E_{n-1}^P + i_n E_n^P + i_{n+1} E_{n+1}^P + \dots)/k_B T}$$

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

$$\begin{aligned} \iota_n &= \frac{1}{Z} \underbrace{\sum_{i_1=0}^I \dots \sum_{i_{n-1}=0}^I \sum_{i'_n=0}^{I-1} \sum_{i_{n+1}=0}^I \dots}_{i_1+\dots+i_{n-1}+i'_n+i_{n+1}+\dots=I-1} \\ &\quad (1 + i'_n) e^{-(i_1 E_1^p + \dots + i_{n-1} E_{n-1}^p + E_n^p + i'_n E_n^p + i_{n+1} E_{n+1}^p + \dots)/k_B T} \end{aligned}$$

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,

$$\begin{aligned} \iota_n &= \frac{1}{Z} e^{-E_n^p/k_B T} \underbrace{\sum_{i_1=0}^{I-1} \dots \sum_{i_{n-1}=0}^{I-1} \sum_{i_n=0}^{I-1} \sum_{i_{n+1}=0}^{I-1} \dots}_{i_1+\dots+i_{n-1}+i_n+i_{n+1}+\dots=I-1} \\ &\quad (1 + i_n) e^{-(i_1 E_1^p + \dots + i_{n-1} E_{n-1}^p + i_n E_n^p + i_{n+1} E_{n+1}^p + \dots)/k_B T} \end{aligned}$$

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

$$\iota_n = \frac{1}{Z} e^{-E_n^p/k_B T} Z^- [1 + \iota_n^-]$$

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:

$$\iota_n = 1 / \left[\frac{Z}{Z^-} e^{E_n^p/k_B T} - 1 \right]$$

The final formula is the Bose-Einstein distribution with

$$e^{-\mu/k_B T} = \frac{Z}{Z^-}$$

Solve for μ :

$$\mu = -k_B T \ln \left(\frac{Z}{Z^-} \right) = \frac{-k_B T \ln(Z) + k_B T \ln(Z^-)}{I - (I - 1)}$$

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 (9.39) 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 $I_A = 6.0221 \cdot 10^{26}$ 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:

$$\nu_n = \frac{1}{Z} \sum_{i_1=0}^1 \dots \underbrace{\sum_{i_{n-1}=0}^1 \sum_{i_n=0}^1 \sum_{i_{n+1}=0}^1 \dots}_{i_1+\dots+i_{n-1}+i_n+i_{n+1}+\dots=I} i_n e^{-(i_1 E_1^p + \dots + i_{n-1} E_{n-1}^p + i_n E_n^p + i_{n+1} E_{n+1}^p + \dots)/k_B T}$$

Again, all terms with $i_n = 0$ are zero, so you can set $i_n = 1 + i'_n$ and get

$$\nu_n = \frac{1}{Z} \sum_{i_1=0}^1 \dots \underbrace{\sum_{i_{n-1}=0}^1 \sum_{i'_n=0}^0 \sum_{i_{n+1}=0}^1 \dots}_{i_1+\dots+i_{n-1}+i'_n+i_{n+1}+\dots=I-1} (1 + i'_n) e^{-(i_1 E_1^p + \dots + i_{n-1} E_{n-1}^p + E_n^p + i'_n E_n^p + i_{n+1} E_{n+1}^p + \dots)/k_B T}$$

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 $Ze^{E_1^p/k_B T}/Z^-$ 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.

A.85 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

$$\frac{I}{\mathcal{V}} = \int_0^\infty \iota^f \mathcal{D} dE^p \quad \frac{E}{\mathcal{V}} = \int_0^\infty E^p \iota^f \mathcal{D} dE^p$$

where the Fermi-Dirac distribution and the density of states are:

$$\iota^f = \frac{1}{e^{(E^p - \mu)/k_B T} + 1} \quad \mathcal{D} = \frac{n_s}{4\pi^2} \left(\frac{2m}{\hbar^2}\right)^{3/2} \sqrt{E^p}$$

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

$$u = \frac{E^p}{k_B T} \quad u_0 = \frac{\mu}{k_B T}$$

to give

$$\begin{aligned} \frac{I}{\mathcal{V}} &= \frac{n_s}{4\pi^2} \left(\frac{2m}{\hbar^2}\right)^{3/2} \mu^{3/2} \int_{u=0}^\infty \frac{(u/u_0)^{1/2}}{e^{u-u_0} + 1} d(u/u_0) \\ \frac{E}{\mathcal{V}} &= \frac{n_s}{4\pi^2} \left(\frac{2m}{\hbar^2}\right)^{3/2} \mu^{5/2} \int_{u=0}^\infty \frac{(u/u_0)^{3/2}}{e^{u-u_0} + 1} d(u/u_0) \end{aligned}$$

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

$$\int_{-1}^0 \sqrt{1+v} dv - \int_{-1}^0 \frac{\sqrt{1+v} e^{u_0 v}}{e^{u_0 v} + 1} dv + \int_0^\infty \frac{\sqrt{1+v} dv}{e^{u_0 v} + 1}$$

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

$$\int_{-1}^0 \sqrt{1+v} dv + \int_0^1 \frac{(\sqrt{1+v} - \sqrt{1-v}) dv}{e^{u_0 v} + 1} + \int_1^\infty \frac{\sqrt{1+v} dv}{e^{u_0 v} + 1}$$

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 $\int_0^\infty x dx/(e^x + 1) = \pi^2/12$, see [28, 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

$$\mu = E_F^p - \frac{\pi^2}{12} \left(\frac{k_B T}{E_F^p} \right)^2 E_F^p + \dots \quad E_F^p = \left(\frac{6\pi^2}{n_s} \right)^{2/3} \frac{\hbar^2}{2m} \left(\frac{I}{V} \right)^{2/3} \quad (\text{A.87})$$

This used the approximations that $\mu \approx E_F^p$ and u_0^{-2} is small, so

$$u_0^{-2} = \left(\frac{k_B T}{\mu} \right)^2 \approx \left(\frac{k_B T}{E_F^p} \right)^2 \quad \left(1 + \frac{\pi^2}{8} u_0^{-2} \right)^{-2/3} \approx 1 - \frac{2}{3} \frac{\pi^2}{8} u_0^{-2}$$

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

$$\frac{E}{I} = E_{\text{ave}}^p = \frac{3}{5} E_F^p + \frac{\pi^2}{4} \left(\frac{k_B T}{E_F^p} \right)^2 E_F^p + \dots \quad (\text{A.88})$$

To get the specific heat at constant volume, divide by m and differentiate with respect to temperature:

$$C_v = \frac{\pi^2}{2} \frac{k_B T}{E_F^p} \frac{k_B}{m} + \dots$$

A.86 Physics of the fundamental commutation relations

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.

Chapter 6.2 explains that the angular momentum operators correspond to small rotations of the axis system through space. So, the commutator $[\hat{L}_x, \hat{L}_y]$ 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, for position coordinates this difference is equivalent to 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.

To verify the stated effect of small rotations around the x and y -axes requires a bit of linear algebra. If you never had a course in it, you will want to skip the rest of this note.

The matrices that describe rotations around the x and y -axes are:

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & \sin \alpha \\ 0 & -\sin \alpha & \cos \alpha \end{pmatrix} \quad R_y = \begin{pmatrix} \cos \beta & 0 & -\sin \beta \\ 0 & 1 & 0 \\ \sin \beta & 0 & \cos \beta \end{pmatrix}$$

Dropping the factors \hbar/i for conciseness, the commutator writes out in terms of the rotation matrices as

$$\hat{L}_x \hat{L}_y - \hat{L}_y \hat{L}_x = \frac{R_x - I}{\alpha} \frac{R_y - I}{\beta} - \frac{R_y - I}{\beta} \frac{R_x - I}{\alpha} = \frac{R_x R_y - R_y R_x}{\alpha \beta}$$

in which α and β are assumed to be infinitesimally small. Note that this is just the difference between rotations around the x and y -axes executed in different order.

Substituting in the rotation matrices gives

$$\begin{aligned} \hat{L}_x \hat{L}_y - \hat{L}_y \hat{L}_x = \\ \frac{1}{\alpha \beta} \begin{pmatrix} 0 & -\sin \alpha \sin \beta & -\sin \beta (1 - \cos \alpha) \\ \sin \alpha \sin \beta & 0 & -\sin \alpha (1 - \cos \beta) \\ -\sin \beta (1 - \cos \alpha) & -\sin \alpha (1 - \cos \beta) & 0 \end{pmatrix} \end{aligned}$$

On the other hand, the change in position caused by an infinitesimal rotation

γ around the z -axis is

$$\hat{L}_z = \frac{R_z - I}{\gamma} = \frac{1}{\gamma} \begin{pmatrix} \cos \gamma - 1 & \sin \gamma & 0 \\ -\sin \gamma & \cos \gamma - 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

From Taylor series expansion with respect to the infinitesimal angles it is seen that this equals minus the commutator above. And that produces the fundamental commutation relationship in terms of the angular momenta. Just put the $(\hbar/i)^2$ back in.

Interestingly enough, the equality does not depend on the relative amounts of rotation α and β around x and y . They do not need to be equal, just small.

A.87 Multiple angular momentum components

Suppose that an eigenstate, call it $|m\rangle$, of \hat{L}_z is also an eigenstate of \hat{L}_x . Then $[\hat{L}_z, \hat{L}_x]|m\rangle$ must be zero, and the commutator relations say that this is equivalent to $\hat{L}_y|m\rangle = 0$, which makes $|m\rangle$ also an eigenvector of \hat{L}_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 $[\hat{L}_x, \hat{L}_y]$ and $[\hat{L}_y, \hat{L}_z]$ 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.

A.88 Components of vectors are less than the total vector

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 $|l\ m\rangle$ 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 $|l\ m\rangle$ states are not eigenstates of \hat{L}_x and \hat{L}_y , the expectation values of square Hermitian operators such as \hat{L}_x^2 and \hat{L}_y^2 is always positive anyway (as can be seen from writing it out in terms of the eigenstates of them.)

A.89 The spherical harmonics with ladder operators

One application of ladder operators is to find the spherical harmonics, which as noted in chapter 3.1.3 is not an easy problem. To do it with ladder operators, show that

$$\boxed{\hat{L}_x = \frac{\hbar}{i} \left(-\sin \phi \frac{\partial}{\partial \theta} - \frac{\cos \theta \cos \phi}{\sin \theta} \frac{\partial}{\partial \phi} \right) \quad \hat{L}_y = \frac{\hbar}{i} \left(\cos \phi \frac{\partial}{\partial \theta} - \frac{\cos \theta \sin \phi}{\sin \theta} \frac{\partial}{\partial \phi} \right)} \quad (\text{A.89})$$

then that

$$\boxed{L^+ = \hbar e^{i\phi} \left(\frac{\partial}{\partial \theta} + i \frac{\cos \theta}{\sin \theta} \frac{\partial}{\partial \phi} \right) \quad L^- = \hbar e^{-i\phi} \left(-\frac{\partial}{\partial \theta} + i \frac{\cos \theta}{\sin \theta} \frac{\partial}{\partial \phi} \right)} \quad (\text{A.90})$$

Note that the spherical harmonics are of the form $Y_l^m = e^{im\phi} \Theta_l^m(\theta)$, so

$$L^+ Y_l^m = \hbar e^{i(m+1)\phi} \sin^m \theta \frac{d(\Theta_l^m / \sin^m \theta)}{d\theta}$$

$$L^- Y_l^m = -\hbar e^{i(m-1)\phi} \frac{1}{\sin^m \theta} \frac{d(\Theta_l^m \sin^m \theta)}{d\theta}$$

Find the Y_l^l harmonic from $\hat{L}^+ Y_l^l = 0$, then apply \hat{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,

$$H^+ = -i\hat{p} + m\omega\hat{x} \quad H^- = i\hat{p} + m\omega\hat{x}$$

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 [17, pp. 42-47] for details. An equivalent derivation is given in chapter 12.2.2 based on quantum field theory.

A.90 Why angular momenta components can be added

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 chapter 6.2.

A.91 Why the 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.

A.92 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 \hat{L}_{ab}^- a few times on the top of the ladder $|l\ l\rangle_a |l\ l\rangle_b$. Every contribution to the coefficient of a state $|l\ m\rangle_a |l\ m\rangle_b$ comes from applying \hat{L}_a^- for $l_a - m_a$ times and \hat{L}_b^- for $l_b - m_b$ times, so all contributions have compatible square roots. \hat{L}_{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 $\hat{L}^+ \hat{L}^-$ and $\hat{L}^- \hat{L}^+$ 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.

A.93 Machine language version of the Clebsch-Gordan tables

The usual “machine language” form of the tables leaves out the a , b , and ab identifiers, the $l_a =$ and $l_b =$ clarifications from the header, and all square root signs, the l values of particles a and b from the kets, and all ket terminator bars

and brackets, but combines the two m -values with missing l values together in a frame to resemble an lm 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.

A.94 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 \vec{L}_a , not with \vec{L}_b , but with an arbitrary multiple λ of \vec{L}_b :

$$\langle (\vec{L}_a + \lambda \vec{L}_b)^2 \rangle = \langle (L_{x,a} + \lambda L_{x,b})^2 \rangle + \langle (L_{y,a} + \lambda L_{y,b})^2 \rangle + \langle (L_{z,a} + \lambda L_{z,b})^2 \rangle$$

For $\lambda = 1$ this produces the expectation value of $(\vec{L}_a + \vec{L}_b)^2$, for $\lambda = -1$, the one for $(\vec{L}_a - \vec{L}_b)^2$. 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

$$\langle (\vec{L}_a + \lambda \vec{L}_b)^2 \rangle = L_a^2 + 2M\lambda + L_b^2\lambda^2$$

where $L_a \equiv \sqrt{\langle L_{xa}^2 + L_{ya}^2 + L_{za}^2 \rangle}$, $L_b \equiv \sqrt{\langle L_{xb}^2 + L_{yb}^2 + L_{zb}^2 \rangle}$, 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:

$$M^2 - L_a^2 L_b^2 \leq 0$$

which means, taking square roots,

$$-L_a L_b \leq M \leq L_a L_b$$

and so

$$L_a^2 - 2L_a L_b + L_b^2 \leq \langle (\vec{L}_a + \vec{L}_b)^2 \rangle \leq L_a^2 + 2L_a L_b + L_b^2$$

or

$$|L_a - L_b|^2 \leq \langle (\vec{L}_a + \vec{L}_b)^2 \rangle \leq |L_a + L_b|^2$$

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 $|l_a - l_b| \leq l_{ab} \leq l_a + l_b$ can be derived from it if a unique value for l_{ab} exists; the key is to recognize that $L = l + \delta$ where δ is an increasing function of l that stays below $1/2$, and the l values must be half integers. This derivation is not as elegant as using the ladder operators, but the result is the same.

A.95 Momentum of shells

Table 10.1 was originally taken from [23], 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. So table 10.1 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 A.1.

Deducing the table using Clebsch-Gordan coefficients would be a messy exercise indeed. A simpler procedure, [21], 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 $j^P = 5/2$. Then the possibilities for the single-particle angular momentum in the z -direction are $m^P = 5/2, 3/2, 1/2, -1/2, -3/2$, and $-5/2$. So there are 6 different one particle states, and these will give rise to $6!/3!(6-3)! = 20$ different antisymmetric states for 3 particles, chapter 4.7.

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 $j^P = 5/2$ single-particle states. First note that m is the sum of the three m^P values of the individual particles. Next, the highest that m^P can be is $5/2$, but the fermions cannot all three be in the same $m^P = 5/2$ state, only one can. Three fermions need three different states, so the highest the combined m can be is $5/2 + 3/2 + 1/2$. This triplet of values of m^P gives exactly one antisymmetric combination of states with $m = 9/2$. (There is only one Slater determinant for three different given states, chapter 4.7). Since the combined angular momentum of this state in any arbitrary direction can never be observed

		possible combined angular momentum j																	
j^p	I	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$	$\frac{9}{2}$	$\frac{11}{2}$	$\frac{13}{2}$	$\frac{15}{2}$	$\frac{17}{2}$	$\frac{19}{2}$	$\frac{21}{2}$	$\frac{23}{2}$	$\frac{25}{2}$	$\frac{27}{2}$	$\frac{29}{2}$	$\frac{31}{2}$	$\frac{33}{2}$	$\frac{35}{2}$
		$\frac{37}{2}$	$\frac{39}{2}$	$\frac{41}{2}$	$\frac{43}{2}$	$\frac{45}{2}$	$\frac{47}{2}$	$\frac{49}{2}$	$\frac{51}{2}$	$\frac{53}{2}$	$\frac{55}{2}$	$\frac{57}{2}$	$\frac{59}{2}$	$\frac{61}{2}$	$\frac{63}{2}$				
$\frac{13}{2}$	1																		
	3	1	1	1	2	2	2	2	2	1	2	1	1	1	1	1	1	1	
	5	1	3	5	5	7	7	8	8	8	7	8	6	6	5	5	3	3	
		2	1	1	1												2		
	7	3	4	7	9	10	11	13	12	13	12	12	10	11	8	8	6	5	
		4	2	2	1	1		1										4	
$\frac{15}{2}$	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	
		5	5	3	3	2	2	1	1		1							7	
	7	4	10	13	17	21	24	25	29	28	29	29	29	26	27	23	22	18	
		14	14	10	9	7	6	4	4	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
		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																
	6	4	1	7	5	11	7	13	9	13	10	12	8	11	7	8	5	6	3	
		2	2	1	1	1													4	
$\frac{15}{2}$	2	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	
		1	2	1	1	1													3	
	6	6	2	11	9	17	13	22	17	23	19	24	18	23	17	19	15	16	11	
		8	9	6	6	3	4	2	2	1	1	1		1					13	
	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 A.1: Additional combined angular momentum values.

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 $\frac{9}{2}, \frac{7}{2}, \dots, -\frac{9}{2}$.

Next consider what combinations have $m = \frac{7}{2}$. The only combination of three different m^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.

A.96 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 chapter 6.2. 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 $e^{im\varphi}$ 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 (1.5) for the exponential, the eigenstates change by a factor

$$\cos\left(\frac{1}{2}\varphi\right) \pm i \sin\left(\frac{1}{2}\varphi\right)$$

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 \hat{L}_n , which multiplies the angular momentum eigenstates by $\pm\frac{1}{2}\hbar$; it is seen to be $2i \sin\left(\frac{1}{2}\varphi\right) \hat{L}_n/\hbar$. So the operator that describes rotation of the coordinate system over an angle φ around the n -axis is

$R_{n,\varphi} = \cos\left(\frac{1}{2}\varphi\right) + i \sin\left(\frac{1}{2}\varphi\right) \frac{2}{\hbar} \hat{L}_n$

(A.91)

Further, in terms of the x , y , and z angular momentum operators, the angular momentum in the n -direction is

$$\hat{L}_n = n_x \hat{L}_x + n_y \hat{L}_y + n_z \hat{L}_z$$

If you put it in terms of the Pauli spin matrices, \hbar drops out:

$$R_{n,\varphi} = \cos\left(\frac{1}{2}\varphi\right) + i \sin\left(\frac{1}{2}\varphi\right) (n_x \sigma_x + n_y \sigma_y + n_z \sigma_z)$$

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 {A.97}.

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 non positive ladder factors would mess up various formulae. The negative spin states found by rotation go out of the window. Bye, bye.

A.97 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.)

A.98 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,

$$\sqrt{(m_0 c^2)^2 + \sum_{i=1}^3 (\hat{p}_i c)^2} = \alpha_0 m_0 c^2 + \sum_{i=1}^3 \alpha_i \hat{p}_i c,$$

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:

$$\alpha_i^2 = 1 \text{ for every } i \quad \alpha_i\alpha_j + \alpha_j\alpha_i = 0 \text{ for } i \neq j \quad (\text{A.92})$$

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

$$\alpha_1 = \begin{pmatrix} 0 & \sigma_1^H \\ \sigma_1 & 0 \end{pmatrix} \quad \alpha_2 = \begin{pmatrix} 0 & \sigma_2^H \\ \sigma_2 & 0 \end{pmatrix} \quad \alpha_3 = \begin{pmatrix} 0 & \sigma_3^H \\ \sigma_3 & 0 \end{pmatrix}.$$

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 (A.92), you see that

$$\alpha_i\alpha_i = 1 \implies \sigma_i^H\sigma_i = \sigma_i\sigma_i^H = 1$$

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 0 \implies \sigma_i^H\sigma_j + \sigma_j^H\sigma_i = \sigma_i\sigma_j^H + \sigma_j\sigma_i^H = 0.$$

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 $e^{i\phi_i}$ for some real angle ϕ_i . The second condition is then according to the Euler formula (1.5) 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

$$\sigma_1 = \begin{pmatrix} 0 & \tau_1^H \\ \tau_1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & \tau_2^H \\ \tau_2 & 0 \end{pmatrix}$$

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 $\tau_1 = e^{i\phi_1}$ and then $\tau_2 = e^{i(\phi_1+\pi/2)}$ or $e^{i(\phi_1-\pi/2)}$. 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 4-dimensional wave function vector takes the form (a_1, a_2, a_3, a_4) , define $\bar{a}_1 = e^{i\phi_1/2}a_1$, $\bar{a}_2 = e^{-i\phi_1/2}a_2$ and similar for \bar{a}_3 and \bar{a}_4 .

In that case, the final cleaned-up σ matrices are

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad (\text{A.93})$$

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 section 10.1.9.

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 $e^{i\pi/2}$, to get the normalized coefficients, you must include an additional factor $e^{i\pi/4}$ in \bar{a}_1 , which by the fundamental definition of angular momentum discussed in chapter 6.2 means that it describes a state with angular momentum $1/2\hbar$. Similarly a_3 corresponds to a state with angular momentum $1/2\hbar$ and a_2 and a_4 to ones with $-1/2\hbar$.

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 εp is a small number in the non-relativistic 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 $1/2\hbar$ state and a spin $-1/2\hbar$ state for rotation around the momentum vector.

A.99 Electromagnetic evolution of expectation values

The purpose of this note is to identify the two commutators of subsection 10.3; 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 3.4.4.

The Hamiltonian is

$$H = \frac{1}{2m} (\hat{\vec{p}} - q\vec{A}) \cdot (\hat{\vec{p}} - q\vec{A}) + q\varphi = \frac{1}{2m} \sum_{j=1}^3 (\hat{p}_j - qA_j)^2 + q\varphi$$

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 6.1.7 given by

$$\frac{d\langle r_i \rangle}{dt} = \left\langle \frac{i}{\hbar} [H, r_i] \right\rangle$$

so you need the commutator

$$[H, r_i] = \left[\frac{1}{2m} \sum_{j=1}^3 (\hat{p}_j - qA_j)^2 + q\varphi, r_i \right]$$

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 $\hat{p}_j - qA_j$ comes out at its own side of the commutator, to give

$$[H, r_i] = \frac{1}{2m} \sum_{j=1}^3 \left\{ (\hat{p}_j - qA_j) [\hat{p}_j - qA_j, r_i] + [\hat{p}_j - qA_j, r_i] (\hat{p}_j - qA_j) \right\}$$

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 \hat{p}_i is inverted), and the canonical commutator has value $i\hbar$, so

$$[H, r_i] = -\frac{1}{m}i\hbar(\hat{p}_i - qA_i)$$

Plugging this in the time derivative of the expectation value of position, you get

$$\frac{d\langle r_i \rangle}{dt} = \frac{1}{m} \langle \hat{p}_i - qA_i \rangle$$

so the normal momentum mv_i is indeed given by the operator $\hat{p}_i - qA_i$.

On to the other commutator! The i -th component of Newton's second law in expectation form,

$$m \frac{d\langle v_i \rangle}{dt} = \left\langle \frac{i}{\hbar} [H, \hat{p}_i - qA_i] \right\rangle - q \left\langle \frac{\partial A_i}{\partial t} \right\rangle$$

requires the commutator

$$[H, \hat{p}_i - qA_i] = \left[\frac{1}{2m} \sum_{j=1}^3 (\hat{p}_j - qA_j)^2 + q\varphi, p_i - qA_i \right]$$

The easiest is the term $q\varphi$, since both φ and A_i are functions of position and commute. And the commutator with \hat{p}_i is the generalized fundamental operator of chapter 3.4.4,

$$[q\varphi, p_i] = i\hbar q \frac{\partial \varphi}{\partial r_i}$$

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 $\hat{p}_j - qA_j$ to its own side of the commutator:

$$\frac{1}{2m} \sum_{j=1}^3 \left\{ (\hat{p}_j - qA_j)[\hat{p}_j - qA_j, p_i - qA_i] + [\hat{p}_j - qA_j, p_i - qA_i](\hat{p}_j - qA_j) \right\}$$

Work out the simpler commutator appearing here first:

$$[\hat{p}_j - qA_j, p_i - qA_i] = -q[p_j, A_i] - q[A_j, p_i] = i\hbar q \frac{\partial A_i}{\partial r_j} - i\hbar q \frac{\partial A_j}{\partial r_i}$$

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 $\hat{p}_j - qA_j$ that this result appears is not relevant and what is left of the Hamiltonian becomes

$$\frac{q\text{i}\hbar}{m} \sum_{j=1}^3 \left\{ \frac{\partial A_i}{\partial r_j} - \frac{\partial A_j}{\partial r_i} \right\} (\hat{p}_j - qA_j)$$

Now let \bar{i} be the index following i in the sequence 123123... and $\bar{\bar{i}}$ 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 $j = \bar{i}$, and a term where $j = \bar{\bar{i}}$. The total is then:

$$\frac{q\text{i}\hbar}{m} \left\{ (\hat{p}_{\bar{i}} - qA_{\bar{i}}) \left(\frac{\partial A_i}{\partial r_{\bar{i}}} - \frac{\partial A_{\bar{i}}}{\partial r_i} \right) - (\hat{p}_{\bar{\bar{i}}} - qA_{\bar{\bar{i}}}) \left(\frac{\partial A_{\bar{i}}}{\partial r_i} - \frac{\partial A_i}{\partial r_{\bar{\bar{i}}}} \right) \right\}$$

and that is

$$-\frac{q\text{i}\hbar}{m} \left\{ (\hat{p}_{\bar{i}} - qA_{\bar{i}}) (\nabla \times \vec{A})_{\bar{i}} - (\hat{p}_{\bar{\bar{i}}} - qA_{\bar{\bar{i}}}) (\nabla \times \vec{A})_{\bar{\bar{i}}} \right\}$$

and the expression in brackets is the i -th component of $(\hat{\vec{p}} - q\vec{A}) \times (\nabla \times \vec{A})$ and produces the $q\vec{v} \times \vec{B}$ term in Newton's equation provided that $\vec{B} = \nabla \times \vec{A}$.

A.100 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.

A.101 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 10.11.

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.

A.102 Various electrostatic derivations.

This section gives various derivations for the electromagnetostatic solutions of section 10.5.

A.102.1 Existence of a potential

This subsection shows that if the curl of the electric field \vec{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

$$\varphi(\vec{r}) = - \int_{\vec{r}_0}^{\vec{r}} \vec{E}(\underline{\vec{r}}) \, d\underline{\vec{r}} \quad (\text{A.94})$$

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 $\nabla \times \vec{E}$ is zero.

Now if you evaluate φ at a neighboring point $\vec{r} + i\partial x$ by a path first going to \vec{r} and from there straight to $\vec{r} + i\partial x$, the difference in integrals is just the integral over the final segment:

$$\varphi(\vec{r} + i\partial x) - \varphi(\vec{r}) = - \int_{\vec{r}}^{\vec{r} + i\partial x} \vec{E}(\underline{\vec{r}}) \, d\underline{\vec{r}} \quad (\text{A.95})$$

Dividing by ∂x and then taking the limit $\partial x \rightarrow 0$ 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 .

A.102.2 The Laplace equation

The homogeneous Poisson equation,

$$\nabla^2 \varphi = 0 \quad (\text{A.96})$$

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

$$\begin{aligned} 0 &= \int_{\text{sphere}} \nabla^2 \varphi d^3 \vec{r} \\ &= \iiint \frac{\partial \varphi}{\partial r} r^2 \sin \theta d\theta d\phi \\ &= \frac{1}{4\pi} \iint \frac{\partial \varphi}{\partial r} \sin \theta d\theta d\phi \\ &= \frac{\partial}{\partial r} \frac{1}{4\pi} \iint \varphi \sin \theta d\theta d\phi \end{aligned}$$

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.

A.102.3 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 x, z -plane satisfy

$$\frac{dz}{dx} = \frac{E_z}{E_x} = \frac{2z^2 - x^2}{3zx}$$

Change to a new variable u by replacing z by xu to get:

$$x \frac{du}{dx} = -\frac{1+u^2}{3u} \implies \int \frac{3u \, du}{1+u^2} = -\int \frac{dx}{x}$$

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 x, z -plane.

A.102.4 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 E_i integrates to

$$\int_{\text{sphere}} E_i \, d^3 \vec{r} = - \int_{\text{sphere}} \nabla \cdot (\varphi \hat{i}_i) \, d^3 \vec{r} = - \int_{\text{sphere surface}} \varphi n_i \, dA$$

where \hat{i}_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

$$\int_{\text{sphere}} E_i d^3 \vec{r} = -\frac{1}{4\pi\epsilon_0} \int_{\text{sphere surface}} \frac{\vec{\phi} \cdot \vec{r} r_i}{r^3} dA$$

For simplicity, take the z -axis along the dipole moment; then $\vec{\phi} \cdot \vec{r} = \phi z$. For the x -component 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 $\vec{\phi}$, has a delta function. For 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 10.4 is obtained.

A.102.5 Integrals of the current density

In subsequent derivations, various integrals of the current density \vec{j} 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 $\int r_i^m r_{\bar{i}}^n j_i d^3 \vec{r}$ where j_i is any component j_1 , j_2 , or j_3 of the current density, \bar{i} is the index following i in the sequence $\dots 123123\dots$, 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,

$$\int r_i^m r_{\bar{i}}^n j_i d^3 \vec{r} = - \int \frac{r_i^{m+1}}{m+1} r_{\bar{i}}^n \frac{\partial j_i}{\partial r_i} d^3 \vec{r}$$

Now use the fact that the divergence of the current density is zero since the charge density is constant for electromagnetostatic solutions:

$$\int r_i^m r_{\bar{i}}^n j_i d^3 \vec{r} = \int \frac{r_i^{m+1}}{m+1} r_{\bar{i}}^n \frac{\partial j_{\bar{i}}}{\partial r_{\bar{i}}} d^3 \vec{r} + \int \frac{r_i^{m+1}}{m+1} r_{\bar{i}}^n \frac{\partial j_{\bar{i}}}{\partial r_{\bar{i}}} d^3 \vec{r}$$

where \bar{i} is the index preceding i in the sequence $\dots 123123\dots$. The final integral can be integrated in the \bar{i} -direction and is then seen to be zero because \vec{j} vanishes at infinity.

The first integral in the right hand side can be integrated by parts in the \bar{i} -direction to give the final result:

$$\int r_i^m r_{\bar{i}}^n j_i d^3 \vec{r} = - \int \frac{r_i^{m+1}}{m+1} n r_{\bar{i}}^{n-1} j_{\bar{i}} d^3 \vec{r} \quad (\text{A.97})$$

It follows from this equation with $m = 0, n = 1$ that

$$\int r_i j_{\bar{i}} d^3 \vec{r} = - \int r_{\bar{i}} j_i d^3 \vec{r} = \mu_{\bar{i}} \quad \vec{\mu} \equiv \frac{1}{2} \int \vec{r} \times \vec{j} d^3 \vec{r} \quad (\text{A.98})$$

with $\vec{\mu}$ the current distribution's dipole moment. In these expressions, you can swap indices as

$$(i, \bar{i}, \bar{\bar{i}}) \rightarrow (\bar{i}, \bar{\bar{i}}, i) \quad \text{or} \quad (i, \bar{i}, \bar{\bar{i}}) \rightarrow (\bar{\bar{i}}, i, \bar{i})$$

because only the relative ordering of the indices in the sequence ...123123... 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_c/q_c of current carrier mass over charge. Then the angular momentum is

$$\vec{L} = \int \vec{r} \times \frac{m_c}{q_c} \vec{j} d^3 \vec{r} = \frac{2m_c}{q_c} \vec{\mu}$$

A.102.6 Lorentz forces on a current distribution

Next is the derivation of the Lorentz forces on a given current distribution \vec{j} in a constant external magnetic field \vec{B}_{ext} . The Lorentz force law says that the force \vec{F} on a charge q moving with speed \vec{v} equals

$$\vec{F} = q \vec{v} \times \vec{B}_{\text{ext}}$$

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 $d^3 \vec{r}$ is:

$$d\vec{F} = \vec{j} \times \vec{B}_{\text{ext}} d^3 \vec{r}$$

The net force on the current distribution is therefore zero, because according to (A.97) 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

$$\vec{M} = \int \vec{r} \times (\vec{j} \times \vec{B}_{\text{ext}}) d^3 \vec{r}$$

According to the vectorial triple product rule, that is

$$\vec{M} = \int (\vec{r} \cdot \vec{B}_{\text{ext}}) \vec{j} d^3 \vec{r} - \int (\vec{r} \cdot \vec{j}) \vec{B}_{\text{ext}} d^3 \vec{r}$$

The second integral is zero because of (A.97) with $m = 1, n = 0$. What is left is can be written in index notation as

$$M_i = \int r_i B_{\text{ext},i} j_i d^3 \vec{r} + \int r_{\bar{i}} B_{\text{ext},\bar{i}} j_i d^3 \vec{r} + \int r_{\bar{i}} B_{\text{ext},\bar{i}} j_i d^3 \vec{r}$$

The first of the three integrals is zero because of (A.97) with $m = 1, n = 0$. The other two can be rewritten using (A.98):

$$M_i = -\mu_{\bar{i}} B_{\text{ext},\bar{i}} + \mu_{\bar{i}} B_{\text{ext},\bar{i}}$$

and in vector notation that reads

$$\vec{M} = \vec{\mu} \times \vec{B}_{\text{ext}}$$

When the (frozen) current distribution is slowly rotated around the axis aligned with the moment vector, the work done is

$$-M d\alpha = -\mu B_{\text{ext}} \sin \alpha d\alpha = d(\mu B_{\text{ext}} \cos \alpha)$$

where α is the angle between $\vec{\mu}$ and \vec{B}_{ext} . By integration, it follows that the work done corresponds to a change in energy for an energy given by

$$E_{\text{ext}} = -\vec{\mu} \cdot \vec{B}_{\text{ext}}$$

A.102.7 Field of a current dipole

A current density \vec{j} creates a magnetic field because of Maxwell's second and fourth equations for the divergence and curl of the magnetic field:

$$\nabla \cdot \vec{B} = 0 \quad \nabla \times \vec{B} = \frac{1}{\epsilon_0 c^2} \vec{j}$$

where \vec{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

$$\nabla^2 \vec{A} = -\frac{1}{\epsilon_0 c^2} \vec{j}$$

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

$$\vec{B} = \nabla \times \vec{A}$$

because it satisfies the equations for \vec{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, (A.97) with $m = n = 0$, the asymptotic expansion (10.47) of the Green's function integral shows that at large distances, the components of \vec{A} behave as a dipole potential. Specifically,

$$A_i \sim \frac{1}{4\pi\epsilon_0 c^2 r^3} \sum_{i=1}^3 r_i \int \underline{r}_i j_i d^3 \underline{r}$$

Now the term $\underline{i} = i$ in the sum does not give a contribution, because of (A.97) with $m = 1, n = 0$. The other two terms are

$$A_i \sim \frac{1}{4\pi\epsilon_0 c^2 r^3} \left[r_{\bar{i}} \int \underline{r}_{\bar{i}} j_i d^3 \underline{r} + r_{\bar{i}} \int \underline{r}_i j_{\bar{i}} d^3 \underline{r} \right]$$

with \bar{i} following i in the sequence $\dots 123123 \dots$ and $\bar{\bar{i}}$ preceding it. These two integrals can be rewritten using (A.98) to give

$$A_i \sim -\frac{1}{4\pi\epsilon_0 c^2 r^3} [r_{\bar{i}} \mu_{\bar{i}} - r_{\bar{\bar{i}}} \mu_{\bar{i}}]$$

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

$$B_i = \frac{\partial A_{\bar{i}}}{\partial r_{\bar{i}}} - \frac{\partial A_{\bar{\bar{i}}}}{\partial r_{\bar{i}}}$$

and substituting in for the vector potential from above, differentiating, and cleaning up produces

$$B_i = \frac{3(\vec{\mu} \cdot \vec{r}) \vec{r} - \vec{\mu} r^2}{4\pi\epsilon_0 c^2 r^5}$$

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:

$$\int B_i d^3\vec{r} = \int \frac{\partial A_{\bar{i}}}{\partial r_{\bar{i}}} d^3\vec{r} - \int \frac{\partial A_{\bar{i}}}{\partial r_{\bar{i}}} d^3\vec{r}$$

Using the divergence theorem, the right hand side becomes an integral over the surface of the sphere:

$$\int B_i d^3\vec{r} = \int A_{\bar{i}} \frac{r_{\bar{i}}}{r} dA - \int A_{\bar{i}} \frac{r_{\bar{i}}}{r} dA$$

Substituting in the asymptotic expression for A_i above,

$$\int B_i d^3\vec{r} = -\frac{1}{4\pi\epsilon_0 c^2 r^4} \left[\int (r_i \mu_{\bar{i}} - r_{\bar{i}} \mu_i) r_{\bar{i}} dA - \int (r_{\bar{i}} \mu_i - r_i \mu_{\bar{i}}) r_{\bar{i}} dA \right]$$

The integrals of $r_i r_{\bar{i}}$ and $r_i r_{\bar{i}}$ are zero, for one because the integrand is odd in r_i . The integrals of $r_{\bar{i}} r_{\bar{i}}$ and $r_{\bar{i}} r_{\bar{i}}$ 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 $4\pi r^2$,

$$\int B_i d^3\vec{r} = \frac{2}{3\epsilon_0 c^2} \mu_i$$

That gives the strength of the delta function for an ideal current dipole.

A.102.8 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

$$\nabla^2 \vec{A} = -\frac{1}{\epsilon_0 c^2} \vec{j}$$

The solution for the vector potential can be written explicitly in terms of the current density using the Green's function integral (10.45):

$$A_i = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{1}{|\vec{r} - \underline{\vec{r}}|} j_i(\underline{\vec{r}}) d^3\underline{\vec{r}}$$

The magnetic field is the curl of \vec{A} ,

$$B_i = \frac{\partial A_{\bar{i}}}{\partial r_{\bar{i}}} - \frac{\partial A_{\bar{i}}}{\partial r_{\bar{i}}}$$

or substituting in and differentiating under the integral

$$B_i = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{r_{\bar{i}} - \underline{r}_{\bar{i}}}{|\vec{r} - \underline{\vec{r}}|^3} j_{\bar{i}}(\underline{\vec{r}}) - \frac{\underline{r}_{\bar{i}} - r_{\bar{i}}}{|\vec{r} - \underline{\vec{r}}|^3} j_i(\underline{\vec{r}}) d^3\underline{\vec{r}}$$

In vector notation that gives the Biot-Savart law

$$\vec{B} = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{\vec{r} - \underline{\vec{r}}}{|\vec{r} - \underline{\vec{r}}|^3} \times \vec{j} d^3 \underline{\vec{r}}$$

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

$$\vec{j} d^3 \underline{\vec{r}} = I d\underline{\vec{r}}$$

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

$$\vec{B} = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{\vec{r} - \underline{\vec{r}}}{|\vec{r} - \underline{\vec{r}}|^3} \times I(\underline{\vec{r}}) d\underline{\vec{r}}$$

where the integration is over all infinitesimal segments $d\underline{\vec{r}}$ of the wires.

A.103 Energy due to 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 subsection 10.3, the Hamiltonian is

$$H = \frac{1}{2m} (\hat{\vec{p}} - q\vec{A})^2 + V$$

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{B} by $\vec{B} = \nabla \times \vec{A}$. The potential energy V is of no particular interest in this note. The first term is, and it can be multiplied out as:

$$H = \frac{1}{2m} \hat{\vec{p}}^2 - \frac{q}{2m} (\hat{\vec{p}} \cdot \vec{A} + \vec{A} \cdot \hat{\vec{p}}) + \frac{q^2}{2m} (\vec{A})^2 + V$$

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:

$$H_{BL} \equiv -\frac{q}{2m} (\hat{\vec{p}} \cdot \vec{A} + \vec{A} \cdot \hat{\vec{p}}) \quad H_{BD} \equiv \frac{q^2}{2m} (\vec{A})^2$$

Now to simplify the analysis, align the z -axis with \vec{B} so that $\vec{B} = \hat{k}B_z$. Then an appropriate vector potential \vec{A} is

$$\vec{A} = -\hat{i}\frac{1}{2}yB_z + \hat{j}\frac{1}{2}xB_z.$$

The vector potential is not unique, but a check shows that indeed $\nabla \times \vec{A} = \hat{k}B_z = \vec{B}$ for the one above. Also, the canonical momentum is

$$\hat{\vec{p}} = \frac{\hbar}{i}\nabla = \hat{i}\frac{\hbar}{i}\frac{\partial}{\partial x} + \hat{j}\frac{\hbar}{i}\frac{\partial}{\partial y} + \hat{k}\frac{\hbar}{i}\frac{\partial}{\partial z}$$

Therefore, in the term H_{BL} above,

$$H_{BL} = -\frac{q}{2m}(\hat{\vec{p}} \cdot \vec{A} + \vec{A} \cdot \hat{\vec{p}}) = -\frac{q}{2m}B_z \left(x\frac{\hbar}{i}\frac{\partial}{\partial y} - y\frac{\hbar}{i}\frac{\partial}{\partial x} \right) = -\frac{q}{2m}B_z \hat{L}_z$$

the latter equality being true because of the definition of angular momentum as $\vec{r} \times \hat{\vec{p}}$. Because the z -axis was aligned with \vec{B} , $B_z \hat{L}_z = \vec{B} \cdot \hat{\vec{L}}$, so, finally,

$$H_{BL} = -\frac{q}{2m}\vec{B} \cdot \hat{\vec{L}}.$$

Similarly, in the part H_{BD} of the Hamiltonian, substitution of the expression for \vec{A} produces

$$\frac{q^2}{2m}(\vec{A})^2 = \frac{q^2}{8m}B_z^2(x^2 + y^2),$$

or writing it so that it is independent of how the z -axis is aligned,

$$H_{BD} = \frac{q^2}{8m}(\vec{B} \times \vec{r})^2$$

A.104 Energy due to 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, section 10.2, assumes that Einstein's energy square root falls apart in a linear combination of terms:

$$H = \sqrt{(m_0c^2)^2 + \sum_{i=1}^3 (\hat{p}_i c)^2} = \alpha_0 m_0 c^2 + \sum_{i=1}^3 \alpha_i \hat{p}_i c$$

which works for the $4 \times 4 \alpha$ matrices given in that section. For an electron in a magnetic field, according to subsection 10.3 you want to replace $\hat{\vec{p}}$ with $\hat{\vec{p}} - q\vec{A}$ 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 $\hat{\vec{p}} - q\vec{A}$ 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 section 10.2 and its note. You get, in index notation,

$$H^2 = (m_0 c^2)^2 I + \sum_{i=1}^3 ((\hat{p}_i - qA_i)c)^2 I + \sum_{i=1}^3 [\hat{p}_{\bar{i}} - qA_{\bar{i}}, \hat{p}_{\bar{i}} - qA_{\bar{i}}] c^2 \alpha_{\bar{i}} \alpha_{\bar{i}}$$

where I is the four by four unit matrix, \bar{i} is the index following i in the sequence 123123..., and \bar{i} is the one preceding i . The final sum represents the additional squared energy that you get by substituting $\hat{\vec{p}} - q\vec{A}$ in the linear combination instead of the square root. The commutator arises because $\alpha_{\bar{i}} \alpha_{\bar{i}} + \alpha_{\bar{i}} \alpha_{\bar{i}} = 0$, giving the terms with the indices reversed the opposite sign. Working out the commutator using the formulae of chapter 3.4.4, and the definition of the vector potential \vec{A} ,

$$H^2 = (m_0 c^2)^2 I + \sum_{i=1}^3 ((\hat{p}_i - qA_i)c)^2 I + q\hbar c^2 i \sum_{i=1}^3 B_i \alpha_{\bar{i}} \alpha_{\bar{i}}.$$

By multiplying out the expressions for the α_i of section 10.2, using the fundamental commutation relation for the Pauli spin matrices that $\sigma_{\bar{i}} \sigma_{\bar{i}} = i\sigma_i$,

$$H^2 = (m_0 c^2)^2 I + \sum_{i=1}^3 ((\hat{p}_i - qA_i)c)^2 I - q\hbar c^2 \sum_{i=1}^3 B_i \begin{pmatrix} \sigma_i & 0 \\ 0 & \sigma_i \end{pmatrix}$$

It is seen that due to the interaction of the spin with the magnetic field, the square energy changes by an amount $-q\hbar c^2 \sigma_i B_i$. Since $\frac{1}{2}\hbar$ times the Pauli spin matrices gives the spin $\hat{\vec{S}}$, the square energy due to the magnetic field acting on the spin is $-2qc^2 \hat{\vec{S}} \cdot \vec{B}$.

In the nonrelativistic case, the rest mass energy $m_0 c^2$ is much larger than the other terms, and in that case, if the change in square energy is $-2qc^2 \hat{\vec{S}} \cdot \vec{B}$, the change in energy itself is smaller by a factor $2m_0 c^2$, so the energy due to the magnetic field is

$$H_{SB} = -\frac{q}{m} \hat{\vec{S}} \cdot \vec{B} \tag{A.99}$$

which is what was to be proved.

A.105 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 3.1.4, 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 definite nonzero values, rather than just being uncertain. But the latter is the truth.

A.106 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

$$a = Ae^{i\omega t/2}, \quad b = Be^{-i\omega t/2}.$$

Then find the eigenvalues and eigenvectors of the now constant matrix. The eigenvalues can be written as $\pm i\omega_1/f$, where f is the resonance factor given in the main text. The solution is then

$$\begin{pmatrix} A \\ B \end{pmatrix} = C_1 \vec{v}_1 e^{i\omega_1 t/f} + C_2 \vec{v}_2 e^{-i\omega_1 t/f}$$

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.

A.107 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 note {A.17}, 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

$$-\frac{1}{R} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) + l(l+1) + \frac{2m_e}{\hbar^2} \frac{1}{2} m_e \omega^2 r^4 = \frac{2m_e}{\hbar^2} r^2 E$$

The difference from {A.17} is that a harmonic oscillator potential $\frac{1}{2}m_e\omega^2r^2$ has replaced the Coulomb potential. A suitable rescaling is now $r = \rho\sqrt{\hbar/m_e\omega}$,

which produces

$$-\frac{1}{R} \frac{d}{d\rho} \left(\rho^2 \frac{dR}{d\rho} \right) + l(l+1) + \rho^4 = \rho^2 \epsilon$$

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

$$\rho^2 f'' + 2\rho f' - l(l+1)f = 2\rho^3 f' + (3 - \epsilon)\rho^2 f$$

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

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega$$

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.

A.108 Impenetrable spherical shell

To solve the problem of particles stuck inside an impenetrable shell of radius a , refer to note {A.56}. According to that note, the solutions without unacceptable singularities at the center are of the form

$$\psi_{Elm}(r, \theta, \phi) \propto j_l(p_{rmc}r/\hbar)Y_l^m(\theta, \phi) \quad p_{rmc} \equiv \sqrt{2m(E - V)} \quad (\text{A.100})$$

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

$$E_{\bar{n}l} = \frac{\hbar^2}{2ma^2}\beta_{\bar{n}l}^2 + V_0 \quad (\text{A.101})$$

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

$$n = l - 1 + 2\bar{n}$$

A.109 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.109.1 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 3.1.

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

$$\rho_m = \frac{Am_p}{\frac{4}{3}\pi R_0^3}$$

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.109.2 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 dS of that chunk gobbles up an amount of volume while it moves given by $\vec{v} \cdot \vec{n} dS$, 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:

$$\int_S \vec{v} \cdot \vec{n} dS = 0$$

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 10.4.

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 $d^3\vec{r}$ 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, [27, 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

$$R(\theta_0, \phi_0) = R_0 + \delta(\theta_0, \phi_0) \quad \delta = \sum_{l,m} r_{lm}(t) l \bar{Y}_l^m(\theta_0, \phi_0) \quad (\text{A.102})$$

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

$$T = \int \frac{1}{2} \rho_m \vec{v} \cdot \vec{v} d^3 \vec{r}_0$$

Putting in the expression for the velocity field in terms of the r_{lm} position coefficients gives

$$T = \int \frac{1}{2} \rho_m \sum_{l,m} \dot{r}_{lm} R_0 \left(\nabla \frac{r_0^l}{R_0^l} \bar{Y}_l^m \right) \cdot \sum_{\underline{l},\underline{m}} \dot{r}_{\underline{l}\underline{m}} R_0 \left(\nabla \frac{r_0^{\underline{l}}}{R_0^{\underline{l}}} \bar{Y}_{\underline{l}}^{\underline{m}} \right) d^3 \vec{r}_0$$

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:

$$\int_V (\nabla F) \cdot (\nabla G) d^3\vec{r}_0 = \int_V \nabla (F \cdot \nabla G) d^3\vec{r}_0 = \int_S F \frac{\partial G}{\partial r_0} dS_0 \quad (\text{A.103})$$

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:

$$dS_0 = R_0^2 \sin \theta_0 d\theta_0 d\phi_0$$

Applying these results to the integral for the kinetic energy, noting that $r_0 = R_0$ on the surface of the droplet, gives

$$T = \frac{1}{2} \rho_m \sum_{l,m} \sum_{\underline{l},\underline{m}} \dot{r}_{lm} \dot{r}_{\underline{l}\underline{m}} l R_0^3 \int \int \bar{Y}_l^m \bar{Y}_{\underline{l}}^{\underline{m}} \sin \theta_0 d\theta_0 d\phi_0$$

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

$$T = \frac{1}{2} \rho_m R_0^3 \sum_{l,m} l \dot{r}_{lm}^2 \quad (\text{A.104})$$

A.109.3 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 $d\theta d\phi$ corresponds to a physical surface element that is approximately a parallelogram. Specifically, the sides of that parallelogram are

$$d\vec{r}_1 = \frac{\partial \vec{r}_{\text{surface}}}{\partial \theta} d\theta \quad d\vec{r}_2 = \frac{\partial \vec{r}_{\text{surface}}}{\partial \phi} d\phi$$

To get the surface area dS , 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,

$$\vec{r}_{\text{surface}} = \hat{i}_r R \quad \frac{\partial \hat{i}_r}{\partial \theta} = \hat{i}_\theta \quad \frac{\partial \hat{i}_r}{\partial \phi} = \sin \theta \hat{i}_\phi$$

That way, the surface area works out to be

$$S = \int \int \sqrt{1 + \left(\frac{1}{R} \frac{\partial R}{\partial \theta} \right)^2 + \left(\frac{1}{R \sin \theta} \frac{\partial R}{\partial \phi} \right)^2} R^2 \sin \theta d\theta d\phi$$

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

$$\begin{aligned} V_s &= \sigma \int \int R_0^2 \sin \theta d\theta d\phi + \sigma \int \int 2R_0 \delta \sin \theta d\theta d\phi + \sigma \int \int \delta^2 \sin \theta d\theta d\phi \\ &\quad + \sigma \int \int \frac{1}{2} \left[\left(\frac{1}{R_0} \frac{\partial \delta}{\partial \theta} \right)^2 + \left(\frac{1}{R_0 \sin \theta} \frac{\partial \delta}{\partial \phi} \right)^2 \right] R_0^2 \sin \theta d\theta d\phi \end{aligned}$$

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

$$\frac{4\pi}{3} R_0^3 = \int \int \int r^2 \sin \theta dr d\theta d\phi = \int \int \frac{1}{3} R^3 \sin \theta d\theta d\phi$$

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

$$-\int \int R_0^2 \delta \sin \theta d\theta d\phi = \int \int R_0 \delta^2 \sin \theta d\theta d\phi$$

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

$$\begin{aligned} V_s &= V_{s,0} - \sigma \int \int \delta^2 \sin \theta d\theta d\phi \\ &\quad + \sigma \int \int \frac{1}{2} \left[\left(\frac{1}{R_0} \frac{\partial \delta}{\partial \theta} \right)^2 + \left(\frac{1}{R_0 \sin \theta} \frac{\partial \delta}{\partial \phi} \right)^2 \right] R_0^2 \sin \theta d\theta d\phi \end{aligned}$$

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

$$I = \int \int \left[\frac{\partial \bar{Y}_l^m}{\partial \theta} \frac{\partial \bar{Y}_{\underline{l}}^{\underline{m}}}{\partial \theta} + \frac{1}{\sin^2 \theta} \frac{\partial \bar{Y}_l^m}{\partial \phi} \frac{\partial \bar{Y}_{\underline{l}}^{\underline{m}}}{\partial \phi} \right] \sin \theta d\theta d\phi$$

This integral can be simplified by using the same theorem (A.103) 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

$$V_s = V_{s,0} + \sum_{l,m} \frac{1}{2}(l-1)(l+2)\sigma\delta_{lm}^2 \quad (\text{A.105})$$

A.109.4 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 $d\delta'$ until the full surface perturbation δ is achieved. (With adding negative amounts $d\delta'$ 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:

$$\delta' = \sum_{l,m} \delta'_{lm} \bar{Y}_l^m \quad d\delta' = \sum_{l,m} d\delta'_{lm} \bar{Y}_l^m$$

The amount of charge added per unit solid angle $d\Omega = \sin\theta d\theta d\phi$ will be called γ' . It is given in terms of the charge density ρ_c and δ' as

$$\gamma' = \rho_c \left(\frac{1}{3}(R_0 + \delta')^3 - \frac{1}{3}R_0^3 \right)$$

To first approximation, the incremental amount of charge laid down per unit solid angle is

$$d\gamma' \sim \rho_c R_0^2 d\delta'$$

However, if $d\gamma$ is written in terms of spherical harmonics,

$$d\gamma' = \sum_{l,m} d\gamma'_{lm} \bar{Y}_l^m \quad d\gamma'_{lm} \sim \rho_c R_0^2 d\delta'_{lm}$$

then the coefficient $d\gamma'_{00}$ 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 $d\delta'_{00}$ 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

$$V_{0,\text{surface}} \sim \frac{Ze}{4\pi\epsilon_0(R_0 + \delta')} \sim \frac{Ze}{4\pi\epsilon_0 R_0} - \frac{Ze}{4\pi\epsilon_0 R_0^2} \delta'$$

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

$$V_{1,\text{inside}} = \sum_{l,m} V_{lm} \frac{r^l}{R_0^l} \bar{Y}_l^m$$

though the coefficients V_{lm} are still unknown. The potential outside the shell takes the form

$$V_{1,\text{outside}} = \sum_{l,m} V_{lm} \frac{R_0^{l+1}}{r^{l+1}} \bar{Y}_l^m$$

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

$$\rho_c \delta' dS = (E_{r,\text{immediately outside}} - E_{r,\text{immediately inside}}) dS$$

where dS is the area of the shell element. Note that $E_r = -\partial V_1 / \partial r$, 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

$$V_{1,\text{surface}} = \sum_{l,m} \frac{\rho_c R_0}{(2l+1)\epsilon_0} \delta'_{lm} \bar{Y}_l^m \quad \rho_c = \frac{Ze}{\frac{4}{3}\pi R_0^3}$$

Multiplying the two surface potentials by the amount of charge laid down gives the incremental change in Coulomb potential of the drop as

$$dV_c = \left[\frac{Ze}{4\pi\epsilon_0 R_0} - \frac{Ze}{4\pi\epsilon_0 R_0^2} \delta' + \sum_{l,m} \frac{3Ze}{(2l+1)4\pi\epsilon_0 R_0^2} \delta'_{lm} \bar{Y}_l^m \right] d\gamma' \sin\theta d\theta d\phi$$

Substituting in $\delta' = \sum_{l,m} \delta'_{lm} \bar{Y}_l^m$ and $d\gamma' = \sum_{l,m} d\gamma'_{lm} \bar{Y}_l^m$, 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 $d\gamma'_{00}$ is zero. For the other terms, it is accurate enough to set $d\gamma'_{lm} = \rho_c R_0^2 d\delta'_{lm}$ and then δ' can be integrated from zero to δ to give the Coulomb potential of the fully deformed sphere:

$$V_c = V_{c,0} - \sum_{l,m} \frac{l-1}{2l+1} \frac{Ze}{4\pi\epsilon_0} \rho_c \delta_{lm}^2 \quad (\text{A.106})$$

A.109.5 Frequency of vibration

Having found the kinetic energy, (A.104), and the potential energy, (A.105) plus (A.106), the motion of the drop can be determined.

A rigorous analysis would so using a Lagrangian analysis, {A.3}. It would use the coefficients r_{lm} as generalized coordinates, getting rid of the δ_{lm} in the potential energy terms using (A.102). 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

$$\omega^2 = \frac{(l-1)l(l+2)}{3} \frac{C_s}{R_A^2 m_p A} - \frac{2(l-1)l}{2l+1} \frac{e^2}{4\pi\epsilon_0 R_A^3 m_p A^2} Z^2$$

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 an 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 $l \geq 2$, and the most important ones are the ones with the lowest frequency, which means $l = 2$.

A.110 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 10.1.7, 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 {A.89}, 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 $i \geq 3$ 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, [21, 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.

A.111 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 11.52. It also describes the Fermi integral plotted in figure 11.50, as well as Fermi's (second) golden rule.

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 11.52.

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 [16] 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.112}. *That is news to the author. But it was supposed to be there, I think.*

A.111.1 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

$$\psi_1(\vec{r}_1, S_{z,1}, \vec{r}_2, S_{z,2}, \dots, \vec{r}_A, S_{z,A})$$

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

$$\psi_2(\vec{r}_1, S_{z,1}, \vec{r}_2, S_{z,2}, \dots, \vec{r}_A, S_{z,A}, \vec{r}_e, S_{z,e}, \vec{r}_{\bar{\nu}}, S_{z,\bar{\nu}})$$

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

$$\psi_2 = \psi_{2,\text{nuc}} |1e, \vec{p}_e\rangle |1\bar{\nu}, \vec{p}_{\bar{\nu}}\rangle \quad (\text{A.107})$$

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

$$\psi_1 = \psi_{1,\text{nuc}} |0e, \vec{p}_e\rangle |0\bar{\nu}, \vec{p}_{\bar{\nu}}\rangle \quad (\text{A.108})$$

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

$$\Psi_N(\vec{r}_1, S_{z,1}, T_{3,1}, \vec{r}_2, S_{z,2}, T_{3,2}, \dots, \vec{r}_A, S_{z,A}, T_{3,A}; t)$$

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.111.2 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 6.1.5 and 6.3. 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 $E_1 = m_{\text{N1}}c^2$.

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

$$E_2 = \langle 1\bar{\nu}, \vec{p}_{\bar{\nu}} | \langle 1e, \vec{p}_e | \psi_{2,\text{nuc}} | H \psi_{2,\text{nuc}} | 1e, \vec{p}_e \rangle | 1\bar{\nu}, \vec{p}_{\bar{\nu}} \rangle$$

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:

$$H_{21} = \langle 1\bar{\nu}, \vec{p}_{\bar{\nu}} | \langle 1e, \vec{p}_e | \psi_{2,\text{nuc}} | H \psi_{1,\text{nuc}} | 0e, \vec{p}_e \rangle | 0\bar{\nu}, \vec{p}_{\bar{\nu}} \rangle$$

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 \hat{a}^\dagger :

$$H' \propto e \psi_{\text{photon}}(\vec{r}) \hat{a}^\dagger$$

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 \hat{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 chapter 12.2 for more details on electromagnetic interactions, including a more precise description of H' . See also {A.112}.)

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 6.4.2, 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 8.

In a periodic box, the wave function of the antineutrino is

$$\psi_{\bar{\nu}, \vec{p}_{\bar{\nu}}} = \frac{1}{\sqrt{\mathcal{V}}} e^{i\vec{p}_{\bar{\nu}} \cdot \vec{r}/\hbar} = \frac{1}{\sqrt{\mathcal{V}}} e^{i(p_{x,\bar{\nu}}x + p_{y,\bar{\nu}}y + p_{z,\bar{\nu}}z)/\hbar}$$

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 $\hbar\partial/i\partial x$ 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:

$$\psi_{e, \vec{p}_e}(\vec{r}) = \frac{1}{\sqrt{\mathcal{V}}} e^{i\vec{p}_e \cdot \vec{r}/\hbar} = \frac{1}{\sqrt{\mathcal{V}}} e^{i(p_{x,e}x + p_{y,e}y + p_{z,e}z)/\hbar}$$

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

$$H_{21} = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\text{nuc}} | \sum_{i=1}^A h_i e^{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i / \hbar} \psi_{1,\text{nuc}} \rangle \quad (\text{A.109})$$

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 $\langle \psi_{2,\text{nuc}} | h_i \psi_{1,\text{nuc}} \rangle$ should never be greater than about one, though it could be much less if $\psi_{2,\text{nuc}}$ and $h_i \psi_{1,\text{nuc}}$ turn out to be almost orthogonal. It is found that g has a rough value of about 100 eV fm³, depending a bit on whether it is a Fermi or Gamow-Teller decay. Figure 11.52 simply used 100 MeV fm³.

A.111.3 Allowed or forbidden

The question of allowed and forbidden decays is directly related to the Hamiltonian coefficient H_{21} , (A.109), 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

$$\frac{1}{\mathcal{V}} e^{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i / \hbar}$$

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:

$$e^{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i / \hbar} \approx 1 + \frac{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\hbar} + \frac{1}{2!} \left(\frac{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\hbar} \right)^2 + \dots$$

Since the first term in the Taylor series is by far the largest, you would expect that the value of H_{21} , (A.109), can be well approximated by replacing the exponential by 1, giving:

$$H_{21}^0 = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\text{nuc}} | \sum_{i=1}^A h_i \psi_{1,\text{nuc}} \rangle$$

However, clearly this approximation does not work if the value of H_{21}^0 is zero for some reason:

If the simplified coefficient H_{21}^0 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 $\psi_{2,\text{nuc}}$ and $h_i \psi_{1,\text{nuc}}$ be *exactly* orthogonal? If you took two random wave functions for $\psi_{2,\text{nuc}}$ and $\psi_{1,\text{nuc}}$, they definitely would not be. But $\psi_{2,\text{nuc}}$ and $\psi_{1,\text{nuc}}$ 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, chapter 6.2. 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 $\psi_{1,\text{nuc}}$ in the inner product in H_{21}^0 , 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_{21}^0 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_{21}^0 . Therefore, their orbital angular momentum is approximated to be zero too. Under these condition H_{21}^0 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 subsection 11.19.4.

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_{21}^0 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, subsection 11.19.6. 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,

$$H_{21}^1 = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\text{nuc}} | \sum_{i=1}^A h_i \frac{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\hbar} \psi_{1,\text{nuc}} \rangle$$

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, {A.15}

$$r Y_1^m = \sum_j c_j r_j$$

with the c_j some constants. Therefore, the factor \vec{r}_i in H_{21}^1 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_{21}^1 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 ℓ th-forbidden decay,

$$H_{21}^\ell = \frac{g}{\ell! \sqrt{\mathcal{V}}} \langle \psi_{2,\text{nuc}} | \sum_{i=1}^A h_i \left(\frac{i(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\hbar} \right)^\ell \psi_{1,\text{nuc}} \rangle \quad (\text{A.110})$$

must be the first nonzero inner product. Note that an ℓ th-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 ℓ th-forbidden decay is slowed down by a factor of order $(pR/\hbar)^{2\ell}$, making highly forbidden decays extremely slow.

A.111.4 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 pseudo-scalar (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, subsection 11.19.6, 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_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_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, [21, 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 10.1.9 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],

$$h_i = \tau_1 \pm i\tau_2 \quad h_i = (\tau_1 \pm i\tau_2) \sum_{j=1}^3 \sigma_j$$

for Fermi and Gamow-Teller decays respectively. Here the three σ_j are the

Pauli spin matrices of chapter 10.1.9. 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 $\langle \psi_{2,\text{nuc}} | h_i \psi_{1,\text{nuc}} \rangle$ 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, section 11.18. 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 $T_{3,\text{net}}$. 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 (10.9) and (10.10) in chapter 10.1 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 fm³, 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.111.5 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_{21}^ℓ 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

$$i\hbar \dot{\bar{b}} = H_{21}^\ell e^{i(E_2 - E_1)t/\hbar} \bar{a}$$

(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 6.3.1 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

$$\bar{b} = -H_{21}^\ell \frac{e^{i(E_2 - E_1)t_c/\hbar} - 1}{(E_2 - E_1)}$$

Half of the exponential can be factored out to produce a real ratio:

$$\bar{b} = -H_{21}^\ell e^{i\frac{1}{2}(E_2 - E_1)t_c/\hbar} \frac{i \sin\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)}{\frac{1}{2}(E_2 - E_1)t_c/\hbar} t_c$$

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 given 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:

$$\lambda_{\text{single final state}} = \frac{|H_{21}^\ell|^2}{\hbar^2} \frac{\sin^2\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)}{\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar\right)^2} t_c \quad (\text{A.111})$$

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 $1 \rightarrow 2$ will not be the same as the one for $2 \rightarrow 1$: 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_{21}^ℓ assures that, chapter 4.3, 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.

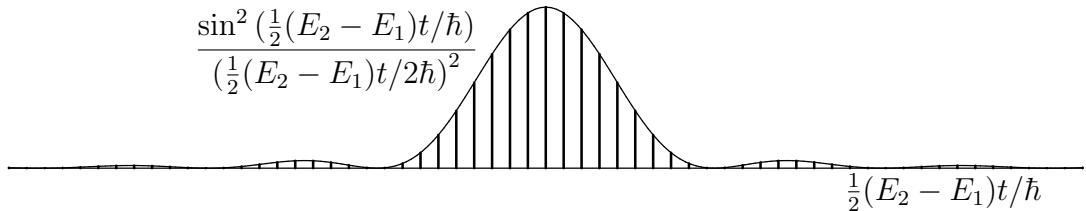


Figure A.24: Energy slop diagram.

To narrow this effect down more precisely, the fraction is plotted in figure A.24. 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:

$$\lambda = \int_{\text{all } E_2} \frac{|H_{21}^\ell|^2}{\hbar^2} \frac{\sin^2 \left(\frac{1}{2}(E_2 - E_1)t_c/\hbar \right)}{\left(\frac{1}{2}(E_2 - E_1)t_c/\hbar \right)^2} t_c \frac{dN}{dE_2} dE_2$$

where dN/dE_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 $|H_{21}^\ell|$ is about constant. It can then be taken out of the integral. Also, the range of energy in figure A.24 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 dN/dE_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, [28, 18.36]. That gives:

$$\lambda = \frac{|H_{21}^\ell|^2}{\hbar^2} t_c \frac{dN}{dE} \frac{2\pi\hbar}{t_c}$$

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.

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:

$\lambda_{\text{single final state}} = \frac{2\pi}{\hbar\varepsilon} |H_{21}^\ell|^2 \quad \Delta E_{\text{slop}} \equiv \varepsilon \quad \varepsilon t_c \sim 2\pi\hbar$

(A.112)

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 dN/dE 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

$$|H_{21}| \ll \varepsilon \ll E \quad \varepsilon \equiv \frac{2\pi\hbar}{t_c} \quad (\text{A.113})$$

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.111.6 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

$$\psi_{\vec{p}} = \frac{1}{\sqrt{\mathcal{V}}} e^{i(p_x x + p_y y + p_z z)/\hbar}$$

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 $e^{ip_x x/\hbar}$ 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.

Graphically this can be 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.25. Each point corresponds 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 $(2\pi\hbar)^3/\mathcal{V}$

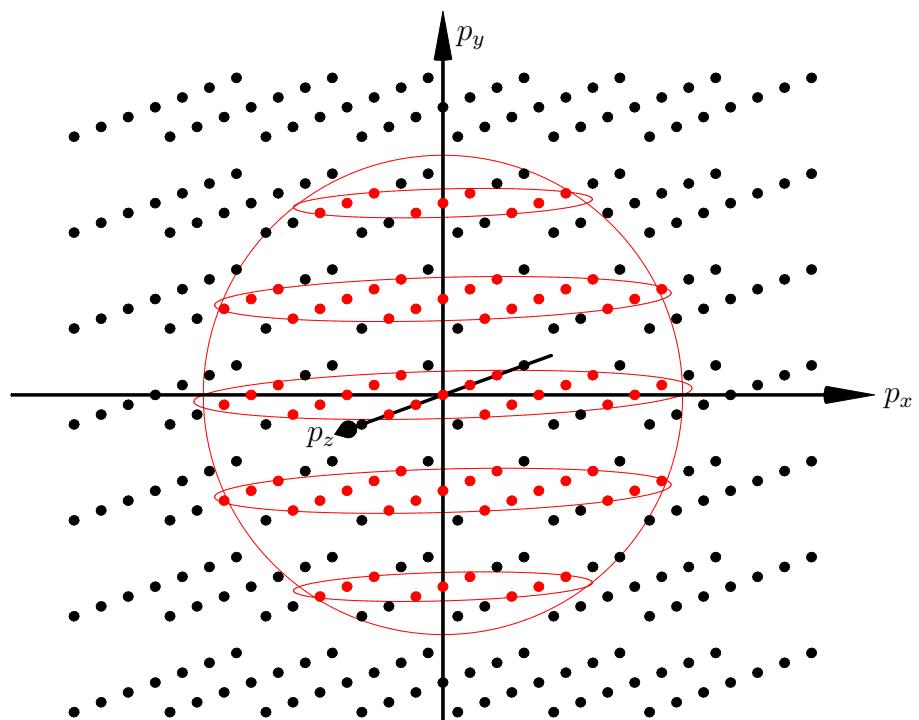


Figure A.25: Possible momentum states for a particle confined to a periodic box. The states are shown as points in momentum space. States that have momentum less than some example maximum value are in red.

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.25. 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:

$$N_{|\vec{p}'| \leq p} = \frac{\frac{4}{3}\pi p^3}{(2\pi\hbar)^2/\mathcal{V}}$$

The number of electron states that have momentum in a range from p_e to $p_e + dp_e$ can be found by taking a differential of the expression above:

$$dN_e = \frac{\mathcal{V}p_e^2}{2\pi^2\hbar^3} dp_e$$

(Here the range dp_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, {A.4} (A.7):

$$T_e = \sqrt{(m_e c^2)^2 + p_e^2 c^2} - m_e c^2 \quad T_{\bar{\nu}} = p_{\bar{\nu}} c \quad (\text{A.114})$$

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 $dp_{\bar{\nu}}$ 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

$$\Delta N_{\bar{\nu}} = \frac{\mathcal{V} p_{\bar{\nu}}^2}{2\pi^2 \hbar^3 c} \frac{1}{c} \varepsilon \quad p_{\bar{\nu}} = \frac{1}{c} (Q - T_e)$$

The number of complete system states in an electron momentum range dp_e is the product of the number of electron states times the number of antineutrino states:

$$dN = dN_e \Delta N_{\bar{\nu}} = \frac{\mathcal{V}^2}{4\pi^4 \hbar^6 c} p_e^2 p_{\bar{\nu}}^2 \varepsilon dp_e$$

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

$$\lambda = \int_{p_e=0}^{p_{e,\max}} \frac{\mathcal{V}^2 |H_{21}^\ell|^2}{2\pi^3 \hbar^7 c} p_e^2 p_{\bar{\nu}}^2 dp_e$$

where the maximum electron momentum $p_{e,\max}$ can be computed from the Q -value of the decay using (A.114). (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.110) for the Hamiltonian coefficient H_{21}^ℓ can be written in the form

$$|H_{21}^\ell|^2 = \frac{g^2}{\mathcal{V}^2} \frac{1}{(\ell!)^2} \left(\frac{\sqrt{p_e^2 + p_{\bar{\nu}}^2} R}{\hbar} \right)^{2\ell} C_N^\ell$$

$$C_N^\ell \equiv \overline{\left| \left\langle \psi_{2,\text{nuc}} \left| \sum_{i=1}^A h_i \left(\frac{(\vec{p}_e + \vec{p}_{\bar{\nu}}) \cdot \vec{r}_i}{\sqrt{p_e^2 + p_{\bar{\nu}}^2} R} \right)^\ell \psi_{1,\text{nuc}} \right| \right\rangle \right|^2}$$

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

$$\lambda = \frac{1}{2\pi^3} \frac{g^2 m_e^4 c^2}{\hbar^6} \frac{m_e c^2}{\hbar} \frac{\tilde{R}^{2\ell}}{(\ell!)^2} C_N^\ell \int_{\tilde{p}_e=0}^{\tilde{p}_{e,\max}} (\tilde{p}_e^2 + \tilde{p}_\nu^2)^\ell \tilde{p}_e^2 \tilde{p}_\nu^2 F^\ell d\tilde{p}_e \quad (\text{A.115})$$

where the \tilde{p} indicate the electron and antineutrino momenta nondimensionalized with $m_e c$. Also,

$$\tilde{Q} \equiv \frac{Q}{m_e c^2} \quad \tilde{p}_{e,\max} = \sqrt{\tilde{Q}^2 + 2\tilde{Q}} \quad \tilde{p}_\nu = \tilde{Q} - \sqrt{1 + \tilde{p}_e^2} + 1 \quad \tilde{R} \equiv \frac{m_e c R}{\hbar} \quad (\text{A.116})$$

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

$$F(\tilde{p}_e, Z_2, A) = \frac{2(1+\xi)}{\Gamma^2(1+2\xi)} \frac{1}{(2\tilde{p}_e \tilde{R})^{2-2\xi}} e^{\pi\eta} |\Gamma(\xi + i\eta)|^2 \quad (\text{A.117})$$

$$\xi \equiv \sqrt{1 - (\alpha Z_2)^2} \quad \eta \equiv \alpha Z_2 \frac{\sqrt{1 + \tilde{p}_e^2}}{\tilde{p}_e} \quad \alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137}$$

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 $p_e = m_e c \tilde{p}_e$ finite. That gives $\xi = 1$ and

$$F(p_e, Z_2) = \frac{2\pi\eta}{1 - e^{-2\pi\eta}} \quad \eta = \frac{e^2}{4\pi\epsilon_0\hbar} \frac{m_e}{p_e}$$

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

$$f = \int_{\tilde{p}_e=0}^{\tilde{p}_{e,\max}} \tilde{p}_e^2 \tilde{p}_{\bar{\nu}}^2 F d\tilde{p}_e \quad (\text{A.118})$$

This quantity is known as the “Fermi integral.” Typical values are shown in figure 11.50.

Note that f also depends a bit on the mass number through the nuclear radius in F . The figure used

$$A = 1.82 + 1.9 Z_2 + 0.01271 Z_2^2 - 0.00006 Z_2^3 \quad (\text{A.119})$$

for beta-minus decay and

$$A = -1.9 + 1.96 Z_2 + 0.0079 Z_2^2 - 0.00002 Z_2^3 \quad (\text{A.120})$$

for beta-plus decay, [16].

A.111.7 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

$$H_{21} = \langle 1\bar{\nu}, \vec{p}_{\bar{\nu}} | \langle 0e, 1s | \psi_{2,\text{nuc}} | H' \psi_{1,\text{nuc}} | 1e, 1s \rangle | 0\bar{\nu}, \vec{p}_{\bar{\nu}} \rangle$$

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

$$H_{21} = \frac{g}{\sqrt{\mathcal{V}}} \langle \psi_{2,\text{nuc}} | \sum_{i=1}^A g h_i \psi_{100}(\vec{r}_i) e^{i\vec{p}_{\bar{\nu}} \cdot \vec{r}_i / \hbar} \psi_{1,\text{nuc}} \rangle$$

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 3.2,

$$\psi_{100}(\vec{r}_i) \approx \frac{1}{\sqrt{\pi a_0^3}} \quad a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2 Z_1} = \frac{\hbar}{m_e c \alpha Z_1}$$

where α is the fine-structure constant.

The square Hamiltonian coefficient for ℓ th-forbidden decays then becomes

$$|H_{21}^\ell|^2 = \frac{g^2}{\mathcal{V}} \frac{m_e^3 c^3 \alpha^3 Z^3}{\pi \hbar^3} \frac{1}{(\ell!)^2} \left(\frac{p_\nu R}{\hbar} \right)^{2\ell} C_N^\ell$$

$$C_N^\ell = \overline{\left| \left\langle \psi_{2,\text{nuc}} \left| \sum_{i=1}^A h_i \left(\frac{\vec{p}_\nu \cdot \vec{r}_i}{p_\nu R} \right)^\ell \psi_{1,\text{nuc}} \right\rangle \right|^2}$$

The decay rate for electron capture is

$$\lambda = \frac{2\pi}{\hbar \varepsilon} |H_{21}^\ell|^2 \frac{\mathcal{V} p_\nu^2}{2\pi^2 \hbar^3} \Delta p_\nu \quad \Delta p_\nu = \frac{\varepsilon}{c}$$

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

$$\lambda = \frac{2}{\pi^2} \frac{g^2 m_e^4 c^2}{\hbar^6} \frac{m_e c^2}{\hbar} (\alpha Z)^3 \frac{\tilde{R}^{2\ell}}{(\ell!)^2} C_N^\ell \tilde{p}_\nu^{2\ell} \tilde{p}_\nu^2 \quad (\text{A.121})$$

where the \tilde{p}_ν indicate the neutrino momentum nondimensionalized with $m_e c$. Also,

$$\tilde{Q} \equiv \frac{Q}{m_e c^2} \quad \tilde{p}_\nu = \tilde{Q} \quad \tilde{R} \equiv \frac{m_e c R}{\hbar} \quad (\text{A.122})$$

A.112 Weisskopf estimates

This note is simply wrong for anything more than first-forbidden decays. It gets the idea across anyway, I think. I made a start to a much more solid and still basic derivation, but I got distracted.

This note derives the Weisskopf ballparks for the decay rate of gamma ray emission. The analysis follows similar ideas as the earlier one for the Fermi theory of beta decay in note {A.111}.

A.112.1 Very loose derivation

It will be assumed that the decay process does not take place in infinite space, but in a very large periodic box. That avoids the problem that momentum eigenfunctions cannot be normalized in infinite space. If the box is big enough, it should not make a difference for the final answer.

Consider now the decay of an excited nucleus in which a photon of linear momentum \vec{p} is released. The state after the decay can be written as

$$\psi_2 = \psi_{N2}|1\gamma, \vec{p}\rangle$$

where ψ_{N2} is the wave function of the nucleus after the decay and the Fock space ket should be read as “one photon in the state with momentum \vec{p} .” Similarly, the state before the decay can be written as

$$\psi_1 = \psi_{N1}|0\gamma, \vec{p}\rangle$$

where the Fock space ket should be read as “zero photons in the state with momentum \vec{p} .”

Next the Hamiltonian is needed. For simplicity, it will be assumed that the y -axis is aligned with the direction of propagation of the photon, and the z -axis with its direction of polarization. In that case, in the simplest case the Hamiltonian can be taken to be of the form, chapter 12.2,

$$H = H_N + H_p - \sum_{i=1}^Z e \hat{E}_z z_i - \sum_{i=1}^Z \frac{e}{2m_p} \hat{B}_x \hat{L}_x - \sum_{i=1}^A \frac{e}{2m_p} \hat{B}_x g_i \hat{S}_x$$

where H_N is the nuclear Hamiltonian, H_p the Hamiltonian giving the photon field energy, the third term is the energy of the charged protons in the electric field of the photon, the fourth the interaction between the orbital angular momentum of the protons with the magnetic field, and the final term is the interaction between the nucleon magnetic moments and the magnetic field. Also m_p is the proton mass, e the proton charge, and g_i is the g -factor of nucleon number i that determines its magnetic moment. It will be assumed that the decay is due to a single proton i that makes a transition to a lower energy level. Therefore, the sums will from now on be omitted. Note that if many protons contribute to the transition through a collective motion effect, the decay rate may be underestimated significantly.

It should be pointed out that for forbidden transitions, except M1 ones, the above expression for the Hamiltonian is *not* correct. The vector potential should be used, {A.47}. But for a ballpark, it does not really make a difference.

In the relativistic approach, the electric and magnetic fields are operators acting on the Fock space ket:

$$\hat{E}_z = i \sqrt{\frac{\hbar\omega}{2\epsilon_0\mathcal{V}}} (\hat{a}^\dagger e^{-ipy/\hbar} - \hat{a} e^{ipy/\hbar}) \quad \hat{B}_x = \frac{i}{c} \sqrt{\frac{\hbar\omega}{2\epsilon_0\mathcal{V}}} (\hat{a}^\dagger e^{-ipy/\hbar} - \hat{a} e^{ipy/\hbar})$$

where \hat{a}^\dagger is the photon creation operator and \hat{a} the photon annihilation operator. Also ω is the frequency of the photon, \mathcal{V} is the volume of the periodic box in

which the decay is assumed to take place, and the constant ϵ_0 is the permittivity of space, $8.85 \cdot 10^{-12} \text{ C}^2/\text{J m}$.

The decay process is governed by Hamiltonian coefficients. The first two give the energy before and after the decay:

$$E_1 \equiv H_{11} \equiv \langle \psi_1 | H \psi_1 \rangle \quad E_2 \equiv H_{22} \equiv \langle \psi_2 | H \psi_2 \rangle$$

They should be essentially the same on account of energy conservation. The reduction in rest mass energy of the nucleus, the Q -value, comes out as the energy E_p of the released photon and recoil kinetic energy of the nucleus. The recoil energy is small and will be ignored. Therefore

$$Q = m_{N1}c^2 - m_{N2}c^2 = E_p = pc = \hbar\omega \quad (\text{A.123})$$

where the relativistic energy relation (A.7) for a particle with zero rest mass was used to relate the photon's energy to its momentum p and the Planck-Einstein relation was used to relate it to its frequency ω .

The Hamiltonian coefficient that actually causes the decay is

$$H_{21} = \langle \psi_2 | H \psi_1 \rangle$$

If the wave functions are written out, it is

$$\langle 1\gamma, \vec{p} | \psi_{N2} | H \psi_{N1} | 0\gamma, \vec{p} \rangle$$

Now, kets with different numbers of particles are mutually orthonormal. Therefore, for a term in the Hamiltonian to give a nonzero contribution, it must contain a creation operator \hat{a}^\dagger to match up the right-hand ket to the left-hand one. That means that only two terms produce a contribution. The first is an interaction with the electric field:

$$H_{21,E} = -ie\sqrt{\frac{Q}{2\epsilon_0\mathcal{V}}}\langle \psi_{N2} | e^{-ipy_i/\hbar} z_i \psi_{N1} \rangle$$

while the second is an interaction with the magnetic field:

$$H_{21,M} = -\frac{i}{c} \frac{e}{2m_p} \sqrt{\frac{Q}{2\epsilon_0\mathcal{V}}} \langle \psi_{N2} | e^{-ipy_i/\hbar} (\hat{L}_x + g_i \hat{S}_x) \psi_{N1} \rangle$$

Normally the nuclear size is small compared to the wave length of the emitted photon, and therefore the argument in the exponentials is small. That allows these exponentials to be expanded into a Taylor series. The term of interest in the expansion is the first one for which H_{21} is nonzero. If l is the number of that term, counted from zero, then the exponential can be replaced by

$$e^{-ipy_i/\hbar} \implies \frac{(-i)^l}{l!} \left(\frac{py_i}{\hbar} \right)^l = \frac{(-i)^l}{l!} \left(\frac{QR}{\hbar c} \right)^l \left(\frac{y_i}{R} \right)^l$$

where R is the nuclear radius. Increasing values of l correspond to increasingly forbidden transitions. Physically they correspond to increasing orbital angular momentum of the photon. The reason is that in quantum mechanics angular momentum is related to angular variation, and the higher the power of y/R , the higher the angular variation. In fact, it can be inferred from the form of the spherical harmonics 3.2 that l can be identified with the azimuthal quantum number of the orbital angular momentum of the photon that is seen by the nucleus.

If the Taylor series approximation is substituted in the Hamiltonian coefficients, their magnitudes become

$$|H_{21,E}^l| = eR\sqrt{\frac{Q}{2\epsilon_0\mathcal{V}}}\frac{1}{l!}\left(\frac{QR}{\hbar c}\right)^l|\langle\psi_{N2}|(y_i/R)^l(z_i/R)\psi_{N1}\rangle|$$

$$|H_{21,M}^l| = \frac{\hbar}{c}\frac{e}{2m_p}\sqrt{\frac{Q}{2\epsilon_0\mathcal{V}}}\frac{1}{l!}\left(\frac{QR}{\hbar c}\right)^l|\langle\psi_{N2}|(y_i/R)^l\hbar^{-1}(\hat{L}_x + g_i\hat{S}_x)\psi_{N1}\rangle|$$

The final inner products are not easily estimated. However, they cannot exceed unity by any significant amount, because ψ_{N1} and ψ_{N2} are normalized wave functions that decay quickly beyond the nuclear radius and the factor multiplying ψ_{N1} is less than one within the nuclear radius. While the inner products should really be summed over the magnetic quantum numbers of ψ_{N2} , and over the two polarization directions of the photon, and averaged over its possible directions of propagation, as well as over the magnetic quantum numbers of ψ_{N1} , that should not raise the magnitude of the inner products above order one.

Unfortunately, the inner products can easily be much smaller than one. That happens if there is a large cancellation between opposite contributions in the inner product integrals. It really means that the nuclear state produced by the Hamiltonian decay operator has only a small probability of being found to be the correct final nuclear state. Predicting this probability is hard, however, so it will not be attempted. The inner product in electric decays will simply be taken to be 1. For magnetic decays, the inner product will be taken to be $\sqrt{10}$ to allow for the fact that the g factor and magnetic quantum number are likely to be somewhat more than one. In using these finite values for the inner products, it must be accepted that the Weisskopf estimates will greatly overestimate the decay rates for some nuclei. That are the nuclei for which the nuclear wave function produced by the decay Hamiltonian matches the correct final wave function poorly.

Accurate or not, estimates for the Hamiltonian coefficient H_{21} have been obtained. The next step is to estimate the total decay rate from them. There is more than one decay process possible, and the total decay rate is the sum of all these possibilities. For one, each direction of motion of the emitted photon

corresponds to a separate decay process. Also, the magnitude of the momentum of the emitted photon can vary to some small extent from one decay to the next.

At first, the latter seems to violate energy conservation, since the momentum of the photon is related to its energy by $E_p = pc$, and that energy should be given by the nuclear rest mass energy reduction Q . However, energy eigenstates are stationary. The very fact that decay occurs shows that the initial and final states are not really energy eigenstates when examined closely enough. There will be a slight uncertainty in energy, and as a result energy conservation is not exact either. Fermi's golden rule, {A.111}, says that this effect can be modeled by allowing for an amount ε of slop in energy conservation. This slop can be estimated from an energy-time relation $\varepsilon t_c \sim 2\pi\hbar$, where t_c is a typical time interval between collisions with the general surroundings. However, the precise value of the energy slop is not really important as it drops out of the final expression for the decay rate.

According to the golden rule, the decay rate to a single final state can be taken as

$$\lambda_{\text{single final state}} = \frac{2\pi}{\hbar\varepsilon} |H_{21}^l|^2$$

This is to be summed over all final states, allowing for the energy slop. Now the amount of momentum states in all directions and within a momentum range slop $\Delta p = \varepsilon/c$ is, as shown for beta decay in {A.111},

$$N = \frac{\mathcal{V}p^2}{2\pi^2\hbar^3 c} \frac{\varepsilon}{c}$$

The final decay rates can now be found by collecting the results above together. Multiply the decay rate to a single final state by the number of final states, substitute in the appropriate expression for the Hamiltonian coefficient H_{21}^l , and clean up, to get:

$$\begin{aligned} \lambda_E^l &= \frac{2}{(l!)^2} \alpha \left(\frac{QR}{\hbar c} \right)^{2(l+1)} \frac{Q}{\hbar} \\ \lambda_B^l &= \frac{5}{(l!)^2} \alpha \left(\frac{\hbar c}{m_p c^2 R} \right)^2 \left(\frac{QR}{\hbar c} \right)^{2(l+1)} \frac{Q}{\hbar} \end{aligned}$$

where $\alpha = e^2/4\pi\epsilon_0\hbar c \approx 1/137$ is the fine structure constant. These rates can be written in terms of the mass number instead of the nuclear radius using an approximation of the form $R \approx 1.23\sqrt[3]{A}$ fm.

It is conventional to indicate the electric transitions as $E\ell$ and magnetic ones as $M\ell$, where ℓ is the net angular momentum of the emitted photon. In terms of the above loose derivation, $\ell = l + 1$. Note however that normally physicists do not use the symbol ℓ but either L , because that is great for confusion with orbital angular momentum, or λ , since that is great for confusion with decay rates or the photon wave length.

A.112.2 Official loose derivation

The derived ballparks are as reasonable as any, even though the Hamiltonian used is only sufficiently accurate for $E1$ and $M1$ transitions. However, there are official values for the ballparks

The Weisskopf ballpark for the decay rate of electric multipole transitions is:

$$\lambda_{E\ell} = \alpha \frac{18(\ell+1)}{\ell[(2\ell+1)!!(\ell+3)]^2} \left(\frac{QR}{\hbar c} \right)^{2\ell} \frac{Q}{\hbar} \quad (\text{A.124})$$

The Moszkowski ballpark for the decay rate of magnetic multipole transitions is derived similarly. It is:

$$\lambda_{M\ell} = \alpha \frac{18(\ell+1)}{\ell[(2\ell+1)!!(\ell+2)]^2} \left(\frac{QR}{\hbar c} \right)^{2\ell} \frac{Q}{\hbar} \left(\frac{\hbar c}{m_p c^2 R} \right)^2 \left(\frac{1}{2} g_p \ell - \frac{\ell}{\ell+1} \right)^2 \quad (\text{A.125})$$

It may be noted that [7, p. 9-110] list $\frac{1}{4}g$ instead of $\frac{1}{2}g$ and use that in their plot. This is corrected in the second edition, [8, p. 9-178], but the Moszkowski plot has disappeared.

Since the Moszkowski unit is a rough ballpark anyway, the official Weisskopf magnetic unit makes a few additional simplifications to make it look more like the electric unit. The factor $\ell+2$ is replaced by $\ell+3$, and the final parenthetical factor is ballparked as 10. That gives

$$\lambda_{M\ell} = \alpha \frac{18(\ell+1)}{\ell[(2\ell+1)!!(\ell+3)]^2} \left(\frac{QR}{\hbar c} \right)^{2\ell} \frac{Q}{\hbar} \left(\frac{\sqrt{10} \hbar c}{m_p c^2 R} \right)^2 \quad (\text{A.126})$$

That makes it the same as the electric unit, save for the final factor which has a magnitude $.3/A^{2/3}$ or about 0.01 for a typical nucleus. Therefore, magnetic transitions are typically much slower than electric ones at the same dipole level.

Electric multipole transitions are often expressed as a fraction of this decay rate. That defines the “Weisskopf unit,” W.u. Note however that various sources may split off part of the expression for λ , producing different dimensions than seconds for the W.u. Including a barn (b) is popular. You do not want to know what that is.

It may be noted that [21] misstates all three formulae through two different errors. Both [8] and [21] do not replace the factor $\ell+2$ by $\ell+3$ in the Weisskopf magnetic transition estimate. Of course, the difference is negligible compared to replacing the parenthetical expression by 10, or compared to the orders of magnitude that the estimate is commonly off anyway. However, an additional source of confusion in physics has been achieved, and physicists should be commended for keeping a straight face while doing it.

A.113 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.

A.114 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 :

$$\psi_{\vec{n}} = \psi_{\vec{n},0} + \varepsilon \psi_{\vec{n},\varepsilon} + \varepsilon^2 \psi_{\vec{n},\varepsilon^2} + \dots$$

$$E_{\vec{n}} = E_{\vec{n},0} + \varepsilon E_{\vec{n},\varepsilon} + \varepsilon^2 E_{\vec{n},\varepsilon^2} + \dots$$

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:

$$\begin{aligned} \varepsilon^0 : \quad & (H_0 - E_{\vec{n},0}) \psi_{\vec{n},0} = 0 \\ \varepsilon^1 : \quad & (H_0 - E_{\vec{n},0}) \psi_{\vec{n},\varepsilon} = -H_\varepsilon \psi_{\vec{n},0} + E_{\vec{n},\varepsilon} \psi_{\vec{n},0} \\ \varepsilon^2 : \quad & (H_0 - E_{\vec{n},0}) \psi_{\vec{n},\varepsilon^2} = -H_\varepsilon \psi_{\vec{n},\varepsilon} + E_{\vec{n},\varepsilon} \psi_{\vec{n},\varepsilon} + E_{\vec{n},\varepsilon^2} \psi_{\vec{n},0} \\ \varepsilon^3 : \quad & (H_0 - E_{\vec{n},0}) \psi_{\vec{n},\varepsilon^3} = -H_\varepsilon \psi_{\vec{n},\varepsilon^2} + E_{\vec{n},\varepsilon} \psi_{\vec{n},\varepsilon^2} + E_{\vec{n},\varepsilon^2} \psi_{\vec{n},\varepsilon} + E_{\vec{n},\varepsilon^3} \psi_{\vec{n},0} \\ \vdots & \quad \dots \end{aligned}$$

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

$$(H_0 - E_{\vec{n},0}) \psi_{\vec{n},\dots} = \dots$$

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

$$\psi_{\vec{n}} = \psi_{\vec{n},0} + \psi_{\vec{n},1} + \psi_{\vec{n},2} + \dots$$

$$E_{\vec{n}} = E_{\vec{n},0} + E_{\vec{n},1} + E_{\vec{n},2} + \dots$$

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:

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},0} = 0 \quad (\text{A.127})$$

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},1} = -H_1\psi_{\vec{n},0} + E_{\vec{n},1}\psi_{\vec{n},0} \quad (\text{A.128})$$

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},2} = -H_1\psi_{\vec{n},1} + E_{\vec{n},1}\psi_{\vec{n},1} + E_{\vec{n},2}\psi_{\vec{n},0} \quad (\text{A.129})$$

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},3} = -H_1\psi_{\vec{n},2} + E_{\vec{n},1}\psi_{\vec{n},2} + E_{\vec{n},2}\psi_{\vec{n},1} + E_{\vec{n},3}\psi_{\vec{n},0} \quad (\text{A.130})$$

...

Now consider each of these equations in turn. First, (A.127) 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 $\psi_{\vec{n},0}$ of the *unperturbed* Hamiltonian H_0 . In particular, to solve (A.128), write

$$\psi_{\vec{n},1} = \sum_{\vec{n} \neq \vec{n}} c_{\vec{n},1} \psi_{\vec{n},0}$$

where the coefficients $c_{\vec{n},1}$ 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 $\psi_{\vec{n},2}, \psi_{\vec{n},3}, \dots$, even if it means that the resulting wave function will no longer be normalized.)

The problem (A.128) becomes

$$\sum_{\vec{n} \neq \vec{n}} c_{\vec{n},1} (E_{\vec{n},0} - E_{\vec{n},0}) \psi_{\vec{n},0} = -H_1 \psi_{\vec{n},0} + E_{\vec{n},1} \psi_{\vec{n},0}$$

where the left hand side was cleaned up using the fact that the $\psi_{\vec{n},0}$ 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:

$$0 = -H_{\vec{n}\vec{n},1} + E_{\vec{n},1} \quad H_{\vec{n}\vec{n},1} = \langle \psi_{\vec{n},0} | H_1 \psi_{\vec{n},0} \rangle$$

And that is exactly the first order correction to the energy claimed in subsection 12.1.1; $E_{\vec{n},1}$ 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 $c_{\vec{n},1}$, 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 $\langle \psi_{\vec{n},0} |$ 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

$$c_{\underline{\vec{n}},1}(E_{\underline{\vec{n}},0} - E_{\vec{n},0}) = -H_{\underline{\vec{n}}\vec{n},1} \quad \text{for } \underline{\vec{n}} \neq \vec{n} \quad H_{\underline{\vec{n}}\vec{n},1} = -\langle \psi_{\underline{\vec{n}},0} | H_1 \psi_{\vec{n},0} \rangle$$

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 $\psi_{\vec{n},0}$ that have the same energy $E_{\vec{n},0}$ as the eigenfunction $\psi_{\underline{\vec{n}},0}$ 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

$$\text{all } H_{\vec{n}_i \vec{n}_j,1} \quad \text{with } E_{\vec{n}_i,0} = E_{\vec{n}_j,0} = E_{\vec{n},0}$$

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, \dots) of the submatrix, there is a corresponding replacement unperturbed eigenfunction

$$c_1 \psi_{\vec{n}_1,0,\text{old}} + c_2 \psi_{\vec{n}_2,0,\text{old}} + \dots$$

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

$$\psi_{\vec{n},1} = - \sum_{\substack{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0} \\ \underline{\vec{n}} \neq \vec{n}}} \frac{H_{\underline{\vec{n}}\vec{n},1}}{E_{\vec{n},0} - E_{\underline{\vec{n}},0}} \psi_{\underline{\vec{n}},0} + \sum_{\substack{E_{\underline{\vec{n}},0} = E_{\vec{n},0} \\ \underline{\vec{n}} \neq \vec{n}}} c_{\underline{\vec{n}},1} \psi_{\underline{\vec{n}},0}$$

The entire process repeats for higher order. In particular, to second order (A.129) gives, writing $\psi_{\vec{n},2}$ also in terms of the unperturbed eigenfunctions,

$$\begin{aligned} \sum_{\underline{\vec{n}}} c_{\underline{\vec{n}},2} (E_{\underline{\vec{n}},0} - E_{\vec{n},0}) \psi_{\underline{\vec{n}},0} &= \sum_{\substack{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}}} \frac{H_{\underline{\vec{n}}\vec{n},1}}{E_{\vec{n},0} - E_{\underline{\vec{n}},0}} (H_1 - E_{\vec{n},1}) \psi_{\underline{\vec{n}},0} \\ &\quad - \sum_{\substack{E_{\underline{\vec{n}},0} = E_{\vec{n},0} \\ \underline{\vec{n}} \neq \vec{n}}} c_{\underline{\vec{n}},1} (H_1 - E_{\vec{n},1}) \psi_{\underline{\vec{n}},0} + E_{\vec{n},2} \psi_{\vec{n},0} \end{aligned}$$

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),

$$0 = \sum_{\substack{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}}} \frac{H_{\underline{\vec{n}}\vec{n},1} H_{\vec{n}\underline{\vec{n}},1}}{E_{\vec{n},0} - E_{\underline{\vec{n}},0}} + E_{\vec{n},2}$$

This gives the second order change in the energy stated in subsection 12.1.1, 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 $\langle \psi_{\vec{n},0} |$ 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 $H_{\vec{n}\vec{n},1}$ 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

$$\text{all } \sum_{\substack{E_{\underline{\vec{n}},0} \neq E_{\vec{n},0}}} \frac{H_{\vec{n}_i \vec{n}_j,1} H_{\underline{\vec{n}} \vec{n}_j,1}}{E_{\vec{n},0} - E_{\underline{\vec{n}},0}} \quad \text{with} \quad E_{\vec{n}_i,0} = E_{\vec{n}_j,0} = E_{\vec{n},0} \quad E_{\vec{n}_i,1} = E_{\vec{n}_j,1} = E_{\vec{n},1}$$

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 $\psi_{\vec{n},0}$ 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.

A.115 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 3.2. Following the convention in perturbation theory to append a subscript zero to the

unperturbed state, it can be summarized as:

$$H_0\psi_{100,0} = E_{100,0}\psi_{100,0} \quad H_0 = -\frac{\hbar^2}{2m_e}\nabla^2 + V \quad \psi_{100,0} = \frac{1}{\sqrt{\pi a_0^3}}e^{-r/a_0}$$

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 (12.5),

$$(H_0 - E_{100,0})\psi_{100,1} = -(H_1 - E_{100,1})\psi_{100,0} \quad H_1 = eE_{\text{ext}}z$$

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:

$$(H_0 - E_{100,0})(f\psi_{100,0}) = -eE_{\text{ext}}z\psi_{100,0} \quad H_0 = -\frac{\hbar^2}{2m_e}\nabla^2 + V$$

Differentiating out the Laplacian ∇^2 of the product $f\psi_{100,0}$ into individual terms using Cartesian coordinates, the equation becomes

$$f(H_0 - E_{100,0})\psi_{100,0} - \frac{\hbar^2}{m_e}(\nabla f) \cdot (\nabla\psi_{100,0}) - \frac{\hbar^2}{2m_e}(\nabla^2 f)\psi_{100,0} = -eE_{\text{ext}}z\psi_{100,0}$$

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. [28, 20.74, 20.82],

$$\nabla f = \frac{\partial f}{\partial r}\hat{r} + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\theta} + \frac{1}{r \sin \theta}\frac{\partial f}{\partial \phi}\hat{\phi} \quad \nabla\psi_{100,0} = -\psi_{100,0}\frac{1}{a_0}\hat{r}$$

Substituting that into the equation, it reduces to

$$\frac{\hbar^2}{m_e}\left(\frac{1}{a_0}\frac{\partial f}{\partial r} - \frac{1}{2}\nabla^2 f\right)\psi_{100,0} = -eE_{\text{ext}}z\psi_{100,0}$$

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:

$$f = -\frac{m_e e E_{\text{ext}} a_0}{2\hbar^2} r^2 \cos \theta + \dots$$

but then, using [28, 20.83], the $\nabla^2 f$ term in the left hand side produces an $eE_{\text{ext}}a_0 \cos \theta$ term. So add another term to f for its r -derivative to eliminate it:

$$f = -\frac{m_e e E_{\text{ext}} a_0}{2\hbar^2} r^2 \cos \theta - \frac{m_e e E_{\text{ext}} a_0^2}{\hbar^2} r \cos \theta$$

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

$$\psi_{100,1} = -\frac{m_e e E_{\text{ext}} a_0}{2\hbar^2 \sqrt{\pi a_0^3}} (r^2 + 2a_0 r) e^{-r/a_0} \cos \theta$$

(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 (12.5), the second order energy change can be found as

$$E_{100,2} = \langle \psi_{100,0} | H_1 \psi_{100,1} \rangle \quad H_1 = eE_{\text{ext}} r \cos \theta$$

Doing the inner product integration in spherical coordinates produces

$$E_{100,2} = -\frac{9m_e e^2 E_{\text{ext}}^2 a_0^4}{4\hbar^2}$$

A.116 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 3.2. The derivation is based on the relativistic Dirac equation from chapter 10.2 and uses nontrivial linear algebra.

According to the Dirac equation, the relativistic Hamiltonian and wave function take the form

$$H_D = m_e c^2 \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \sum_{i=1}^3 c \hat{p}_i \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix} + V \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \vec{\psi}_D = \begin{pmatrix} \vec{\psi}^p \\ \vec{\psi}^n \end{pmatrix}$$

where m_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 10.1.9. 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 4.5.1, 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 note {A.114}. 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 note {A.114}, take the Hamiltonian apart into successive powers of $1/c$ as $H_D = H_{D,0} + H_{D,1} + H_{D,2}$ with

$$H_{D,0} = \begin{pmatrix} m_e c^2 & 0 \\ 0 & -m_e c^2 \end{pmatrix} \quad H_{D,1} = \sum_{i=1}^3 \begin{pmatrix} 0 & c \hat{p}_i \sigma_i \\ c \hat{p}_i \sigma_i & 0 \end{pmatrix} \quad H_{D,2} = \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix}$$

and similarly for the wave function vector:

$$\vec{\psi}_D = \begin{pmatrix} \vec{\psi}_0^p \\ \vec{\psi}_0^n \end{pmatrix} + \begin{pmatrix} \vec{\psi}_1^p \\ \vec{\psi}_1^n \end{pmatrix} + \begin{pmatrix} \vec{\psi}_2^p \\ \vec{\psi}_2^n \end{pmatrix} + \begin{pmatrix} \vec{\psi}_3^p \\ \vec{\psi}_3^n \end{pmatrix} + \begin{pmatrix} \vec{\psi}_4^p \\ \vec{\psi}_4^n \end{pmatrix} + \dots$$

and the energy:

$$E_D = E_{D,0} + E_{D,1} + E_{D,2} + E_{D,3} + E_{D,4} + \dots$$

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 note {A.114}:

$$\begin{aligned} c^2 : \quad & \left[\begin{pmatrix} m_e c^2 & 0 \\ 0 & -m_e c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_0^p \\ \vec{\psi}_0^n \end{pmatrix} = 0 \\ c^1 : \quad & \left[\begin{pmatrix} m_e c^2 & 0 \\ 0 & -m_e c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_1^p \\ \vec{\psi}_1^n \end{pmatrix} = \\ & - \left[\sum_{i=1}^3 \begin{pmatrix} 0 & c \hat{p}_i \sigma_i \\ c \hat{p}_i \sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_0^p \\ \vec{\psi}_0^n \end{pmatrix} \end{aligned}$$

$$c^0 : \quad \left[\begin{pmatrix} m_e c^2 & 0 \\ 0 & -m_e c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_2^p \\ \vec{\psi}_2^n \end{pmatrix} =$$

$$- \left[\sum_{i=1}^3 \begin{pmatrix} 0 & c\hat{p}_i \sigma_i \\ c\hat{p}_i \sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_1^p \\ \vec{\psi}_1^n \end{pmatrix}$$

$$- \left[\begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} - \begin{pmatrix} E_{D,2} & 0 \\ 0 & E_{D,2} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_0^p \\ \vec{\psi}_0^n \end{pmatrix}$$

$$c^{-1} : \quad \left[\begin{pmatrix} m_e c^2 & 0 \\ 0 & -m_e c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_3^p \\ \vec{\psi}_3^n \end{pmatrix} =$$

$$- \left[\sum_{i=1}^3 \begin{pmatrix} 0 & c\hat{p}_i \sigma_i \\ c\hat{p}_i \sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_2^p \\ \vec{\psi}_2^n \end{pmatrix}$$

$$- \left[\begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} - \begin{pmatrix} E_{D,2} & 0 \\ 0 & E_{D,2} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_1^p \\ \vec{\psi}_1^n \end{pmatrix}$$

$$+ \begin{pmatrix} E_{D,3} & 0 \\ 0 & E_{D,3} \end{pmatrix} \begin{pmatrix} \vec{\psi}_0^p \\ \vec{\psi}_0^n \end{pmatrix}$$

$$c^{-2} : \quad \left[\begin{pmatrix} m_e c^2 & 0 \\ 0 & -m_e c^2 \end{pmatrix} - \begin{pmatrix} E_{D,0} & 0 \\ 0 & E_{D,0} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_4^p \\ \vec{\psi}_4^n \end{pmatrix} =$$

$$- \left[\sum_{i=1}^3 \begin{pmatrix} 0 & c\hat{p}_i \sigma_i \\ c\hat{p}_i \sigma_i & 0 \end{pmatrix} - \begin{pmatrix} E_{D,1} & 0 \\ 0 & E_{D,1} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_3^p \\ \vec{\psi}_3^n \end{pmatrix}$$

$$- \left[\begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} - \begin{pmatrix} E_{D,2} & 0 \\ 0 & E_{D,2} \end{pmatrix} \right] \begin{pmatrix} \vec{\psi}_2^p \\ \vec{\psi}_2^n \end{pmatrix}$$

$$+ \begin{pmatrix} E_{D,3} & 0 \\ 0 & E_{D,3} \end{pmatrix} \begin{pmatrix} \vec{\psi}_1^p \\ \vec{\psi}_1^n \end{pmatrix} + \begin{pmatrix} E_{D,4} & 0 \\ 0 & E_{D,4} \end{pmatrix} \begin{pmatrix} \vec{\psi}_0^p \\ \vec{\psi}_0^n \end{pmatrix}$$

$$c^{-3} : \quad \dots$$

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

$$E_{D,0} = m_e c^2 \quad \vec{\psi}_0^p = \text{still arbitrary} \quad \vec{\psi}_0^n = 0$$

Plug that into the order c^1 equation to give, for top and bottom subvectors

$$0 = E_{D,1}\vec{\psi}_0^p - 2m_e c^2 \vec{\psi}_1^n = -\sum_i c \hat{p}_i \sigma_i \vec{\psi}_0^p$$

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

$$E_{D,1} = 0 \quad \vec{\psi}_1^p = \text{still arbitrary} \quad \vec{\psi}_1^n = \sum_j \frac{1}{2m_e c} \hat{p}_j \sigma_j \vec{\psi}_0^p$$

where the summation index i was renamed to j to avoid ambiguity later.

Plug all that in the order c^0 equation to give

$$0 = -\frac{1}{2m_e} \sum_i \sum_j \hat{p}_i \hat{p}_j \sigma_i \sigma_j \vec{\psi}_0^p - V \vec{\psi}_0^p + E_{D,2} \vec{\psi}_0^p \quad \vec{\psi}_2^n = \sum_j \frac{1}{2m_e c} \hat{p}_j \sigma_j \vec{\psi}_1^p$$

The first of these two equations is the non-relativistic Hamiltonian eigenvalue problem of chapter 3.2. 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

$$\frac{1}{2m_e} \sum_i \sum_j \hat{p}_i \hat{p}_j \sigma_i \sigma_j + V = \frac{1}{2m_e} \sum_i \hat{p}_i^2 + V \equiv H_0$$

where H_0 is the nonrelativistic hydrogen Hamiltonian of chapter 3.2.

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 non-relativistic energy eigenfunction with that energy:

$$E_{D,2} = E_n \quad \vec{\psi}_0^p = \sum_l \sum_m c_{lm+} \psi_{nlm} \uparrow + \sum_l \sum_m c_{lm-} \psi_{nlm} \downarrow$$

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 3.2 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

$$\vec{\psi}_3^n = \sum_j \frac{1}{2m_e c} \hat{p}_j \sigma_j \vec{\psi}_2^p + \frac{V - E_n}{2m_e c^2} \sum_j \frac{1}{2m_e c} \hat{p}_j \sigma_j \vec{\psi}_0^p$$

To order c^{-2} , you get for the top vector

$$0 = -(H_0 - E_n)\vec{\psi}_2^p - \sum_i \sum_j \hat{p}_i \sigma_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_j \sigma_j \vec{\psi}_0^p + E_{D,4} \vec{\psi}_0^p$$

and that determines the approximate relativistic energy correction.

Now recall from note {A.114} that if you do a nonrelativistic expansion of an eigenvalue problem $(H_0 + H_1)\psi = E\psi$, the equations to solve are (A.127) and (A.128);

$$(H_0 - E_{\vec{n},0})\psi_{\vec{n},0} = 0 \quad (H_0 - E_{\vec{n},0})\psi_{\vec{n},1} = -(H_1 - E_{\vec{n},1})\psi_{\vec{n},0}$$

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

$$H_1 \stackrel{?}{=} \sum_i \sum_j \hat{p}_i \sigma_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_j \sigma_j$$

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 $j = \bar{i}$, where \bar{i} is the number following i in the cyclic sequence $\dots 123123 \dots$, and it involves a term $j = \bar{i}$ where \bar{i} precedes i in the sequence. So the Dirac operator falls apart into three pieces:

$$H_1 \stackrel{?}{=} \sum_i \hat{p}_i \sigma_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_i \sigma_i + \sum_i \hat{p}_i \sigma_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_{\bar{i}} \sigma_{\bar{i}} + \sum_i \hat{p}_i \sigma_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_{\bar{i}} \sigma_{\bar{i}}$$

or using the properties of the Pauli matrices that $\sigma_i \sigma_i = 1$, $\sigma_i \sigma_{\bar{i}} = i\sigma_{\bar{i}}$, and $\sigma_i \sigma_{\bar{i}} = -i\sigma_{\bar{i}}$ for any i ,

$$H_1 \stackrel{?}{=} \sum_i \hat{p}_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_i + i \sum_i \hat{p}_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_{\bar{i}} \sigma_{\bar{i}} - i \sum_i \hat{p}_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_{\bar{i}} \sigma_{\bar{i}} \quad (\text{A.131})$$

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:

$$H_{1,\text{spin-orbit}} = i \sum_i \hat{p}_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_{\bar{i}} \sigma_i - i \sum_i \hat{p}_{\bar{i}} \frac{V - E_n}{4m_e^2 c^2} \hat{p}_i \sigma_i$$

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;

$$H_{1,\text{spin-orbit}} = \frac{i}{4m_e^2 c^2} \sum_i [\hat{p}_i (V - E_n) \hat{p}_{\bar{i}} - \hat{p}_{\bar{i}} (V - E_n) \hat{p}_i] \sigma_i$$

Now according to the generalized canonical commutator of chapter 3.4.4:

$$\hat{p}_i (V - E_n) = (V - E_n) \hat{p}_i - i\hbar \frac{\partial(V - E_n)}{\partial r_i}$$

where E_n is a constant that produces a zero derivative. So $\hat{p}_{\bar{i}}$, respectively \hat{p}_i 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, $(V - E_n) \hat{p}_i \hat{p}_{\bar{i}}$ and $-(V - E_n) \hat{p}_{\bar{i}} \hat{p}_i$ cancel since linear momentum operators commute. What is left are just the added derivative terms:

$$H_{1,\text{spin-orbit}} = \frac{\hbar}{4m_e^2 c^2} \sum_i \left[\frac{\partial V}{\partial r_{\bar{i}}} \hat{p}_{\bar{i}} - \frac{\partial V}{\partial r_i} \hat{p}_i \right] \sigma_i$$

Note that the errant eigenvalue E_n mercifully dropped out. Now the hydrogen Hamiltonian 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

$$H_{1,\text{spin-orbit}} = -\frac{\hbar V}{4m_e^2 c^2 r^2} \sum_i [r_{\bar{i}} \hat{p}_{\bar{i}} - r_i \hat{p}_i] \sigma_i$$

The term between the square brackets can be recognized as the i th component of the angular momentum operator; also the Pauli spin matrix σ_i is defined as $\hat{S}_i / \frac{1}{2}\hbar$, so

$$H_{1,\text{spin-orbit}} = -\frac{V}{2m_e^2 c^2 r^2} \sum_i \hat{L}_i \hat{S}_i$$

Get rid of c^2 using $|E_1| = \frac{1}{2}\alpha^2 m_e c^2$, of V using $V = -2|E_1|a_0/r$, and m_e using $|E_1| = \hbar^2/2m_e a_0^2$ to get the spin-orbit interaction as claimed in the section on fine structure.

That leaves the term

$$\sum_i \hat{p}_i \frac{V - E_n}{4m_e^2 c^2} \hat{p}_i$$

in (A.131). Since $V = H_0 - \hat{p}^2/2m_e$, it can be written as

$$\sum_i \hat{p}_i \frac{H_0 - E_n}{4m_e^2 c^2} \hat{p}_i - \frac{(\hat{p}^2)^2}{8m_e^3 c^2}$$

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,

$$H_{1,\text{Darwin}} \stackrel{?}{=} \sum_i \hat{p}_i \frac{H_0 - E_n}{4m_e^2 c^2} \hat{p}_i$$

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

$$H_{\vec{n}\vec{n},1,\text{Darwin}} \equiv \langle \psi_{\vec{n},0} | H_{1,\text{Darwin}} \psi_{\vec{n},0} \rangle = \frac{1}{4m_e^2 c^2} \sum_i \langle \psi_{\vec{n},0} | \hat{p}_i (H_0 - E_n) \hat{p}_i \psi_{\vec{n},0} \rangle$$

for any arbitrary pair of unperturbed hydrogen energy eigenfunctions $\psi_{\vec{n},0}$ and $\psi_{\vec{n},0}$ with energy E_n . To see what that means, the leading Hermitian operator \hat{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:

$$H_{\vec{n}\vec{n},1,\text{Darwin}} = \frac{1}{8m_e^2 c^2} \sum_i \left(\langle \hat{p}_i \psi_{\vec{n},0} | (H_0 - E_n) \hat{p}_i \psi_{\vec{n},0} \rangle + \langle (H_0 - E_n) \hat{p}_i \psi_{\vec{n},0} | \hat{p}_i \psi_{\vec{n},0} \rangle \right)$$

Now if you simply swap the order of the factors in $(H_0 - E_n) \hat{p}_i$ in this expression, you get zero, because both eigenfunctions have energy E_n . However, swapping the order of $(H_0 - E_n) \hat{p}_i$ brings in the generalized canonical commutator $[V, \hat{p}_i]$ that equals $i\hbar \partial V / \partial r_i$. Therefore, writing out the remaining inner product you get

$$H_{\vec{n}\vec{n},1,\text{Darwin}} = \frac{-\hbar^2}{8m_e^2 c^2} \sum_i \int_{\text{all } \vec{r}} \frac{\partial V}{\partial r_i} \frac{\partial \psi_{\vec{n},0}^* \psi_{\vec{n},0}}{\partial r_i} d^3 \vec{r}$$

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

$$H_{\vec{n}\vec{n},1,\text{Darwin}} = \frac{-\hbar^2}{8m_e^2 c^2} \sum_i \int_{\text{all } \vec{r}} \left[\frac{\partial}{\partial r_i} \left(\frac{\partial V}{\partial r_i} \psi_{\vec{n},0}^* \psi_{\vec{n},0} \right) - \frac{\partial^2 V}{\partial r_i^2} \psi_{\vec{n},0}^* \psi_{\vec{n},0} \right] d^3 \vec{r}$$

and the first term integrates away since $\psi_{\vec{n},0}^* \psi_{\vec{n},0}$ 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 (10.23) 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

$$H_{1,\text{Darwin}} = \frac{\hbar^2 e^2}{8m_e^2 c^2 \epsilon_0} \delta^3(\vec{r})$$

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.

A.117 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 10.21 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 x,y -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 10.21, 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 $\vec{d} = d_x \hat{i} + d_z \hat{k}$.

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 10.4

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 10.18, 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 $\vec{\phi}' = q' d_c \hat{j}$, where q' is the electric charge per unit length. According to table 10.3, a single electric charge along the z -axis creates an electric field whose potential is

$$\varphi = \frac{q'}{2\pi\epsilon_0} \ln \frac{1}{r} = -\frac{q'}{4\pi\epsilon_0} \ln(x^2 + y^2)$$

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

$$\varphi = -\frac{q'}{4\pi\epsilon_0} \ln(x^2 + (y - d)^2) + \frac{q'}{4\pi\epsilon_0} \ln(x^2 + y^2)$$

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 $\vec{\phi}'$ is $q'_m d_x v / c^2 \hat{j}$. And since in this coordinate system the magnetic dipole moment

is $\vec{\mu}' = q'_m(d_x\hat{i} + d_z\hat{k})$ and the velocity $v\hat{k}$, it follows that the generated electric dipole strength is

$$\vec{\phi}' = -\vec{\mu}' \times \vec{v}/c^2$$

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 $-\vec{\phi} \cdot \vec{E}$ where \vec{E} is the electric field of the nucleus, $e\vec{r}/4\pi\epsilon_0 r^3$ according to table 10.3. 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:

$$-\frac{e}{4\pi\epsilon_0 c^2} \frac{1}{r^3} \vec{\mu} \cdot (\vec{r} \times \vec{v}) = -\frac{e}{4\pi\epsilon_0 c^2 m_e} \frac{1}{r^3} \vec{\mu} \cdot \vec{L}$$

To get the correct spin-orbit interaction, the magnetic dipole moment $\vec{\mu}$ used in this expression must be the classical one, $-e\vec{S}/2m_e$. 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

$$\frac{e^2}{8\pi\epsilon_0 m_e^2 c^2} \frac{1}{r^3} \vec{S} \cdot \vec{L}$$

Getting rid of c^2 using $|E_1| = \frac{1}{2}\alpha^2 m_e c^2$, of e^2/ϵ_0 using $e^2/4\pi\epsilon_0 = 2|E_1|a_0$, and of m_e using $|E_1| = \hbar^2/2m_e a_0^2$, the claimed expression for the spin-orbit energy is found.

A.118 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 derived are:

$$\begin{aligned} \langle \psi_{nlm} | (a_0/r)^3 \psi_{nlm} \rangle &= \frac{1}{l(l+\frac{1}{2})(l+1)n^3} \\ \langle \psi_{nlm} | (a_0/r)^2 \psi_{nlm} \rangle &= \frac{1}{(l+\frac{1}{2})n^3} \\ \langle \psi_{nlm} | (a_0/r) \psi_{nlm} \rangle &= \frac{1}{n^2} \\ \langle \psi_{nlm} | 1 \psi_{nlm} \rangle &= 1 \\ \langle \psi_{nlm} | (r/a_0) \psi_{nlm} \rangle &= \frac{3n^2 - l(l+1)}{2} \\ \langle \psi_{nlm} | (r/a_0)^2 \psi_{nlm} \rangle &= \frac{n^2(5n^2 - 3l(l+1) + 1)}{2} \\ \dots \end{aligned} \tag{A.132}$$

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 3.2,

$$H = -\frac{\hbar^2}{2m_e r^2} \left\{ \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right\} - \frac{e^2}{4\pi\epsilon_0 r} \frac{1}{r}$$

Its energy eigenfunctions of given square and z angular momentum and their energy are

$$\psi_{nlm} = R_{nl}(r) Y_l^m(\theta, \phi) \quad E_n = -\frac{\hbar^2}{2n^2 m_e a_0^2} \quad a_0 = \frac{4\pi\epsilon_0 \hbar^2}{m_e e^2}$$

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 $l(l + 1)$:

$$H_{\text{DT}} = -\frac{\hbar^2}{2m_e r^2} \left\{ \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) - l(l + 1) \right\} - \frac{e^2}{4\pi\epsilon_0} \frac{1}{r}$$

The reason is that the angular derivatives are essentially the square angular momentum operator of chapter 3.1.3. 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 12.1.1:

$$\frac{\partial E_{n,\text{DT}}}{\partial l} = \left\langle \psi_{nlm} \left| \frac{\partial H_{\text{DT}}}{\partial l} \right| \psi_{nlm} \right\rangle$$

(Yes, the eigenfunctions ψ_{nlm} are good, because the purely radial H_{DT} commutes with both \hat{L}_z and \hat{L}^2 , which are angular derivatives.) Substituting in the dirty trick Hamiltonian,

$$\frac{\partial E_{n,\text{DT}}}{\partial l} = \frac{\hbar^2(2l + 1)}{2m_e a_0^2} \left\langle \psi_{nlm} \left| \left(\frac{a_0}{r} \right)^2 \right| \psi_{nlm} \right\rangle$$

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_{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 $H_{\text{DT}}R_{nl} = E_{\text{DT}}R_{nl}$. If you skim back through chapter 3.2 and its note, you see that that eigenvalue problem was solved in note A.17. 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 $E_{n,\text{DT}} = -\hbar^2/2n^2m_e a_0^2$. 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 $dn/dl = 1$, and then according to the chain rule the derivative of E_{DT} is $\hbar^2/n^3 m_e a_0^2$. Substitute it in and there you have that nasty expectation value as given in (A.132).

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:”

$$4(q + 1)\langle q \rangle - 4n^2(2q + 1)\langle q - 1 \rangle + n^2q[(2l + 1)^2 - q^2]\langle q - 2 \rangle = 0 \quad (\text{A.133})$$

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 (A.132). It may be noted that this can instead be derived from the virial theorem of chapter 6.1.7, or from the Hellmann-Feynman theorem by differentiating the hydrogen Hamiltonian with respect to

the charge e . Substituting in $q = 1, 2, \dots$ produces the expectation values for $r/a_0, (r/a_0)^2, \dots$. 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 $q = -2, -3, \dots$ 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 , {A.17}, 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

$$\langle q \rangle \equiv \langle \psi_{nlm} | (r/a_0)^q \psi_{nlm} \rangle = \int_{\text{all } \vec{r}} (r/a_0)^q |\psi_{nlm}|^2 d^3\vec{r} = \int_{\text{all } \vec{r}} (r/a_0)^q |R_{nl} Y_l^m|^2 d^3\vec{r}$$

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

$$\langle q \rangle = \int_{r=0}^{\infty} (r/a_0)^q R_{nl}^2 r^2 dr$$

To simplify the notations, a non-dimensional 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

$$\langle q \rangle = \int_0^{\infty} \rho^q f^2 d\rho$$

To further shorten the notations, from now on the limits of integration and $d\rho$ 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 note {A.17}, 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

$$f'' = \left[\frac{1}{n^2} - \frac{2}{\rho} + \frac{l(l+1)}{\rho^2} \right] f \quad (\text{A.134})$$

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 :

$$\int \rho^q f'' f = \frac{1}{n^2} \langle q \rangle - 2\langle q-1 \rangle + l(l+1)\langle q-2 \rangle \quad (\text{A.135})$$

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

$$\int \rho^q f f' = \int \rho^q \left(\frac{1}{2} f^2 \right)' = \rho^q \frac{1}{2} f^2 \Big| - \int q \rho^{q-1} \frac{1}{2} f^2,$$

the latter from integration by parts, it follows that

$$\int \rho^q f f' = \frac{1}{2} \rho^q f^2 \Big| - \frac{q}{2} \langle q-1 \rangle \quad (\text{A.136})$$

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 (A.136). 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 $f' f'$:

$$\int \rho^q f f'' = \rho^q f f' \Big| - \frac{q}{2} \rho^{q-1} f^2 \Big| + \frac{q(q-1)}{2} \langle q-2 \rangle - \frac{\rho^{q+1}}{q+1} f'^2 \Big| + 2 \int \frac{\rho^{q+1}}{q+1} f' f'' \quad (\text{A.137})$$

In the final integral, according to the differential equation (A.134), 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 (A.136), 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 (A.137) and then equating that to (A.135) 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 $q \rightarrow -1$ to cover that case.

A.119 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.

Bibliography

- [1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, third edition, 1965.
- [2] A. Aharoni. *Introduction to the Theory of Ferromagnetism*. Oxford University Press, second edition, 2000.
- [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.
- [4] R. Baierlein. *Thermal Physics*. Cambridge University Press, Cambridge, UK, 1999.
- [5] C.A. Bertulani. *Nuclear Physics in a Nutshell*. Princeton University Press, 2007.
- [6] S.H. Chue. *Thermodynamics : a rigorous postulatory approach*. Wiley, 1977.
- [7] E.U. Condon and H. Odishaw, editors. *Handbook of Physics*. McGraw-Hill, 1958.
- [8] E.U. Condon and H. Odishaw, editors. *Handbook of Physics*. McGraw-Hill, 2nd edition, 1967.
- [9] E.A. Desloge. *Thermal Physics*. Holt, Rinehart and Winston, New York, 1968.
- [10] A.M. Ellis. Spectroscopic selection rules: The role of photon states. *J. Chem. Educ.*, 76:1291–1294, 1999.
- [11] 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.

- [12] R. P. Feynman. *QED, the Strange Theory of Light and Matter*. Princeton, expanded edition, 2006.
- [13] R. P. Feynman. *Statistical Mechanics*. Westview/Perseus, 1998.
- [14] R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics*, volume I. Addison-Wesley, 1965.
- [15] R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics*, volume III. Addison-Wesley, 1965.
- [16] N.B. Gove and M.J. Martin. Log f tables. *Nucl. Data Tables A*, 10:206, 1971.
- [17] David J. Griffiths. *Introduction to Quantum Mechanics*. Pearson Prentice-Hall, second edition, 2005.
- [18] B.R. Holstein. The Van der Waals interaction. *Am. J. Phys.*, 69:441–449, 2001.
- [19] C. Kittel. *Introduction to Solid State Physics*. Wiley, 7th edition, 1996.
- [20] W. Koch and M. C. Holthausen. *A chemist’s guide to density functional theory*. Wiley-VCH, Weinheim, second edition, 2000.
- [21] K.S. Krane. *Introductory Nuclear Physics*. Wiley, 1988.
- [22] R. G. Parr and W. Yang. *Density Functional Theory of Atoms and Molecules*. Oxford, New York, 1989.
- [23] M.A. Preston and R.K. Bhaduri. *Structure of the Nucleus*. Addison-Wesley, 1975.
- [24] 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.
- [25] 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/hsp/33/2>.
- [26] A. J. M. Schmets and W. Montfrooij. Teaching superfluidity at the introductory level, 2008. URL <http://arxiv.org/abs/0804.3086>.
- [27] A. Sitenko and V. Tartakovskii. *Theory of Nucleus*. Kluwer, 1997.

- [28] M.R. Spiegel and J. Liu. *Mathematical Handbook of Formulas and Tables*. Schaum's Outline Series. McGraw-Hill, second edition, 1999.
- [29] R. L. Sproull. *Modern Physics, a textbook for engineers*. Wiley, first edition, 1956.
- [30] M. Srednicki. *Quantum Field Theory*. Cambridge University Press, Cambridge, UK, 2007.
- [31] 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.
- [32] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry*. Dover, first, revised edition, 1996.
- [33] D. R. Tilley and J. Tilley. *Superfluidity and Superconductivity*. Institute of Physics Publishing, Bristol and Philadelphia, third edition, 1990.
- [34] A. Yariv. *Theory and Applications of Quantum Mechanics*. Wiley & Sons, 1982.
- [35] A. Zee. *Quantum Field Theory in a Nutshell*. Princeton University Press, Princeton, NJ, 2003.

Web Pages

Below is a list of relevant web pages. Some of the discussions were based on them.

1. Amber Schilling's page¹

One of the info sources for chemical bonds, with lots of good pictures.

2. chemguide.co.uk²

Jim Clarke's UK site with lots of solid info.

3. Citizendium³

The Citizen's Compendium. Had a rather good write up on the quantization of the electromagnetic field.

4. ENSDF data⁴

Authoritative and comprehensive data on nuclei.

5. Hyperphysics⁵

An extensive source of info on chemical bonds and the periodic table.

6. ICC program⁶

Program to compute internal conversion coefficients.

7. J. Jäckle⁷

This web site includes a good description of the Peltier and Seebeck effects.

¹http://wulfenite.fandm.edu/Intro_to_Chem/table_of_contents.htm

²<http://www.chemguide.co.uk/>

³<http://en.citizendium.org/>

⁴<http://www-nds.iaea.org/relnsd/NdsEnsdf/QueryForm.html>

⁵<http://hyperphysics.phy-astr.gsu.edu/hbase/hph.html>

⁶<http://ie.lbl.gov/programs/icc/icc.htm>

⁷<http://www.uni-konstanz.de/FuF/Physik/Jaeckle/>

8. Mayer, M. Goeppert: Nobel Prize lecture⁸
An excellent introduction to the shell model of nuclear physics written for a general audience.
9. Middlebury College Modern Physics Laboratory Manual⁹
Gives a very understandable introduction to NMR with actual examples (item XIX.)
10. NIST data¹⁰
Authoritative values of physical constants from NIST.
11. NuDat 2 database¹¹
Extensive information about nuclei provided by the National Nuclear Data Center.
12. Purdue chemistry review¹²
This book's source for the electronegativity values.
13. Quantum Exchange¹³
Lots of stuff.
14. Rainwater, J.: Nobel Prize lecture¹⁴
An introduction to distorted nuclei written for a general audience.
15. T. Tritt¹⁵
Thermoelectric materials: principles, structure, properties, and applications. From Encyclopedia of Materials: Science and Technology. Elsevier 2002.
16. TUNL Nuclear Data Evaluation Group¹⁶
Extensive data on light nuclei from $A = 3$ to 20.

⁸nobelprize.org/nobel_prizes/physics/laureates/1963/mayer-lecture.html

⁹<http://cat.middlebury.edu/~PHManual/>

¹⁰<http://physics.nist.gov/PhysRefData/contents-constants.html>

¹¹<http://www.nndc.bnl.gov/nudat2/>

¹²<http://chemed.chem.psu.edu/genchem/topicreview/index.html>

¹³<http://www.compadre.org/quantum/>

¹⁴nobelprize.org/nobel_prizes/physics/laureates/1975/rainwater-lecture.pdf

¹⁵<http://virtual.clemson.edu/TMRL/Publications/PDFS/teoverview.pdf>

¹⁶<http://www.tunl.duke.edu/nucldata/>

17. University of Michigan¹⁷

Invaluable source on the hydrogen molecule and chemical bonds. Have a look at the animated periodic table for actual atom energy levels.

18. Wikipedia¹⁸

Probably this book's primary source of information on about every loose end, though somewhat uneven. Some great, some confusing, some overly technical.

¹⁷<http://www.umich.edu/~chem461/>

¹⁸<http://wikipedia.org>

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.

- 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, ...).

- Multiplication symbol. May indicate:

- An emphatic multiplication.
- Multiplication continued on the next line / from the previous line.
- A vectorial product between vectors. In index notation, the i -th component of $\vec{v} \times \vec{w}$ equals

$$(\vec{v} \times \vec{w})_i = v_{\bar{i}} w_{\hat{i}} - v_{\hat{i}} w_{\bar{i}}$$

where \bar{i} is the index following i in the sequence 123123..., and \hat{i} the one preceding it (or second following). Alternatively, evaluate the determinant

$$\vec{v} \times \vec{w} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{vmatrix}$$

- ! 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,

$$n! = \int_0^\infty t^n e^{-t} dt$$

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

$$n! \sim \sqrt{2\pi n} n^n e^{-n} [1 + \dots]$$

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 $\int_0^\infty 2e^{-u^2} du$, or $\int_{-\infty}^\infty e^{-u^2} du$, which equals $\sqrt{\int_{-\infty}^\infty e^{-x^2} dx \int_{-\infty}^\infty e^{-y^2} dy}$ and the integral under the square root can be done analytically using polar coordinates. The result is that

$$-\frac{1}{2}! = \int_{-\infty}^\infty e^{-u^2} du = \sqrt{\pi}$$

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.

$|\dots\rangle$ 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. $\delta(x - \underline{x})$, $|1s\rangle$ for the 1s or ψ_{100} hydrogen state, $|2p_z\rangle$ for the $2p_z$ or ψ_{210} state, etcetera. In short, whatever can indicate some state can be pushed into a ket.

$\langle \dots |$ A “bra” is like a ket $|\dots\rangle$, 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.

Σ 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 $2 + 1 + 4 = 7$.

\int Integration symbol, the continuous version of the summation symbol. For example,

$$\int_{\text{all } x} f(x) dx$$

is the summation of $f(x) dx$ over all little fragments dx that make up the entire x -range.

\rightarrow May indicate:

- An approaching process. $\lim_{\varepsilon \rightarrow 0}$ indicates for practical purposes the value of the expression following the \lim when ε is extremely small, $\lim_{r \rightarrow \infty}$ 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.

^ 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:

$$\nabla \equiv \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) = \hat{i} \frac{\partial}{\partial x} + \hat{j} \frac{\partial}{\partial y} + \hat{k} \frac{\partial}{\partial z}$$

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:

$$\text{grad } f = \nabla f = \hat{i} \frac{\partial f}{\partial x} + \hat{j} \frac{\partial f}{\partial y} + \hat{k} \frac{\partial f}{\partial z}.$$

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:

$$\text{div } \vec{v} = \nabla \cdot \vec{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$$

or in index notation

$$\text{div } \vec{v} = \nabla \cdot \vec{v} = \sum_{i=1}^3 \frac{\partial v_i}{\partial x_i}$$

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

$$(\text{curl } \vec{v})_i = (\text{rot } \vec{v})_i = (\nabla \times \vec{v})_i = \frac{\partial v_{\bar{i}}}{\partial x_i} - \frac{\partial v_i}{\partial x_{\bar{i}}}$$

where \bar{i} is the index following i in the sequence 123123..., and \bar{i} the one preceding it (or second following).

The operator ∇^2 is called the Laplacian. In Cartesian coordinates:

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

In non Cartesian coordinates, don't guess; look these operators up in a table book.

* A superscript star normally indicates a complex conjugate. In the complex conjugate of a number, every i is changed into a $-i$.

< Less than.

$\langle \dots \rangle$ May indicate:

- An inner product.
- An expectation value.

> Greater than.

[...] May indicate:

- A grouping of terms in a formula.
- A commutator. For example, $[A, B] = AB - BA$.

\equiv Emphatic equals sign. Typically means “by definition equal” or “everywhere equal.”

\sim Indicates approximately equal. Normally the approximation applies when something is small or large. Read it as “is approximately equal to.”

\propto Proportional to. The two sides are equal except for some unknown constant factor.

α (alpha) May indicate:

- The fine structure constant, $e^2/4\pi\epsilon_0\hbar c$, about 1/137.036 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.

β (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.

γ (Gamma) May indicate:

- The Gamma function. Look under “!” for details.
- The “width” or uncertainty in energy of an approximate energy eigenstate.

γ (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.

Δ (capital delta) May indicate:

- An increment in the quantity following it.
- A delta particle.
- Often used to indicate the Laplacian ∇^2 .

δ (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 6.6 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 10.5 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 .

∂ (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 .

ϵ (epsilon) May indicate:

- ϵ_0 is the permittivity of space. Equal to $8.854\ 19\ 10^{-12}\ \text{C}^2/\text{J m}$
- Scaled energy.
- Orbital energy.
- Lagrangian multiplier.
- A small quantity, if symbol ε is not available.

ε (variant of epsilon) May indicate:

- A very small quantity.
- The slop in energy conservation during a decay process.

η (eta) 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:

- Wave length.
- Decay constant.
- A generic eigenvalue.
- Scaled square momentum.
- Some multiple of something.

μ (mu) May indicate:

- Magnetic dipole moment.
- $\mu_B = e\hbar/2m_e = 9.27 \cdot 10^{-24} \text{ J/T}$ or $5.788 \cdot 10^{-5} \text{ eV/T}$ is the Bohr magneton.
- A summation index.
- Chemical potential/molar Gibbs free energy.

ν (nu) May indicate:

- Electron neutrino.
- Scaled energy eigenfunction number in solids.
- A summation index.
- Strength of a delta function potential.

ξ (xi) May indicate:

- Scaled argument of the one-dimensional harmonic oscillator eigenfunctions.
- x -position of a particle.
- A summation or integration index.

Π (Pi) Peltier coefficient.

π (pi) May indicate:

- The area of a circle of unit radius. Value 3.141 592...
- Half the perimeter of a circle of unit radius. Value 3.141 592...
- A 180° angle expressed in radians. Note that $e^{\pm i\pi} = -1$. Value 3.141 592...
- A 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.

ρ (rho) May indicate:

- Electric charge per unit volume.
- Scaled radial coordinate.
- Radial coordinate.
- Eigenfunction of a rotation operator R .
- Mass-base density.
- Energy density of electromagnetic radiation.

Σ (Sigma) Seebeck coefficient.

σ (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.

τ (tau) May indicate:

- Life time or half life.
- Some coefficient.

Φ (capital phi) May indicate:

- Some function of ϕ to be determined.
- The momentum-space wave function.
- Relativistic electromagnetic potential.

ϕ (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.

φ (variant of phi) May indicate:

- A change in angle ϕ .
- An alternate symbol for ϕ .
- An electric potential.

χ (chi) Spinor component.

Ψ (capital psi) Upper case psi is used for the wave function.

ψ (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.

ω (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.

A May indicate:

- Repeatedly used to indicate the operator for a generic physical quantity a , with eigenfunctions α .
- Electromagnetic vector potential.
- Einstein A coefficient.
- Some generic matrix.
- Some constant.
- Area.

\AA 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 \hat{a} or creation operator \hat{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 Å. 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 if 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 $-273.15\text{ }^{\circ}\text{C}$ in degrees Centigrade (Celsius).

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 6.1.9. These processes are very slow, to keep 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.

angle According to trigonometry, if the length of a segment of a circle is divided by its radius, it gives the total angular extent of the circle segment. More precisely, it gives the angle, in radians, between the line from the center to the start of the circle segment and the line from the center to the end of the segment. The generalization to three dimensions is called the “solid angle;” the total solid angle over which a segment of a spherical

surface extends, measured from the center of the sphere, is the area of that segment divided by the square radius of the sphere.

B May indicate:

- Repeatedly used to indicate a generic second operator or matrix.
- Magnetic field strength.
- Einstein *B* coefficient.
- Some constant.

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{i} , \hat{j} , 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 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:

- A third operator.
- A variety of different constants.

c May indicate:

- The speed of light, about $2.997\ 92\ 10^8$ m/s.
- 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 we 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 (1.6) 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

$$\langle f + \lambda e^{-i\alpha} g | f + \lambda e^{-i\alpha} g \rangle$$

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 we multiply out, we 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:

$$|\langle f|g \rangle| \leq \sqrt{\langle f|f \rangle} \sqrt{\langle g|g \rangle}$$

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 [28, pp. 40-...].

curl The curl of a vector \vec{v} is defined as $\text{curl } \vec{v} = \text{rot } \vec{v} = \nabla \times \vec{v}$.

D May indicate:

- Difference in wave number values.
- \mathcal{D} is density of states.

\vec{D} Primitive (translation) vector of a reciprocal lattice.

d Indicates a vanishingly small change or interval of the following variable. For example, dx can be thought of as a small segment of the x -axis.

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{d} Primitive (translation) vector of a crystal lattice.

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 df/dx , 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 we 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, we 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 we already know how to work out those 2×2 determinants, so we 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. We are allowed to use the following manipulations without seriously affecting the computed determinant:

1. We may “transpose” the matrix, i.e. change its columns into its rows.
2. We can create zeros in a row by subtracting a suitable multiple of another row.

3. We may also swap rows, as long as we remember that each time that we swap two rows, it will flip over the sign of the computed determinant.
4. We 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 lower triangular form”, we can eliminate all matrix coefficients a_{ij} for which j is greater than i , and that makes evaluating the determinant pretty much trivial.

div(ergence) The divergence of a vector \vec{v} is defined as $\text{div } \vec{v} = \nabla \cdot \vec{v}$.

E May indicate:

- The total energy. Possible values are the eigenvalues of the Hamiltonian.
- $E_n = -\hbar^2/2m_e a_0^2 n^2 = E_1/n^2$ may indicate the nonrelativistic (Bohr) energy levels of the hydrogen atom. The ground state energy E_1 equals -13.605 7 eV.
- Electric field strength. To keep electric field apart from energy, note that the electric field is a vector while energy is a scalar. A vector sign over the E implies it is the electrical field. A subscript denoting a component indicates the same thing, but sometimes a subscript may refer to the part of energy that is in a given direction. Use context to decide.
- Internal energy of a substance.

e May indicate:

- The basis for the natural logarithms. Equal to 2.71 281 828 459... 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 18 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^{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

$$\vec{v}_1 = \begin{pmatrix} a_{12} \\ \lambda_1 - a_{11} \end{pmatrix} \quad \vec{v}_2 = \begin{pmatrix} \lambda_2 - a_{22} \\ a_{21} \end{pmatrix}$$

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 equal to $1.602\,18\,10^{-19}$ J.

exponential function A function of the form e^{\dots} , also written as $\exp(\dots)$. 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.

f May indicate:

- A generic function.
- A generic vector.
- A fraction.
- The resonance factor.
- Specific Helmholtz free energy.

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. 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.

G Gibbs free energy.

g May indicate:

- A second generic function or a second generic vector.
- The strength of gravity, $9.806\,65\text{ m/s}^2$ exactly under standard conditions on the surface of the earth.
- The g-factor, a nondimensional constant that indicates the gyro-magnetic ratio relative to charge and mass. For the electron $g_e = 2.002\,319\,304\,362$. For the proton $g_p = 5.585\,694\,7$. For the neutron, based on the mass and charge of the proton, $g_n = -3.826\,085$.

- 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} ,

$$\int_V \nabla \cdot \vec{v} dV = \int_A \vec{v} \cdot \vec{n} dA$$

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 $\text{grad } f = \nabla f$.

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.

h May indicate:

- The original Planck constant $h = 2\pi\hbar$.
- h_n is a one-dimensional harmonic oscillator eigenfunction.
- Single-electron Hamiltonian.
- Specific enthalpy.

\hbar The reduced Planck constant, equal to $1.05457 \cdot 10^{-34}$ Js. 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 $\hbar c = 197.329$ MeV fm.

I May indicate:

- The number of electrons or particles.
- Electrical current.
- Unit matrix.
- I_A is Avogadro's number, $6.0221 \cdot 10^{26}$ 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 $c = c_r + i c_i$ with c_r and c_i real numbers, then $\Im(c) = c_i$. Note that $c - c^* = 2i\Im(c)$.

\mathcal{I} May indicate:

- Radiation energy intensity.
- 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 .

i The standard square root of minus one: $i = \sqrt{-1}$, $i^2 = -1$, $1/i = -i$, $i^* = -i$.

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, we 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: $\dots, -2, -1, 0, 1, 2, 3, 4, \dots$

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.

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 transferred with the surroundings, no work is done on or by the surroundings.

J May indicate:

- Total angular momentum.
- Number of nuclei in a quantum computation of electronic structure.

j May indicate:

- The quantum number of total square angular momentum.
- \vec{j} is electric current density.
- The number of a nucleus in a quantum computation.
- A summation index.
- A generic index or counter.

K May indicate:

- The atomic states or orbitals with theoretical Bohr energy E_1
- Degrees Kelvin.

K May indicate:

- An exchange integral in Hartree-Fock.
- Maximum wave number value.

K Thomson (Kelvin) coefficient.

k May indicate:

- A wave number. A wave number is a measure for how fast a periodic function oscillates with variations in spatial position.

- A summation index.

k_B Boltzmann constant. Equal to $1.380\,65\,10^{-23}$ J/K. Relates absolute temperature to a typical unit of heat motion energy.

kmol A kilo mole refers to $6.022\,1\,10^{26}$ 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.

L The atomic states or orbitals with theoretical Bohr energy E_2

L May indicate:

- Angular momentum.
- Orbital angular momentum.

L Lagrangian.

l May indicate:

- The azimuthal quantum number.
- A generic summation index.

ℓ May indicate:

- The typical length in the harmonic oscillator problem.
- The dimensions of a solid block (with subscripts).
- A length.
- Multipole level in transitions.

lim Indicates the final result of an approaching process. $\lim_{\varepsilon \rightarrow 0}$ 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 $x\hat{i} + y\hat{j} + z\hat{k}$ with the objects the unit vectors \hat{i} , \hat{j} , and \hat{k} and the coefficients the position coordinates x , y , and z .

M The atomic states or orbitals with theoretical Bohr energy E_3

M May indicate:

- Molecular mass. See separate entry.
- Mirror operator.
- Figure of merit.

m May indicate:

- Mass.
 - m_e : electron mass. Equal to $9.109\,382\,10^{-31}$ kg. The rest mass energy is 0.510 998 910 MeV.
 - m_p : proton mass. Equal to $1.672\,621\,10^{-27}$ 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.
 - m : particle mass.
- The magnetic quantum number.
- Number of a single-electron wave function.
- A generic summation index or generic integer.

matrix A table of numbers.

As a simple example, a two-dimensional matrix A is a table of four numbers called a_{11} , a_{12} , a_{21} , and a_{22} :

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

unlike a two-dimensional (ket) vector \vec{v} , which would consist of only two numbers v_1 and v_2 arranged in a column:

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

(Such a vector can be seen as a “rectangular matrix” of size 2×1 , but let’s not get into that.)

In index notation, a matrix A is a set of numbers $\{a_{ij}\}$ indexed by two indices. The first index i is the row number, the second index j is the column number. A matrix turns a vector \vec{v} into another vector \vec{w} 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 another two-dimensional ket vector.

Note that in matrix multiplications like the example above, in geometric terms we take dot products between the rows of the first factor and the column of the second factor.

To multiply two matrices together, just think of the columns of the second matrix as separate vectors. For example:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

which is another two-dimensional matrix. In index notation, the ij component of the product matrix has value $\sum_k a_{ik}b_{kj}$.

The zero matrix is like the number zero; it does not change a matrix it is added to and turns whatever it is multiplied with into zero. A zero matrix is zero everywhere. In two dimensions:

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

A unit matrix is the equivalent of the number one for matrices; it does not change the quantity it is multiplied with. A unit matrix is one on its “main diagonal” and zero elsewhere. The 2 by 2 unit matrix is:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

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 switch the two indices. Graphically, it turns its rows into its columns and vice versa. The Hermitian “adjoint” A^H is what you get if you switch the two indices and then take the complex conjugate of every element. 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.

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. Corresponding names are yotta, zetta, exa, peta, tera, giga, mega, kilo, 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. “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, that 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.)

N The atomic states or orbitals with theoretical Bohr energy E_4

N May indicate:

- Number of states.
- Number of single-particle states.
- Number of neutrons in 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.
- A generic index.
- A natural number.
- n_s is the number of spin states.

and maybe some other stuff.

natural Natural numbers are the numbers: $1, 2, 3, 4, \dots$

normal A normal operator or matrix is one that has orthonormal eigenfunctions or eigenvectors. Since these are not orthonormal in general, a normal operator or matrix is abnormal! Another example of a highly confusing term. If it would have been called, say, orthonormal, you would have a clue what meaning of “normal” was being referred to. 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 (ortho)normal, it must commute with its Hermitian adjoint, $[A, A^\dagger] = 0$. Therefore, Hermitian, skew-Hermitian, and unitary operators or matrices are (ortho)normal.

opposite The opposite of a number a is $-a$. In other words, it is the additive inverse.

P May indicate:

- The linear momentum eigenfunction.
- A power series solution.
- Probability.
- Pressure.
- Hermitian part of an annihilation operator.

p May indicate:

- Linear momentum.
- Linear momentum in the x -direction.
- Integration variable with units of linear momentum.

\mathbf{p} May indicate:

- 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.

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:

$$(\vec{r} - \vec{r}_P) \cdot (\vec{r}_Q - \vec{r}_P) = \frac{1}{2}(\vec{r}_Q - \vec{r}_P) \cdot (\vec{r}_Q - \vec{r}_P)$$

(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 $c = |c|e^{i\alpha}$ 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 $e^{i(\omega t - \phi)}$, 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.

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 \hat{p}_x which equals $\hbar\partial/\partial x$. (But which becomes canonical momentum in a magnetic field.)

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.

q May indicate:

- Charge.
- Heat flux density.
- The number of an energy eigenfunction of a system of particles.
- Generic index.

R May indicate:

- Some function of r to be determined.
- Some function of (x, y, z) to be determined.
- Rotation operator.
- Ideal gas constant.
- Transition rate.
- Nuclear radius.
- R_{nl} is a hydrogen radial wave function.
- $R_u = 8.314\,472 \text{ kJ/kmol K}$ is the universal gas constant, the equivalent of Boltzmann's constant for a kmol instead of a single atom or molecule.

\Re The real part of a complex number. If $c = c_r + ic_i$ with c_r and c_i real numbers, then $\Re(c) = c_r$. Note that $c + c^* = 2\Re(c)$.

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:

$$m = \frac{m_0}{\sqrt{1 - (v/c)^2}} \quad p = mv \quad T = mc^2 - m_0c^2$$

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...

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{r} The position vector. In Cartesian coordinates (x, y, z) or $x\hat{i}+y\hat{j}+z\hat{k}$. In spherical coordinates $r\hat{r}$. 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.

rot The rot of a vector \vec{v} is defined as $\text{curl } \vec{v} = \text{rot } \vec{v} = \nabla \times \vec{v}$.

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 for generic angular momentum.)
- Entropy.

s Energy state with orbital azimuthal quantum number $l = 0$. Spherically symmetric.

s May indicate:

- Spin value of a particle. Equals $1/2$ 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.

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 [28, pp. 40-]. Good to remember: $\cos^2 \alpha + \sin^2 \alpha = 1$.

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} ,

$$\int_A (\nabla \times \vec{v}) \, dA = \oint \vec{v} \cdot d\vec{r}$$

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 why!

symmetry Symmetries are operations 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 temperature 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.

T Translation operator that translates a wave function a given amount through space.

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,

$$\vec{a} \cdot (\vec{b} \times \vec{c}) = \sum_i a_i (b_{\bar{i}} c_{\bar{i}} - b_{\bar{i}} c_{\bar{i}})$$

where \bar{i} is the index following i in the sequence 123123..., and \bar{i} 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

$$a_{\bar{i}}(b_i c_{\bar{i}} - b_{\bar{i}} c_i) - a_{\bar{i}}(b_{\bar{i}} c_i - b_i c_{\bar{i}})$$

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 the atomic mass unit, equivalent to $1.660\,538\,78 \cdot 10^{-27}$ kg or $931.494\,03$ MeV/c².

u May indicate:

- The first velocity component in a Cartesian coordinate system.
- A complex coordinate in the derivation of spherical harmonics.
- An integration variable.

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, \hat{V} is simply written as *V*.

- \mathcal{V} is 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 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:

$$(1, 3, 5) \begin{pmatrix} 7 \\ 11 \\ 13 \end{pmatrix} = 1 \cdot 7 + 3 \cdot 11 + 5 \cdot 13 = 105$$

To turn a ket into a bra for purposes of taking inner products, write the complex conjugates of its components as a row.

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

$$c_i = a_{\bar{i}} b_{\bar{i}} - a_{\bar{i}} b_{\bar{i}}$$

where \bar{i} is the index following i in the sequence 123123..., and \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{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.

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 generic argument of a function.

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 generic argument of a function.

Index

F , <u>1148</u>	μ , <u>1138</u>
\cdot , <u>1131</u>	ν , <u>1138</u>
\times , <u>1131</u>	ξ , <u>1138</u>
$!$, <u>1131</u>	Π , <u>1139</u>
$ $, <u>1132</u>	π , <u>1139</u>
$ \dots\rangle$, <u>1133</u>	ρ , <u>1139</u>
$\langle\dots $, <u>1133</u>	Σ , <u>1139</u>
\uparrow , <u>1133</u>	σ , <u>1139</u>
\downarrow , <u>1133</u>	τ , <u>1139</u>
Σ , <u>9</u>	21 cm line
\sum , <u>1133</u>	derivation, <u>770</u>
\int , <u>1133</u>	intro, <u>761</u>
\rightarrow , <u>1133</u>	Φ , <u>1140</u>
$\vec{}$, <u>1133</u>	ϕ , <u>1140</u>
$\hat{}$, <u>1133</u>	φ , <u>1140</u>
$'$, <u>1134</u>	χ , <u>1140</u>
∇ , <u>1134</u>	Ψ , <u>1140</u>
$*$, <u>1134</u>	ψ , <u>1140</u>
$<$, <u>1135</u>	ω , <u>1140</u>
$\langle\dots\rangle$, <u>1135</u>	
$>$, <u>1135</u>	A , <u>1141</u>
$[\dots]$, <u>1135</u>	\AA , <u>1141</u>
\equiv , <u>1135</u>	a , <u>1141</u>
\sim , <u>1135</u>	a_0 , <u>1141</u>
\propto , <u>1135</u>	absolute, <u>1141</u>
α , <u>1135</u>	absolute temperature, <u>184</u>
β , <u>1135</u>	absolute value, <u>4</u>
Γ , <u>1136</u>	absolute zero
γ , <u>1136</u>	nonzero energy, <u>53</u>
Δ , <u>1136</u>	requires ground state, <u>441</u>
δ , <u>1136</u>	absorbed dose, <u>578</u>
∂ , <u>1137</u>	absorption
ϵ , <u>1137</u>	incoherent radiation, <u>317</u>
ϵ_0 , <u>1137</u>	single weak wave, <u>314</u>
ε , <u>1137</u>	absorptivity, <u>199</u>
η , <u>1137</u>	acceleration
Θ , <u>1137</u>	in quantum mechanics, <u>292</u>
θ , <u>1137</u>	acceptors
ϑ , <u>1138</u>	semiconductors, <u>250</u>
κ , <u>1138</u>	actinides, <u>163</u>
λ , <u>1138</u>	actinoids, <u>163</u>

action, 835
 relativistic, *see* special relativity
 activation energy
 nuclear fission, 661
 radicals, 122
 active viewpoint, 296
 activity, 578
 adiabatic
 disambiguation, 1142
 quantum mechanics, 292
 thermodynamics, 472
 adiabatic surfaces, 365
 adiabatic theorem
 derivation and implications, 921
 intro, 293
 adjoint, 1142
 matrices, 1156
 Aharonov-Bohm effect, 528
 Airy functions
 application, 943
 connection formulae, 952
 graphs, 951
 software, 948
 alkali metals, 163
 alkaline metals, 163
 allowed transitions
 beta decay, 719
 alpha, *see* α
 alpha decay, 574
 data, 599
 definition, 574
 Gamow/Gurney and Condon theory, 599
 overview of data, 570
 Q -value, 714
 quantum mechanical tunneling, 599
 alpha particle, 599
 ammonia molecule, 123
 amplitude
 quantum, 23
 angle, 1142
 angular frequency, 333
 angular momentum, 64
 addition, 511
 Clebsch-Gordan coefficients, 512
 intro, 310
 advanced treatment, 501
 component, 65
 eigenfunctions, 66
 eigenvalues, 66
 conservation, 309
 definition, 64
 fundamental commutation relations
 as an axiom, 502
 intro, 99
 ladder operators, 503
 ladders, 504
 normalization factors, 508
 nuclei
 data, 665
 operator
 Cartesian, 64
 possible values, 506
 spin, 127
 square angular momentum, 67
 eigenfunctions, 68
 eigenvalues, 69
 symmetry and conservation, 296
 triangle inequality
 intro, 311
 uncertainty, 71
 anomalous magnetic moment, 555
 anti-bonding, 412
 anticommutator, 782
 antisymmetrization requirement, 138
 graphical depiction, 444
 indistinguishable particles, 446
 number of terms, 145
 using groupings, 142
 using occupation numbers, 776
 using Slater determinants, 142
 atomic mass
 conversion to nuclear mass, 580
 versus nuclear mass, 716
 atomic mass unit, 580
 atomic number, 150, 570
 atoms
 eigenfunctions, 151
 eigenvalues, 151
 ground state, 153
 Hamiltonian, 150
 Auger effect
 Meissner, 1100
 Auger electrons, 745
 Avalanche diode, 261
 average
 versus expectation value, 87
 Avogadro's number, 1150
 azimuthal quantum number, 69
 B, 1143

b, [1143](#)
 Balmer transitions, [79](#)
 band gap
 and Bragg reflection, [905](#)
 direct versus indirect, [266](#)
 intro, [231](#)
 band structure
 crossing bands, [411](#)
 nearly-free electrons, [421](#)
 widely spaced atoms, [396](#)
 band theory
 intro, [228](#)
 barn, [684](#)
 baryon, [128](#)
 baryons
 intro, [592](#)
 basis, [1143](#)
 crystal, *see* lattice, basis
 spin states, [136](#)
 vectors or functions, [15](#)
 battery, [219](#)
 bcc, *see* lattice
 becquerel, [578](#)
 Bell's theorem, [806](#)
 cheat, [809](#)
 benzene molecular ring, [123](#)
 Berry's phase, [925](#)
 beryllium-11
 nuclear spin, [640](#)
 Bessel functions
 spherical, [958](#)
 beta, *see* β
 beta decay, [704](#)
 beta-plus decay
 definition, [574](#)
 double
 explanation, [713](#)
 electron capture
 definition, [573](#)
 electron emission
 definition, [573](#)
 energetics
 data, [705](#)
 energy release data
 emph, [704](#)
 Fermi theory, [1072](#)
 forbidden decays, [718](#)
 intermediate vector bosons, [589](#)
 intro, [573](#)
 inverse beta decay

 definition, [573](#)
 K or L capture
 definition, [573](#)
 lone neutron, [568](#)
 momentum conservation, [1090](#)
 negatron emission
 definition, [574](#)
 nuclei that do, [573](#)
 overview of data, [570](#)
 positron emission
 definition, [573](#)
 Q -value, [714](#)
 superallowed decays, [727](#)
 von Weizsaecker predictions
 emph, [711](#)
 beta vibration
 nuclei, [658](#)
 Bethe-von Weizsäcker formula, [596](#)
 binding energy
 definition, [107](#)
 hydrogen molecular ion, [107](#)
 hydrogen molecule, [121](#)
 lithium hydride, [125](#)
 Biot-Savart law, [551](#)
 derivation, [1056](#)
 blackbody radiation, [491](#)
 intro, [196](#)
 blackbody spectrum, [197](#)
 extended derivation, [491](#)
 Bloch function
 nearly-free electrons, [421](#)
 one-dimensional lattice, [402](#)
 three-dimensional lattice, [410](#)
 Bloch wave
 explanation, [336](#)
 intro, [239](#)
 Bloch's theorem, [402](#)
 body-centered cubic, *see* lattice
 Bohm
 EPR experiment, [806](#)
 Bohr energies, [78](#)
 relativistic corrections, [759](#)
 Bohr magneton, [555](#)
 Bohr radius, [81](#)
 Boltzmann constant, [1153](#)
 Boltzmann factor, [454](#)
 bond
 covalent, [165](#)
 hydrogen, [167](#)
 ionic, [172](#)

pi, 166
 polar, 167
 sigma, 165
 Van der Waals, 387
 bond length
 definition, 107
 hydrogen molecular ion, 109
 hydrogen molecule, 121
 Born
 approximation, 961
 series, 964
 Born statistical interpretation, 21
 Born-Oppenheimer approximation
 basic idea, 361
 derivation, 359
 diagonal correction, 977
 hydrogen molecular ion, 101
 hydrogen molecule, 114
 include nuclear motion, 363
 relation to adiabatic theorem, 293
 spin degeneracy, 972
 vibronic coupling terms, 975
 Borromean nucleus, 640
 Bose-Einstein condensation
 derivation, 486
 intro, 186
 rough explanation, 189
 superfluidity, 1002
 Bose-Einstein distribution
 blackbody radiation, 491
 intro, 196
 canonical probability, 453
 for given energy, 451
 identify chemical potential, 484
 intro, 195
 bosons, 127
 ground state, 183
 symmetrization requirement, 138
 bound states
 hydrogen
 energies, 78
 boundary conditions
 acceptable singularity, 864
 hydrogen atom, 878
 across delta function potential, 946
 at infinity
 harmonic oscillator, 867
 hydrogen atom, 878
 impenetrable wall, 33
 radiation, 941
 accelerating potential, 942
 three-dimensional, 955
 unbounded potential, 944
 Bq, 578
 bra, 10, 1133
 Bragg diffraction
 electrons, 438
 Bragg planes
 Brillouin fragment boundaries, 411
 energy singularities, 427
 one-dimensional (Bragg points), 404
 X-ray diffraction, 438
 Bragg reflection
 and band gaps, 905
 Bragg's law, 432
 Brillouin zone
 intro, 244
 one-dimensional, 403
 three-dimensional, 410
 broadband radiation
 intro, 262
 built-in potential, 255
^oC, 1143
 C, 1143
 c, 1143
 canonical commutation relation, 96
 canonical Hartree-Fock equations, 376
 canonical momentum, *see* special relativity
 intro, 295
 with a magnetic field, 528
 canonical probability distribution, 453
 carbon nanotubes
 electrical properties
 intro, 236
 intro, 170
 Carnot cycle, 464
 cat, Schrödinger's, 804
 Cauchy-Schwartz inequality, 1144
 centrifugal stretching, 652
 chain reaction
 emph, 662
 charge
 electrostatics, 537
 charge conjugation
 Wu experiment, 730
 charge independence
 nuclear force, 568
 explanation, 593
 charge symmetry

example, 624
 nuclear force, 568
 explanation, 593
 charge transfer insulators, 234
 chemical bonds, 165
 covalent pi bonds, 166
 covalent sigma bonds, 165
 hybridization, 169
 ionic bonds, 172
 polar covalent bonds, 167
 promotion, 169
 spⁿ hybridization, 169
 chemical equilibrium
 constant pressure, 482
 constant volume, 482
 chemical potential, 480
 and diffusion, 218
 intro, 215
 and distributions, 484
 line up
 Peltier cooler, 268
 microscopic, 484
 chi, *see* χ
 Ci, 578
 classical, 1144
 Clausius-Clapeyron equation, 481
 Clebsch-Gordan coefficients, 511
 coefficient of performance, 466
 coefficients of eigenfunctions
 evaluating, 60
 give probabilities, 29
 time variation, 281
 collapse of the wave function, 28
 collision-dominated regime, 317
 collisionless regime, 316
 color force, 568
 gluon exchange
 intro, 288
 commutation relation
 canonical, 96
 commutator, 93
 definition, 95
 commutator eigenvalue problems, 504
 commuting operators, 94
 common eigenfunctions, 94
 comparative half-life, 721
 complete set, 15
 complex conjugate, 4
 complex numbers, 3
 component waves, 329
 components of a vector, 6
 conduction band
 intro, 231
 conductivity
 effect of light, 264
 electrical, 227
 ionic, 237
 configuration mixing, 632
 confinement, 206
 single particle, 45
 connection formulae, 951, 952
 conservation laws
 and symmetries, 296
 conservation of crystal momentum, 245
 conservation of wavevector, 245
 conserved vector current hypothesis, 1084
 contact potential, 219
 continuity equation
 incompressible flow, 1064
 conventional cell, 408
 conversion electron, 744
 Copenhagen Interpretation, 27
 correlation energy, 383
 cos, 1144
 Coulomb barrier, 601
 Coulomb integrals, 375
 Coulomb potential, 72
 coupling constant, 1076
 covalent bond
 hydrogen molecular ion, 100
 covalent solids, 411
 creationists, 1022
 cross product, 1165
 crystal
 lattice, *see* lattice
 one-dimensional
 primitive translation vector, 401
 three-dimensional
 primitive vectors, 407
 crystal momentum, 337
 light-emitting diodes, 266
 curie, 578
 curl, 1134, 1144
 d, 1145
 D, 1144
 d, 1145
 d-block
 periodic table, 163
 \vec{D} , 1145

\vec{d} , [1145](#)
 Dalton, [580](#)
 Darwin term, [763](#)
 Debye model, [493](#)
 Debye temperature, [493](#), [494](#)
 decay constant, [578](#)
 decay rate, [578](#)
 deformed nuclei
 emph, [646](#)
 degeneracy, [58](#)
 degeneracy pressure, [204](#)
 degenerate semiconductor, [251](#)
 delayed neutrons, [662](#)
 Delta, *see* Δ
 delta, *see* δ
 delta function, [323](#)
 three-dimensional, [323](#)
 density
 mass, [458](#)
 molar, [458](#)
 particle, [458](#)
 density of modes, [182](#)
 density of states, [180](#)
 confined, [206](#)
 periodic box, [224](#)
 depletion layer, [255](#)
 derivative, [1145](#)
 determinant, [1145](#)
 deuterium, [569](#)
 deuteron
 intro, [569](#)
 diamagnetic contribution, [556](#)
 diamond
 band gap, [411](#)
 intro, [170](#)
 differential cross-section, [957](#)
 dimensional analysis, [908](#)
 dineutron, [700](#)
 diode
 semiconductor, [254](#)
 diode laser, [266](#)
 dipole, [542](#)
 dipole strength
 molecules, [389](#)
 diproton, [700](#)
 Dirac delta function, [323](#)
 Dirac equation, [524](#)
 Dirac notation, [18](#)
 direct band gap, [266](#)
 direct gap semiconductor, [245](#)
 discrete spectrum
 versus broadband
 intro, [262](#)
 disintegration constant, [578](#)
 disintegration rate, [578](#)
 dispersion relation, [333](#)
 distinguishable particles
 intro, [188](#), [194](#)
 div, [1134](#)
 div(ergence), [1147](#)
 divergence, [1134](#)
 divergence theorem, [1150](#)
 donors
 semiconductors, [250](#)
 doping
 semiconductors, [247](#)
 dose equivalent, [579](#)
 dot product, [8](#)
 double layer of charges
 contact surfaces, [219](#)
 doublet states, [137](#)
 dpm, [578](#)
 dynamic phase, [924](#)
 E , [1147](#)
 e , [1147](#)
 effective dose, [579](#)
 effective mass
 from equation of motion, [338](#)
 one-dimensional example, [241](#)
 Ehrenfest theorem, [292](#)
 e^{iax} , [1148](#)
 eigenfunction, [13](#)
 eigenfunctions
 angular momentum component, [66](#)
 atoms, [151](#)
 harmonic oscillator, [54](#)
 hydrogen atom, [81](#)
 impenetrable spherical shell, [1062](#)
 linear momentum, [325](#)
 position, [322](#)
 square angular momentum, [68](#)
 eigenvalue, [13](#)
 eigenvalue problems
 commutator type, [504](#)
 ladder operators, [504](#)
 eigenvalues
 angular momentum component, [66](#)
 atoms, [151](#)
 harmonic oscillator, [52](#)

hydrogen atom, 78
 impenetrable spherical shell, 1062
 linear momentum, 325
 position, 322
 square angular momentum, 69
 eigenvector, 13, 1148
Einstein
 dice, 29
 mass-energy relation, *see* special relativity
 summation convention, 850
 swiped special relativity, 839
 Einstein A and B coefficients, 318
 Einstein's derivation, 319
Einstein A coefficients
 quantum derivation, 794
Einstein B coefficients
 quantum derivation, 931
Einstein Podolski Rosen, 806
electric charge
 electron and proton, 72
electric dipole moment
 nuclei, 679
electric dipole transitions
 intro, 307
 selection rules, *see* selection rules
electric potential
 classical derivation
 emph, 529, 1049
 quantum derivation
 emph, 527
 relativistic derivation
 emph, 857
electric quadrupole moment
 nuclei, 679
electric quadrupole transitions
 Hamiltonian, 929
 intro, 309
 selection rules, *see* selection rules
electrical conduction
 intro, 224
electrochemical potential
 definition, 212
electromagnetic field
 energy, 788
 Hamiltonian, 526
 Maxwell's equations, 529
 quantized, 790
electromagnetic force
 photon exchange, 588

electromagnetic potentials
 gauge transformation, 858
electromagnetic waves
 quantized, 794
electron
 in magnetic field, 554
electron affinity, 390
 Hartree-Fock, 380
electron capture, *see* beta decay
electron emission, *see* beta decay
electronegativity, 159, 390
electrons
 lack of intelligence, 205, 260
 emission of radiation, 303
 emissivity, 199
 energy conservation, 285
energy spectrum
 harmonic oscillator, 52
 hydrogen atom, 78
energy-time uncertainty relation
 decay of a state
 emph, 289
 Mandelshtam-Tamm version, 290
 popular version, 591
enthalpy, 460
enthalpy of vaporization, 482
entropy, 469
 descriptive, 462
EPR, 806
epsilon, *see* ϵ , ε
equipartition theorem, 495
equivalent dose, 579
eta, *see* η
Euler formula, 5
eV, 1148
even-even nuclei
 emph, 575
Everett, III, 813
every possible combination, 112, 129
exchange force mechanism
 and two-state systems, 288
 nuclear forces, 588
exchange integrals, 375
exchange operator, 121
exchange terms, *see* twilight terms
exchanged
 Las Vegas interpretation, 177
excited determinants, 384
exciton
 intro, 264

exclusion principle, 144
 exclusion-principle repulsion, 164
 expectation value, 86

- definition, 89
- simplified expression, 90
- versus average, 87

 exponential function, 1148
 exponential of an operator, 294
 exposure, 578
 extended zone scheme, 417

- intro, 244

 extensive variable, 458
 extreme independent particle model, 629
 extreme single-particle model, 629

f , 1149
 f-block

- periodic table, 163

 F-center

- intro, 264

 face centered cubic, *see* lattice
 factorial, 1131
 fcc, *see* lattice
 fermi, 595
 Fermi brim

- definition, 212

 Fermi decay, 718
 Fermi energy

- definition, 212
- electrons in a box, 202

 Fermi factor, 213

- definition, 213

 Fermi function

- intro, 1078
- value, 1092

 Fermi integral

- intro, 721
- value, 1093

 Fermi level

- definition, 212
- line up
 - Peltier cooler, 268

 Fermi surface

- electrons in a box, 202
- periodic boundary conditions, 221
- periodic zone scheme, 420
- reduced zone scheme, 420

 Fermi temperature, 488
 Fermi theory

- comparison with data, 723

 Fermi theory of beta decay, 1072
 Fermi's golden rule, 1084
 Fermi-Dirac distribution

- canonical probability, 453
- for given energy, 451
- identify chemical potential, 484
- intro, 210

 Fermi-Kurie plot, 729
 fermions, 127

- antisymmetrization requirement, 138
- ground state, 200

 Feynman diagrams, 964
 field emission, 218
 field operators, 797
 filled shells, 520
 filtering property, 323
 Fine structure, 760

- fine structure
 - hydrogen atom, 759
- fine structure constant, 760

 first Brillouin zone

- intro, 244

 first law of thermodynamics, 441, 460
 first-forbidden decays

- beta decay, 720

 fission

- energetics, 581
- spontaneous
 - definition, 574
 - overview of data, 570

 flopping frequency, 564
 Floquet theory, 402
 fluorine-19

- nuclear spin, 638

 flux, 911
 Fock operator, 377
 Fock space kets

- beta decay, 1074

 Fock state, 777
 forbidden decays

- beta decay, 718

 forbidden transitions, 307

- alpha decay, 606

 force

- in quantum mechanics, 292

 forces

- particle exchange mechanism, 588

 four-vectors, *see* special relativity
 Fourier analysis, 403
 Fourier coefficients, 859

Fourier integral, 860
 Fourier series, 859
 Fourier transform, 334, 860
 Fourier's law
 heat conduction, 912
 free path, 226
 free-electron gas
 intro, 200
 model for crystal structure, 414
 periodic box
 intro, 220
 specific heat, 1033
 ft -value, 721
 function, 6, 7, 1149
 functional, 1149
 fundamental commutation relations
 as an axiom, 502
 orbital angular momentum, 99
 spin
 introduction, 132
 fusion
 energetics, 581

 G , 1149
 g , 1149
 g -factor, 555
 Galilean transformation, 844
 Galvani potential, 219
 Gamma, *see* Γ
 gamma, *see* γ
 gamma decay
 definition, 574
 gamma function, 1131
 gamma rays
 intro, 731
 gamma vibration
 nuclei, 658
 Gamow theory, 599
 Gamow-Teller decay, 718
 gauge transformation
 electromagnetic potentials, 858
 Gauss' theorem, 1150
 generalized coordinates, 833
 intro, 295
 generator of rotations, 299
 geometric phase, 924
 Gibbs free energy, 477
 microscopic, 484
 grad, 1134
 grad(ient), 1150

 gradient, 1134
 grain, 395
 grain boundaries, 395
 graphene
 electrical properties
 intro, 236
 graphite
 electrical properties
 intro, 236
 intro, 170
 gray, 579
 Green's function
 Laplacian, 548
 ground state
 absolute zero temperature, 184
 atoms, 153
 bosons, 183
 fermions, 200
 harmonic oscillator, 54
 hydrogen atom, 79, 81
 hydrogen molecular ion, 108
 hydrogen molecule, 121, 136, 139
 nonzero energy, 53
 group property
 coordinate system rotations, 1043
 Lorentz transform, 851
 group theory, 300
 group velocity, 332
 intro, 331
 gyromagnetic ratio, 554

 H , 1150
 h , 1150
 half-life, 578
 excited atoms, 321
 halo nucleus, 640
 halogens, 163
 Hamiltonian, 26
 and physical symmetry, 297
 atoms, 150
 classical, 838
 electromagnetic field, 526
 gives time variation, 281
 harmonic oscillator, 48
 partial, 50
 hydrogen atom, 72
 hydrogen molecular ion, 101
 hydrogen molecule, 114
 in matrix form, 148
 numbering of eigenfunctions, 26

one-dimensional free space, 328
 relativistic, non quantum, 858
 Hamiltonian dynamics
 relation to Heisenberg picture, 295
 Hamiltonian perturbation coefficients, 748
 Hankel functions
 spherical, 958
 harmonic oscillator, 47
 classical frequency, 48
 eigenfunctions, 54
 eigenvalues, 52
 energy spectrum, 52
 ground state, 54
 Hamiltonian, 48
 partial Hamiltonian, 50
 particle motion, 342
 Hartree product, 142, 367
 intro, 177
 Hartree-Fock, 366
 Coulomb integrals, 375
 exchange integrals, 375
 restricted
 closed shell, 370
 open shell, 370
 unrestricted, 370
 \hbar , 1150
 heat, 185, 461
 heat capacity
 valence electrons, 212
 heat conduction
 electrons, 236
 heat flux density
 including Peltier effect, 912
 omit density, 912
 heavy water, 570
 Heisenberg
 uncertainty principle, 23
 uncertainty relationship, 96
 helion, 571
 helium ionization energy, 749
 Hellmann-Feynman theorem, 749
 Helmholtz equation, 961
 Green's function solution, 962
 Helmholtz free energy, 477
 microscopic, 483
 Hermitian matrices, 1156
 Hermitian operators, 15
 hidden variables, 29, 807
 hidden versus nonexistent, 71
 hieroglyph, 431, 769
 hole
 nuclear shell model, 625
 holes
 in shells, 520
 light, heavy, split-off, 339
 semiconductors
 holes per state, 248
 holes per unit volume, 249
 intro, 234
 Hund's rules, 430
 hybridization, 169
 hydrogen
 metallic, 233
 nonmetal, 232
 hydrogen atom, 72
 eigenfunctions, 81
 eigenvalues, 78
 energy spectrum, 78
 ground state, 79, 81
 Hamiltonian, 72
 relativistic corrections, 759
 hydrogen bonds, 167, 389
 hydrogen molecular ion, 100
 bond length, 109
 experimental binding energy, 109
 ground state, 108
 Hamiltonian, 101
 shared states, 103
 hydrogen molecule, 114
 binding energy, 121
 bond length, 121
 ground state, 121, 136, 139
 Hamiltonian, 114
 hyperfine splitting, 759

 I , 1150
 i , 3, 1151
 reciprocal, 4
 i index, 7
 \Im , 1150
 i , 1151
 \mathcal{I} , 1151
 i -spin, 699
ideal gas
 quantum derivation, 489
 thermodynamic properties, 479
ideal gas law, 490
ideal magnetic dipole, 544
ideality factor, 258
identical particles, 138

iff, 10, [1151](#)
 imaginary part, 4
 impact parameter, 957
 incompressibility
 intro, 205
 independent particle model, *see* unperturbed shell model
 index notation, [1151](#)
 indirect band gap, 266
 indirect gap semiconductor, 245
 indistinguishable particles, 446
 intro, 188, 194
 inner product
 multiple variables, 18
 inner product of functions, 10
 inner product of vectors, 9
 insulated system, 472
 integer, [1151](#)
 intelligent designers, 1022
 intensive variable, 458
 intermediate vector bosons
 weak force carriers, 589
 internal conversion, 744
 definition, 574
 intro, 731
 internal energy, 459
 internal pair production
 intro, 731
 internal transition
 definition, 574
 interpretation
 interpretations, 28
 many worlds, 813
 orthodox, 27
 relative state, 812
 statistical, 27
 intrinsic quadrupole moment
 nuclei, 690
 intrinsic semiconductor, 247
 intrinsic state
 nuclei, 649
 inverse, [1151](#)
 inversion operator, 313, *see* parity
 ionic bonds, 172
 ionic conductivity, 237
 ionic molecules, 390
 ionic solids, 390
 ionization, 79
 ionization energy, 390
 Hartree-Fock, 379
 helium, 749
 hydrogen atom, 79
 irrotational flow, 1065
 islands of isomerism, 736
 iso, [1151](#)
 isobar
 nuclei, 574
 isobaric analog states, 702
 isobaric spin, 699
 isolated, [1152](#)
 isolated system, 472
 isomer, 735
 isomeric transition
 definition, 574
 isospin, 699
 beta decay, 1074
 isothermal atmosphere, 214
 isotones, 573
 isotope, 570
 isotopic spin, 699
 J , [1152](#)
 j , [1152](#)
 K, [1152](#)
 K , [1152](#)
 k , [1152](#)
 \mathcal{K} , [1152](#)
 K-capture, *see* beta decay
 kappa, *see* κ
 k_B , [1153](#)
 Kelvin coefficient, 277
 Kelvin heat, 277
 Kelvin relationships
 thermoelectrics, 914
 intro, 277
 ket, 10, [1133](#)
 ket notation
 spherical harmonics, 69
 spin states, 128
 kinetic energy
 nuclear decay, 714
 operator, 25
 kinetic energy operator
 in spherical coordinates, 73
 Klein-Gordon equation, 524
 kmol, [1153](#)
 Koopman's theorem, 379
 Kramers relation, 1118
 L, [1153](#)

L , 1153
 \mathcal{L} , 1153
 l , 1153
 ℓ , 1153
 L-capture, *see* beta decay
 ladder operators
 angular momentum, 503
 Lagrangian
 relativistic, 857
 simplest case, 832
 Lagrangian mechanics, 832
 Lagrangian multipliers
 derivations, 969
 for variational statements, 358
 Lamb shift, 759, 769
 lambda, *see* λ
 Landé g -factor, 768
 lanthanides, 163
 lanthanoids, 163
 Laplace equation, 1050
 solution in spherical coordinates, 1064
 Laplacian, 1134
 Larmor frequency
 definition, 561
 Larmor precession, 562
 laser, 304
 laser diode, 266
 latent heat of vaporization, 482
 lattice, 393
 basis, 393
 diamond, 413
 lithium, 395
 NaCl, 393
 bcc, 395
 diamond, 412
 fcc, 393
 lithium, 395
 NaCl, 393
 one-dimensional, 396
 primitive translation vector, 401
 reciprocal
 lithium, 410
 NaCl, 410
 one-dimensional, 403
 primitive vectors, 410
 three-dimensional, 410
 three-dimensional
 primitive vectors, 407
 unit cell, 393
 bcc, 395
 fcc, 393
 law of Dulong and Petit, 494
 law of mass action
 semiconductors, 252
 Lebesgue integration, 935
 LED, 265
 length of a vector, 9
 Lennard-Jones potential, 387
 Casimir-Polder, 388
 lifetime, 577
 excited atom, 321
 light waves
 classical, 536
 light-cone, *see* special relativity
 light-emitting diode, 265
 light-emitting diodes
 crystal momentum, 266
 lim, 1153
 linear combination, 1153
 linear momentum
 classical, 24
 eigenfunctions, 325
 eigenvalues, 325
 operator, 25
 symmetry and conservation, 296
 liquid drop model, *see* nuclei
 localization
 absence of, 330
 London forces, 387
 Casimir-Polder, 388
 Lorentz factor, 843
 Lorentz transformation, *see* special relativity
 Lorentz-Fitzgerald, *see* special relativity
 luminosity, 956
 Lyman transitions, 79
 M, 1154
 M , 1154
 m , 1154
 m_e , 1154
 m_n , 1154
 m_p , 1154
 Madelung constant, 392
 magic numbers
 intro, 576
 shell model, 610
 magnetic dipole moment, 554
 magnetic dipole transitions
 Hamiltonian, 928
 intro, 308

selection rules, *see* selection rules
 magnetic moment
 nuclei, 679
 magnetic quantum number, 66
 magnetic spin anomaly, 555
 magnetic vector potential
 classical derivation, 1054
 in the Dirac equation, 1058
 quantum derivation, 527
 relativistic derivation, 857
 magnitude, 4
 main block
 periodic table, 163
 majority carriers, 251
 maser
 ammonia, 126
 mass number, 570
 mass-energy relation, *see* special relativity
 Dirac equation
 emph, 525
 fine-structure
 emph, 761
 for nuclei, 579
 need for quantum field theory, 772
 matching regions, 952
 mathematicians, 850, 989
 matrix, 12, 1154
 maximum principle
 Laplace equation, 1050
 Maxwell relations, 478
 Maxwell's equations, 529
 Maxwell-Boltzmann distribution
 canonical probability, 453
 for given energy, 451
 intro, 214
 mean lifetime, 578
 mean value property
 Laplace equation, 1050
 measurable values, 27
 measurement, 28
 Meissner
 credit, 1100
 meson, 128
 mesons
 intro, 592
 metalloids, 163
 compared to semimetals, 236
 metals, 394
 method of stationary phase, 936
 metric prefixes, 1156

minority carriers, 251
 mirror nuclei, 595
 beta decay, 1083
 mass difference data, 710
 molar mass
 versus molecular mass etc., 1156
 mole, 458
 molecular mass, 459
 versus molar mass etc., 1156
 molecular solids, 387
 molecules
 ionic, 390
 momentum conservation
 beta decay, 1090
 momentum space wave function, 326
 Moszkowski unit, 1099
 Mott insulators, 234
 moving mass
 seespecial relativity, 840
 mu, *see* μ
 multipole expansion, 547

N, 1157
N, 1157
n, 1157
 n-p-n transistor, 259
 n-type semiconductor, 250
 nabla, 1134
 natural, 1158
 nearly-free electron model, 421
 negatron emission, *see* beta decay
 neon-19
 nuclear spin, 638
 Neumann functions
 spherical, 958
 neutrino
 needed in beta decay, 715
 neutron
 intro, 568
 mixed beta decay, 719
 neutron emission
 definition, 574
 neutron excess, 573
 neutron stars, 205, 573
 Newton's second law
 in quantum mechanics, 292
 Newtonian analogy, 26
 Maxwellian analogy, 790
 Newtonian mechanics, 21
 in quantum mechanics, 291

nitrogen-11
 nuclear spin, 640
 NMR
 spin one-half, 687
 noble gas, 155
 noble gases, 163
 noncanonical Hartree-Fock equations, 988
 nonequilibrium thermodynamics
 emph, 910
 nonexistent versus hidden, 71
 nonholonomic, 925
 Nordheim rules, *see* nuclei
 norm of a function, 10
 normal, 1158
 normalized, 10
 normalized wave functions, 22
 nu, *see* ν
 nuclear decay
 overview of data, 570
 nuclear force, 568
 meson exchange, 592
 nuclear magnetic resonance, 558
 nuclear magneton, 556, 685
 nuclear reactions
 antiparticles, 705
 nuclei
 beta vibration, 658
 do not contain electrons, 1073
 gamma vibration, 658
 internal conversion, 744
 intro, 570
 liquid drop model
 binding energy, 596
 intro, 595
 nuclear radius, 595
 Nordheim rules, 670
 pairing energy
 evidence, 583
 parity
 data, 674
 intro, 617
 perturbed shell model, 627
 rotational bands
 emph, 648
 shell model, 610
 nonspherical nuclei, 652
 Rainwater-type justification, 616
 shells
 evidence, 583
 stable odd-odd ones, 711
 unperturbed shell model, 627
 vibrational states
 emph, 644
 nucleon number, 570
 nucleons, 568
 OBEP, 594
 oblate spheroid, 684
 observable values, 27
 occupation numbers
 beta decay, 1074
 intro, 184
 single-state, 776
 octupole vibration
 nuclei, 646
 odd-odd nuclei
 emph, 575
 odd-particle shell model, 628
 omega, *see* ω
 one-boson exchange potential, 594
 one-dimensional free space
 Hamiltonian, 328
 one-particle shell model, 629
 one-pion exchange potential, 592
 Onsager reciprocal relations, 914
 OPEP, 592
 operator
 exponential of an operator, 294
 operators, 12
 angular momentum component, 65
 Hamiltonian, 26
 kinetic energy, 25
 in spherical coordinates, 73
 linear momentum, 25
 position, 25
 potential energy, 26
 quantum mechanics, 25
 square angular momentum, 67
 total energy, 26
 opposite, 1158
 orbitals, 367
 orthodox interpretation, 27
 orthogonal, 10
 orthonormal, 10
 P , 1158
 p states, 82
 p , 1158
 p-n junction, 253
 p-n-p transistor, 259
 p-state, 1158

- p-type semiconductor, 250
 parity, 312
 - alpha decay, 607
 - as a symmetry, 300
 - conservation in atomic transitions, 313
 - inversion operator, 313
 - multiplies instead of adds, 313
 - nuclei
 - data, 674
 - intro, 617
 - orbital
 - derivation, 1061
 - orbital angular momentum, 313
 - spherical harmonics
 - derivation, 873
 - symmetry and conservation, 296
 - intro, 312- parity violation
 - Wu experiment, 729
- Parseval's relation, 935
- partial wave analysis, 957
- partition function, 454
- Paschen transitions, 79
- passive viewpoint, 296
- Pasternack relation, 1118
- Pauli exclusion principle, 144
 - atoms, 155
 - common phrasing, 156
- Pauli repulsion, 164
- Pauli spin matrices, 520
 - generalized, 523
- Peltier coefficient, 269
- Peltier effect, 268
- periodic box, 220
 - beta decay, 1077
- periodic table, 155
 - full, 161
- periodic zone scheme, 420
 - intro, 245
- permanents, 143
- permittivity of space, 72
- perpendicular bisector, 1158
- perturbation theory
 - helium ionization energy, 749
 - second order, 748
 - time dependent, 315
 - time-independent, 747
 - weak lattice potential, 422
- perturbed shell model, 625
- phase angle, 1159
- phase equilibrium, 481
- phase speed, 330
- Phi, *see* Φ
- phi, *see* ϕ , φ
- phonons, 494
 - nuclei, 644
- photoconductivity
 - intro, 264
- photon, 79, 1159
 - energy, 79
 - spin value, 127
- photon packet, 792
- photons
 - density of modes, 182
- photovoltaic cell, 265
- physical symmetry
 - commutes with Hamiltonian, 297
- physicists, 28, 29, 163, 200, 204, 212, 219, 239, 245, 251, 274, 292, 307, 318, 365, 384, 402, 411, 431, 536, 537, 555, 571, 573–575, 577, 578, 580, 589, 591, 601, 602, 613, 629, 683, 703, 730, 731, 744, 777, 790, 847, 849, 873, 913, 956, 957, 959, 995, 1075, 1083, 1088, 1098–1100, 1156, 1159
- pi, *see* π
- pi bonds, 166
- Plancherel's theorem, 935
- Planck's blackbody spectrum, 197
- Planck's constant, 25
- Planck-Einstein relation, 79
- point charge
 - static, 537
- pointer states, 83
- Poisson bracket, 295
- Poisson equation, 548
- polar bonds, 167
- poly-crystalline, 395
- population inversion, 304
- position
 - eigenfunctions, 322
 - eigenvalues, 322
 - operator, 25
- positron emission, *see* beta decay
- possible values, 27
- potassium-40
 - decay modes, 713
- potential, 1159
 - existence, 1049
- potential energy

operator, 26
 potential energy surfaces, 365
 Poynting vector, 788
 prefixes
 YZEPTGMkmunpfazy, 1156
 pressure, 459
 primitive cell, 408
 primitive translation vector
 one-dimensional, 401
 reciprocal lattice, 410
 three-dimensional, 407
 principal quantum number, 75
 principle of relativity, 841
 probabilities
 evaluating, 60
 from coefficients, 29
 probability current, 966
 probability density, 117
 probability to find the particle, 21
 prolate spheroid, 684
 promotion, 169
 nuclei, 640
 prompt neutrons, 662
 proper distance, *see* special relativity
 proper time, *see* special relativity
 proton
 intro, 568
 proton emission
 definition, 574
 Psi, *see* Ψ
 psi, *see* ψ
 pure substance, 439
 p_x , 1159
 Pythagorean theorem, 846

 Q , 1159
 q , 1160
 Q-value
 alpha and beta decay, 714
 nuclei, 602
 quadrupole moment
 spin one-half, 687
 quadrupole vibration
 nuclei, 646
 quality factor, 579
 quantum chromodynamics, 568
 quantum confinement, 206
 single particle, 45
 quantum dot, 47
 density of states, 208
 quantum electrodynamics
 electron g factor, 555
 Feynman's book, 773
 intro, 588
 photon exchange
 intro, 288
 quantum interference
 intro, 22
 quantum mechanics
 acceleration, 292
 force, 292
 Newton's second law, 292
 Newtonian mechanics, 291
 velocity, 292
 quantum well, 47
 density of states, 207
 quantum wire, 47
 density of states, 208
 quark
 spin, 128
 quarks, 568
 Dirac equation, 525
 proton and neutron, 556

 R , 1160
 \Re , 1160
 r , 1161
 \vec{r} , 1162
 Rabi flopping frequency, 564
 rad, 579
 radiation
 emission and absorption, 300
 radiation weighting factor, 579
 radioactivity
 intro, 570
 radium emanation, 575
 radium X, 575
 random number generator, 29
 rare earths, 163
 RaX, 575
 Rayleigh formula
 partial wave expansion, 960
 spherical Bessel functions, 959
 RE, 575
 real part, 4
 reciprocal, 1162
 reciprocal lattice, *see* lattice, reciprocal
 recombination
 semiconductors, 252
 reduced mass

hydrogen atom electron, 73
 reduced zone scheme, 419
 intro, 244
 reflection coefficient, 350, 351, 968
 relative state formulation, 816
 relative state interpretation, 812
 relativistic corrections
 hydrogen atom, 759
 Relativistic effects
 Dirac equation, 524
 relativistic mass
 seespecial relativity, 840
 relativistic quantum mechanics
 beta decay, 1074
 relativity, *see* special relativity, 1160
 rem, 579
 residual strong force, 568
 resistivity
 electrical, 227
 resonance factor, 563
 rest mass
 seespecial relativity, 840
 restricted Hartree-Fock, 370
 reversibility, 464
 RHF, 370
 rho, *see* ρ
 roentgen, 578
 röntgen, 578
 rot, 1134, 1162
 rotational band
 nuclei, 651
 rotational bands
 seenuclei, 648

S, 1162
 s state, 1162
 s states, 82
 s, 1162
 saturated, 481
 scalar, 1162
 scattering, 347
 one-dimensional coefficients, 350
 three-dimensional coefficients, 955
 scattering amplitude, 956
 Schmidt lines, 688
 Schottky effect, 217
 Schrödinger equation, 280
 Schrödinger equation
 failure?, 810
 integral form, 963

Schrödinger's cat, 804
 second law of thermodynamics, 462
 second quantization, 790
 Seebeck coefficient, 274
 Seebeck effect, 272
 selection rules
 angular momentum conservation, 309
 derivation, 926
 electric dipole transitions, 308
 electric quadrupole transitions, 309
 magnetic dipole transitions, 308
 parity conservation, 312
 self-adjoint, 1142
 self-consistent field method, 378
 semi-conductors
 band gap, 411
 lattice, *see* lattice, diamond
 semi-empirical mass formula, 596
 semiconductor
 degenerate, 251
 direct gap, 245
 intrinsic, 247
 intro, 235
 n and p-type, 250
 semiconductor laser, 265, 266
 semiconductors
 compensation, 253
 conduction electrons per state, 247
 conduction electrons per volume, 249
 doping, 247
 holes
 intro, 234
 holes per state, 248
 holes per unit volume, 249

semimetal
 intro, 235
 separation of variables, 49
 for atoms, 151
 linear momentum, 325
 position, 322
 shell model of nuclei, 610
 shielding approximation, 151
 Shockley diode equation, 258
 SI prefixes, 1156
 sievert, 579
 sigma, *see* σ
 sigma bonds, 165
 simple cubic lattice, 415
 sin, 1162
 singlet state, 136

derivation, 509
 skew-Hermitian, 1142
 Slater determinants, 143
 small perturbation theory, 422
 solar cell, 265
 solid angle, 1142
 solids, 387

- covalent, 411
- ionic, 390
- molecular, 387

 spⁿ hybridization, 169
 space charge region, 255
 space-like, *see* special relativity
 space-time, *see* special relativity
 space-time interval, *see* special relativity
 special relativity, 839

- action, 858
- canonical momentum, 857
- causality and proper time, 847
- four-vectors, 848
 - dot product, 849
- in terms of momentum, 840
- light-cone, 848
- Lorentz force, 857
 - derivation, 859
- Lorentz transformation, 843
 - basic derivation, 844
 - group derivation, 851
 - group property, 851
 - index notation, 849
- Lorentz-Fitzgerald contraction, 842
 - derivation, 845
- mass-energy relation, 840
 - derivation, 855
- Lagrangian derivation, 858

- mechanics
- intro, 852
- Lagrangian, 856
- momentum four-vector, 854
- moving mass, 840
- derivation, 853
- Lagrangian derivation, 857
- proper distance, 846
- as dot product, 849
- proper time, 846
- relativistic mass, 840
- rest mass energy, 840
- derivation, 855
- space-like, 846
- space-time, 848
- space-time interval, 846
- causality, 847
- superluminal interaction, 847
- time dilation
- derivation, 845
- time-dilation, 842
- time-like, 846
- velocity transformation, 844
- derivation, 845
- warp factor, 847
- specific activity, 578
- specific decay rate, 577
- specific heat
- constant pressure, 461
- constant volume, 461
- values, 494
- specific volume, 458
- molar, 458
- spectral line broadening, 317
- spectrum
- hydrogen, 80
- spherical Bessel functions, 958
- spherical coordinates, 65
- spherical Hankel functions, 958
- spherical harmonics
- derivation, 1036
- derivation from the ODE, 871
- derivation using ladders, 1036
- generic expression, 873
- intro, 68
- Laplace equation derivation, 873
- parity, 873
- spherical Neumann functions, 958
- spheroid, 684
- spin, 127
- fundamental commutation relations
 - introduction, 132
- nuclei
 - data, 665
 - value, 127
 - x*- and *y*-eigenstates, 522
- spin down, 128
- spin orbitals, 367
- spin states
 - ambiguity in sign, 1042
 - axis rotation, 1041
- spin up, 128
- spin-orbit interaction
 - nucleons, 617
- spinor, 130

spontaneous emission
 quantum derivation, 794
 spontaneous emission rate, 318
 spontaneous fission, 661
 standard deviation, 86
 definition, 87
 simplified expression, 90
 Stark effect, 756
 stationary states, 286
 statistical interpretation, 27
 Stefan-Boltzmann formula, 492
 Stefan-Boltzmann law, 199
 Stern-Gerlach apparatus, 557
 stoichiometric coefficient, 482
 Stokes' theorem, 1162
 string theory
 gravity, 822
 superallowed beta decays, 1083
 superallowed decay
 beta decay, 727
 superconductivity, 228
 superfluidity
 Feynman argument, 194
 superluminal interaction
 Bell's theorem, 805
 hidden variables, 807
 many worlds interpretation, 816
 quantum, 22
 do not allow communication, 808
 produce paradoxes, 809
 relativistic paradoxes, 847
 surface tension, 642
 symmetrization requirement
 graphical depiction, 444
 identical bosons, 138
 identical fermions, *see* antisymmetrization
 indistinguishable particles, 446
 using groupings, 143
 using occupation numbers, 776
 using permanents, 143
 symmetry, 1163

T , 1163
 t , 1163
 \mathcal{T} , 1163
 tantalum-180m, 733
 tau, *see* τ
 temperature, 440, 1163
 definition, 452

Carnot, 469
 definition using entropy, 480
 intro, 184
 thermal de Broglie wavelength, 485
 thermal efficiency, 467
 thermal equilibrium, 440
 thermionic emission, 217
 thermocouple, 273
 thermodynamics
 first law, 460
 second law, 462
 third law, 474
 thermoelectric generator, 273
 thermoelectrics
 figure of merit, 908
 macroscopic equations, 910
 thermogenerator, 273
 Theta, *see* Θ
 theta, *see* θ, ϑ
 third law of thermodynamics, 474
 Thomson coefficient, 277
 Thomson effect, 277
 Thomson relationships
 thermoelectrics, 914
 intro, 277
 throw the dice, 29
 TID, 578
 time
 directionality, 819
 time dependent perturbation theory, 315
 time variation
 Hamiltonian, 281
 time-dilation, *see* special relativity
 time-like, *see* special relativity
 tissue weighting factor, 579
 total energy
 operator, 26
 total ionizing dose, 578
 TPEP, 594
 transistor, 259
 transition elements, 163
 transition metals, 163
 transition probability, 316
 transition rate, 318
 transitions
 hydrogen atom, 79
 translation operator
 crystal period, 336
 transmission coefficient, 350, 351, 968
 transparent crystals, 263

transpose of a matrix, [1146](#)
 triangle inequality, [311](#)
 triple alpha process, [583](#)
 triple product, [1163](#)
 triplet states, [136](#)
 derivation, [509](#)
 tritium, [571](#)
 triton, [571](#)
 tunneling, [348](#)
 field emission, [218](#)
 Stark effect, [759](#)
 WKB approximation, [350](#)
 Zener diodes, [261](#)
 turning point, [341](#)
 turning points
 WKB approximation, [345](#)
 twilight terms, [124](#)
 exchange terms, [125](#)
 Hartree-Fock, [375](#)
 Lennard-Jones/London force, [999](#)
 lithium hydride, [125](#)
 particle exchange rate, [288](#)
 spontaneous emission, [796](#)
 two state systems
 ground state energy, [122](#)
 time variation, [287](#)
 unsteady perturbations, [300](#)
 two-pion exchange potential, [594](#)
 two-state systems
 atom-photon model, [794](#)

u , [1164](#)
 u , [1164](#)
 UHF, [370](#)
 uncertainty principle
 angular momentum, [71](#)
 energy, [56](#), [286](#)
 Heisenberg, [23](#)
 position and linear momentum, [23](#)
 uncertainty relationship
 generalized, [95](#)
 Heisenberg, [96](#)
 unified atomic mass unit, [580](#)
 unit cell
 seelattice, [393](#)
 unit matrix, [1155](#)
 unitary
 matrix, [990](#)
 time advance operator, [294](#)
 unitary operators, [1142](#)

universal gas constant, [479](#), [494](#)
 universal mass unit, [580](#)
 unperturbed shell model, [625](#)
 unrestricted Hartree-Fock, [370](#)

V , [1164](#)
 v , [1165](#)
 \vec{v} , [1165](#)
 vacuum energy, [320](#)
 vacuum state, [779](#)
 valence band
 intro, [231](#)
 values
 measurable, [27](#)
 observable, [27](#)
 possible, [27](#)
 Van der Waals forces, [387](#)
 Casimir-Polder, [388](#)
 variational method, [107](#)
 helium ionization energy, [752](#)
 hydrogen molecular ion, [106](#)
 hydrogen molecule, [120](#)
 variational principle, [355](#)
 basic statement, [355](#)
 differential form, [356](#)
 Lagrangian multipliers, [357](#)
 vector, ℓ , [1165](#)
 vectorial product, [1165](#)
 velocity
 in quantum mechanics, [292](#)
 vibrational states
 seenuclei, [644](#)
 vibronic coupling terms, [975](#)
 virial theorem, [290](#)
 virtual work, [835](#)
 viscosity, [465](#)
 Volta potential, [219](#)
 von Weizsäcker formula, [596](#)

W , [1165](#)
 w , [1166](#)
 \vec{w} , [1166](#)
 warp factor, *see* special relativity
 wave function, [20](#)
 multiple particles, [112](#)
 multiple particles with spin, [133](#)
 with spin, [129](#)
 wave number, [333](#)
 Floquet, [402](#)
 Fourier versus Floquet, [403](#)
 simple, [14](#)

wave number vector
 Bloch function, 410
 wave packet
 accelerated motion, 340
 definition, 330
 free space, 327, 340
 harmonic oscillator, 342
 partial reflection, 348
 physical interpretation, 331
 reflection, 341
 weak force
 boson exchange
 intro, 288
 particle exchange, 589
 Weisskopf estimates, 736
 derivation, 1094
 Weisskopf unit
 electric, 1099
 Wigner-Seitz cell, 408
 WKB approximation
 connection formulae, 951
 WKB connection formulae, 952
 WKB theory, 343
 Woods-Saxon potential, 616
 work function, 217
 Wronskian, 967

X , 1166
 x , 1166
 X-ray diffraction, 432
 xi, *see* ξ

Y , 1166
 y , 1166
 Y_l^m , 1166
 yrast line, 658
 Yukawa potential, 591

Z , 1166
 z , 1166
 Zeeman effect, 755
 intermediate, 767
 weak, 767
 Zener diode, 261
 zero matrix, 1155
 zero point energy, 364
 zeroth law of thermodynamics, 440