

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-0-8493-3750-5 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Parker, Michael A.

Solid state and quantum theory for optoelectronics / author, Michael A. Parker.
p. cm.

"A CRC title."

Includes bibliographical references and index.

ISBN 978-0-8493-3750-5 (hardcover : alk. paper)

1. Optoelectronics. 2. Quantum theory. 3. Solid state physics. I. Title.

TA1750.P3725 2010

621.381'045--dc22

2009030736

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Contents

Preface.....	xvii
Author	xix
Chapter 1 Introduction to the Solid State	1
1.1 Brief Preview	1
1.2 Introduction to Matter and Bonds	3
1.2.1 Gasses and Liquids.....	3
1.2.2 Solids	4
1.2.3 Bonding and the Periodic Table.....	5
1.2.4 Dopant Atoms.....	8
1.3 Introduction to Bands and Transitions	9
1.3.1 Intuitive Origin of Bands	9
1.3.2 Indirect Bands and Light- and Heavy-Hole Bands.....	11
1.3.3 Introduction to Transitions	13
1.3.4 Introduction to Band-Edge Diagrams	14
1.3.5 Bandgap States and Defects	15
1.4 Introduction to the pn Junction	16
1.4.1 Junction Technology	17
1.4.2 Band-Edge Diagrams and the pn Junction.....	18
1.4.3 Nonequilibrium Statistics	19
1.5 Device Trends.....	21
1.5.1 Monolithic Integration of Device Types	21
1.5.2 Year 2000 Benchmarks	21
1.5.3 Small Optical Signals	22
1.5.4 Fabrication Challenges	23
1.6 Vacuum Tubes and Transistors	23
1.6.1 Vacuum Tube	23
1.6.2 Bipolar Transistor.....	24
1.6.3 Field-Effect Transistor.....	25
1.7 Brief Summary of Some Early Nanometer-Scale Devices	26
1.7.1 Resonant-Tunnel Device	26
1.7.2 Resonant-Tunneling Transistor	26
1.7.2.1 Single-Electron Transistors.....	27
1.7.2.2 Quantum Cellular Automation (QCA)	27
1.7.2.3 Aharonov–Bohm Effect Device.....	27
1.7.2.4 Quantum Interference Devices	28
1.7.2.5 Josephson Junction	28
1.8 Review Exercises.....	28
References and Further Readings.....	29
Chapter 2 Vector and Hilbert Spaces.....	31
2.1 Vector and Hilbert Spaces.....	31
2.1.1 Motivation for Linear Algebra in Quantum Theory	31
2.1.2 Definition of Vector Space.....	33

2.1.3	Hilbert Space	34
2.1.4	Comment on the Length of a Vector for Quantum Theory	36
2.1.5	Linear Isomorphism.....	37
2.1.6	Antilinear Isomorphism.....	37
2.2	Dirac Notation and Euclidean Vector Spaces	37
2.2.1	Kets, Bras, and Brackets for Euclidean Space.....	38
2.2.2	Basis and Completeness for Euclidean Space.....	39
2.2.3	Closure Relation for the Euclidean Vector Space.....	40
2.2.4	Euclidean Dual Vector Space.....	41
2.2.5	Inner Product and Norm.....	44
2.3	Introduction to Coordinate and Vector Representation of Functions	45
2.3.1	Initial View of the Coordinate Representation of Functions	46
2.3.2	Coordinate Basis Set	47
2.3.3	Introduction to the Inner Product for Functions	49
2.3.4	Representations of Functions	49
2.4	Function Space with Discrete Basis Sets	50
2.4.1	Introduction to Hilbert Space	50
2.4.2	Hilbert Space of Functions with Discrete Basis Vectors	51
2.4.3	Closure Relation for Functions with a Discrete Basis	53
2.4.4	Norms and Inner Products for Function Spaces with Discrete Basis Sets	54
2.4.5	Discussion of Weight Functions	55
2.4.6	Some Miscellaneous Notes on Notation	58
2.5	Function Spaces with Continuous Basis Sets	59
2.5.1	Continuous Basis Set of Functions	59
2.5.2	Coordinate Space.....	61
2.5.3	Representations of the Dirac Delta Using Basis Vectors.....	64
2.6	Graham–Schmidt Orthonormalization Procedure.....	65
2.6.1	Simplest Case of Two Vectors.....	65
2.6.2	More than Two Vectors	66
2.7	Fourier Basis Sets.....	66
2.7.1	Fourier Cosine Series	67
2.7.2	Fourier Sine Series	68
2.7.3	Fourier Series.....	69
2.7.4	Alternate Basis for the Fourier Series	71
2.7.5	Fourier Transform.....	71
2.8	Closure Relations, Kronecker Delta, and Dirac Delta Functions.....	73
2.8.1	Alternate Closure Relations and Representations of the Kronecker Delta Function for Euclidean Space	74
2.8.2	Cosine Basis Functions	75
2.8.3	Sine Basis Functions	77
2.8.4	Fourier Series Basis Functions	77
2.8.5	Some Notes.....	78
2.9	Introduction to Direct Product Spaces.....	79
2.9.1	Overview of Direct Product Spaces	79
2.9.2	Introduction to Dyadic Notation for the Tensor Product of Two Euclidean Vectors.....	82
2.9.3	Direct Product Space from the Fourier Series	82
2.9.4	Components and Closure Relation for the Direct Product of Functions with Discrete Basis Sets.....	84
2.9.5	Notes on the Direct Products of Continuous Basis Sets.....	85

2.10	Introduction to Minkowski Space	86
2.10.1	Coordinates and Pseudo-Inner Product	86
2.10.2	Pseudo-Orthogonal Vector Notation	86
2.10.3	Tensor Notation	86
2.10.4	Derivatives	87
2.11	Brief Discussion of Probability and Vector Components	88
2.11.1	Simple 2-D Space for Starters	88
2.11.2	Introduction to Applications of the Probability	90
2.11.3	Discrete and Continuous Hilbert Spaces	91
2.11.4	Contrast with Random Vectors	92
2.12	Review Exercises	92
	References and Further Readings	98
Chapter 3	Operators and Hilbert Space	99
3.1	Introduction to Operators and Groups	99
3.1.1	Linear Operator	100
3.1.2	Transformations of the Basis Vectors Determine the Linear Operator	100
3.1.3	Introduction to Isomorphisms	101
3.1.4	Comments on Groups and Operators	101
3.1.5	Permutation Group and a Matrix Representation: An Example	103
3.2	Matrix Representations	104
3.2.1	Definition of Matrix for an Operator with Identical Domain and Range Spaces	105
3.2.2	Matrix of an Operator with Distinct Domain and Range Spaces	106
3.2.3	Dirac Notation for Matrices	107
3.2.4	Operating on an Arbitrary Vector	109
3.2.5	Matrix Equation	110
3.2.6	Matrices for Function Spaces	113
3.2.7	Introduction to Operator Expectation Values	114
3.2.8	Matrix Notation for Averages	115
3.3	Common Matrix Operations	116
3.3.1	Composition of Operators	116
3.3.2	Isomorphism between Operators and Matrices	117
3.3.3	Determinant	118
3.3.4	Introduction to the Inverse of an Operator	120
3.3.5	Trace	122
3.3.6	Transpose and Hermitian Conjugate of a Matrix	123
3.4	Operator Space	124
3.4.1	Concepts and Section Summary	124
3.4.2	Basis Expansion of a Linear Operator	126
3.4.3	Introduction to the Inner Product for a Hilbert Space of Operators	129
3.4.4	Proof of the Inner Product	131
3.4.5	Basis for Matrices	132
3.5	Operators and Matrices in Direct Product Space	133
3.5.1	Review of Direct Product Spaces	133
3.5.2	Operators	134

3.5.3	Matrices of Direct Product Operators	134
3.5.4	Matrix Representation of Basis Vectors for Direct Product Space	137
3.6	Commutators and Algebra of Operators	138
3.6.1	Initial Discussion of Operator Algebra	139
3.6.2	Introduction to Commutators	140
3.6.3	Some Commutator Theorems.....	141
3.7	Unitary Operators and Similarity Transformations	143
3.7.1	Orthogonal Rotation Matrices	143
3.7.2	Unitary Transformations.....	146
3.7.3	Visualizing Unitary Transformations	147
3.7.4	Trace and Determinant	148
3.7.5	Similarity Transformations	148
3.7.6	Equivalent and Reducible Representations of Groups.....	150
3.8	Hermitian Operators and the Eigenvector Equation.....	151
3.8.1	Adjoint, Self-Adjoint, and Hermitian Operators	152
3.8.2	Adjoint and Self-Adjoint Matrices	154
3.9	Relation between Unitary and Hermitian Operators	156
3.9.1	Relation between Hermitian and Unitary Operators	156
3.10	Eigenvectors and Eigenvalues for Hermitian Operators	158
3.10.1	Basic Theorems for Hermitian Operators	158
3.10.2	Direct Product Space	162
3.11	Eigenvectors, Eigenvalues, and Diagonal Matrices	162
3.11.1	Motivation for Diagonal Matrices	162
3.11.2	Eigenvectors and Eigenvalues	164
3.11.3	Diagonalize a Matrix	165
3.11.4	Relation between a Diagonal Operator and the Change-of-Basis Operator	169
3.12	Theorems for Hermitian Operators.....	170
3.12.1	Common Theorems	171
3.12.2	Bounded Hermitian Operators Have Complete Sets of Eigenvectors.....	172
3.12.3	Derivation of the Heisenberg Uncertainty Relation.....	176
3.13	Raising–Lowering and Creation–Annihilation Operators	179
3.13.1	Definition of the Ladder Operators	179
3.13.2	Matrix and Basis-Vector Representations of the Raising and Lowering Operators	180
3.13.3	Raising and Lowering Operators for Direct Product Space.....	182
3.14	Translation Operators	183
3.14.1	Exponential Form of the Translation Operator	183
3.14.2	Translation of the Position Operator	184
3.14.3	Translation of the Position-Coordinate Ket	185
3.14.4	Example Using the Dirac Delta Function	185
3.14.5	Relation among Hilbert Space and the 1-D Translation, and Lie Group	186
3.14.6	Translation Operators in Three Dimensions	186
3.15	Functions in Rotated Coordinates	186
3.15.1	Rotating Functions	186
3.15.2	Rotation Operator	188
3.15.3	Rectangular Coordinates for the Generator of Rotations about z	189

3.15.4	Rotation of the Position Operator	189
3.15.5	Structure Constants and Lie Groups	190
3.15.6	Structure Constants for the Rotation Lie Group	191
3.16	Dyadic Notation.....	192
3.16.1	Notation	192
3.16.2	Equivalence between the Dyad and the Matrix	192
3.17	Review Exercises.....	193
	References and Further Reading	199
Chapter 4	Fundamentals of Classical Mechanics	201
4.1	Constraints and Generalized Coordinates.....	201
4.1.1	Constraints	201
4.1.2	Generalized Coordinates.....	202
4.1.3	Phase Space Coordinates.....	204
4.2	Action, Lagrangian, and Lagrange's Equation	204
4.2.1	Origin of the Lagrangian in Newton's Equations	205
4.2.2	Lagrange's Equation from a Variational Principle.....	207
4.3	Hamiltonian	210
4.3.1	Hamiltonian from the Lagrangian	210
4.3.2	Hamilton's Canonical Equations	211
4.4	Poisson Brackets.....	213
4.4.1	Definition of the Poisson Bracket and Relation to the Commutator.....	213
4.4.2	Basic Properties for the Poisson Bracket	214
4.4.3	Constants of the Motion and Conserved Quantities	215
4.5	Lagrangian and Normal Coordinates for a Discrete Array of Particles.....	216
4.5.1	Lagrangian and Equations of Motion.....	216
4.5.2	Transformation to Normal Coordinates	217
4.5.3	Lagrangian and the Normal Modes.....	222
4.6	Classical Field Theory	224
4.6.1	Lagrangian and Hamiltonian Density	225
4.6.2	Lagrange Density for 1-D Wave Motion	227
4.7	Lagrangian and the Schrödinger Equation	230
4.7.1	Schrödinger Wave Equation.....	230
4.7.2	Hamiltonian Density.....	231
4.8	Brief Summary of the Structure of Space-Time.....	232
4.8.1	Introduction to Space-Time Warping.....	232
4.8.2	Minkowski Space	233
4.8.3	Lorentz Transformation	236
4.8.4	Some Examples	238
4.9	Review Exercises.....	239
	References and Further Readings.....	243
Chapter 5	Quantum Mechanics.....	245
5.1	Relation between Quantum Mechanics and Linear Algebra	245
5.1.1	Observables and Hermitian Operators	246
5.1.2	Eigenstates	247
5.1.3	Meaning of Superposition of Basis States and the Probability Interpretation.....	249

5.1.4	Probability Interpretation.....	250
5.1.5	Averages.....	252
5.1.6	Motion of the Wave Function.....	254
5.1.7	Collapse of the Wave Function.....	255
5.1.8	Interpretations of the Collapse.....	257
5.1.9	Noncommuting Operators and the Heisenberg Uncertainty Relation.....	259
5.1.10	Complete Sets of Observables.....	262
5.2	Fundamental Operators and Procedures for Quantum Mechanics.....	263
5.2.1	Summary of Elementary Facts.....	263
5.2.2	Momentum Operator.....	264
5.2.3	Hamiltonian Operator and the Schrödinger Wave Equation.....	264
5.2.4	Introduction to Commutation Relations and Heisenberg Uncertainty Relations.....	266
5.2.5	Derivation of the Heisenberg Uncertainty Relation.....	267
5.2.6	Program.....	269
5.3	Examples for Schrödinger's Wave Equation.....	271
5.3.1	Discussion of Quantum Wells.....	272
5.3.2	Solutions to Schrödinger's Equation for the Infinitely Deep Well.....	273
5.3.3	Finitely Deep Square Well.....	279
5.4	Harmonic Oscillator.....	285
5.4.1	Introduction to Classical and Quantum Harmonic Oscillators.....	285
5.4.2	Hamiltonian for the Quantum Harmonic Oscillator.....	288
5.4.3	Introduction to the Ladder Operators for the Harmonic Oscillator.....	288
5.4.4	Ladder Operators in the Hamiltonian.....	290
5.4.5	Properties of the Raising and Lowering Operators.....	292
5.4.6	Energy Eigenvalues.....	294
5.4.7	Energy Eigenfunctions.....	294
5.5	Introduction to Angular Momentum.....	296
5.5.1	Classical Definition of Angular Momentum.....	296
5.5.2	Origin of Angular Momentum in Quantum Mechanics.....	297
5.5.3	Angular Momentum Operators.....	298
5.5.4	Pictures for Angular Momentum in Quantum Mechanics.....	299
5.5.5	Rotational Symmetry and Conservation of Angular Momentum.....	301
5.5.6	Eigenvalues and Eigenvectors.....	303
5.5.7	Eigenvectors as Spherical Harmonics.....	305
5.6	Introduction to Spin and Spinors.....	309
5.6.1	Basic Idea of Spin.....	309
5.6.2	Link between Physical Space and Hilbert Space.....	312
5.6.3	Pauli Spin Matrices.....	315
5.6.4	Rotations.....	317
5.6.5	Direct Product Space for a Single Electron.....	318
5.6.6	Spin Hamiltonian.....	319
5.7	Angular Momentum for Multiple Systems.....	323
5.7.1	Adding Angular Momentum.....	323
5.7.2	Clebsch–Gordon Coefficients.....	326

5.8	Quantum Mechanical Representations	330
5.8.1	Discussion of the Schrödinger, Heisenberg, and Interaction Representations	331
5.8.2	Schrödinger Representation.....	333
5.8.3	Rate of Change of the Average of an Operator in the Schrödinger Picture.....	334
5.8.4	Ehrenfest's Theorem for the Schrödinger Representation	335
5.8.5	Heisenberg Representation	337
5.8.6	Heisenberg Equation	338
5.8.7	Newton's Second Law from the Heisenberg Representation.....	339
5.8.8	Interaction Representation.....	340
5.9	Time-Independent Perturbation Theory.....	341
5.9.1	Initial Discussion of Perturbations	341
5.9.2	Nondegenerate Perturbation Theory.....	342
5.9.3	Unitary Operator for Time-Independent Perturbation Theory.....	349
5.10	Time-Dependent Perturbation Theory.....	352
5.10.1	Physical Concept	353
5.10.2	Time-Dependent Perturbation Theory Formalism in the Schrödinger Picture.....	355
5.10.3	Example for Further Thought and Questions.....	359
5.10.4	Time-Dependent Perturbation Theory in the Interaction Representation	362
5.10.5	Evolution Operator in the Interaction Representation	364
5.11	Introduction to Optical Transitions	365
5.11.1	EM Interaction Potential.....	365
5.11.2	Integral for the Probability Amplitude	367
5.11.3	Rotating Wave Approximation	369
5.11.4	Absorption	370
5.11.5	Emission	371
5.11.6	Discussion of the Results	372
5.12	Fermi's Golden Rule.....	373
5.12.1	Introductory Concepts on Probability	373
5.12.2	Definition of the Density of States.....	374
5.12.3	Equations for Fermi's Golden Rule	377
5.13	Density Operator.....	382
5.13.1	Introduction to the Density Operator	382
5.13.2	Density Operator and the Basis Expansion.....	386
5.13.3	Ensemble and Quantum Mechanical Averages.....	390
5.13.4	Loss of Coherence.....	394
5.13.5	Some Properties.....	396
5.14	Introduction to Multiparticle Systems	397
5.14.1	Introduction	397
5.14.2	Permutation Operator	399
5.14.3	Simultaneous Eigenvectors of the Hamiltonian and the Interchange Operator	401
5.14.4	Introduction to Fock States	403
5.14.5	Origin of Fock States	404
5.14.5.1	Bosons.....	406
5.14.5.2	Fermions	408

5.15	Introduction to Second Quantization.....	408
5.15.1	Field Commutators	409
5.15.2	Creation and Annihilation Operators	410
5.15.3	Introduction to Fock States	412
5.15.4	Interpretation of the Amplitude and Field Operators.....	414
5.15.5	Fermion–Boson Occupation and Interchange Symmetry	415
5.15.6	Second Quantized Operators	416
5.15.7	Operator Dynamics.....	418
5.15.8	Origin of Boson Creation and Annihilation Operators	418
5.16	Propagator.....	422
5.16.1	Idea of the Green Function	422
5.16.2	Propagator for a Conservative System	423
5.16.3	Alternate Formulation.....	424
5.16.4	Propagator and the Path Integral.....	425
5.16.5	Free-Particle Propagator	426
5.17	Feynman Path Integral.....	428
5.17.1	Derivation of the Feynman Path Integral	428
5.17.2	Classical Limit.....	430
5.17.3	Schrödinger Equation from the Propagator.....	431
5.18	Introduction to Quantum Computing	432
5.18.1	Turing Machines.....	432
5.18.2	Block Diagrams for the Quantum Computer	434
5.18.3	Memory Register with Multiple Spins.....	435
5.18.4	Feynman Computer for Negation without a Program Counter	436
5.18.5	Example Physical Realizations of Quantum Computers	439
5.19	Introduction to Quantum Teleportation.....	440
5.19.1	Local versus Nonlocal	440
5.19.2	EPR Paradox.....	441
5.19.3	Bell’s Theorem	442
5.19.4	Quantum Teleportation.....	443
5.20	Review Exercises.....	445
	References and Further Reading	458

Chapter 6	Solid-State: Structure and Phonons.....	461
6.1	Origin of Crystals	461
6.1.1	Orbitals and Spherical Harmonics.....	461
6.1.2	Hybrid Orbital	463
6.2	Crystal, Lattice, Atomic Basis, and Miller Notation.....	464
6.2.1	Lattice	464
6.2.2	Translation Operator.....	465
6.2.3	Atomic Basis	467
6.2.4	Unit Cells.....	467
6.2.5	Miller Indices.....	468
6.3	Special Unit Cells	469
6.3.1	Body-Centered Cubic Lattice.....	469
6.3.2	Face-Centered Cubic Lattice	470
6.3.3	Wigner–Seitz Primitive Cell.....	470
6.3.4	Diamond and Zinc Blende Lattice	471
6.3.5	Tetrahedral Bonding and the Diamond Structure	472

6.4	Reciprocal Lattice	472
6.4.1	Primitive Reciprocal Lattice Vectors	473
6.4.2	Discussion of Reciprocal Lattice Vector in the Fourier Series	474
6.4.3	Fourier Series and General Lattice Translations	475
6.4.4	Application to X-Ray Diffraction.....	476
6.4.5	Comment on Band Diagrams and Dispersion Curves	478
6.5	Comments on Crystal Symmetries	479
6.5.1	Space and Point Groups	479
6.5.2	Rotations.....	481
6.5.3	Defects	484
6.5.4	Introduction to Symmetries in Quantum Mechanics	484
6.6	Phonon Dispersion Curves for Monatomic Crystal	486
6.6.1	Introduction to Normal Modes for Monatomic Linear Crystal	487
6.6.2	Equations of Motion.....	491
6.6.3	Phonon Group Velocity for Monatomic Crystal.....	494
6.6.4	Three-Dimensional Monatomic Crystals.....	496
6.6.5	Longitudinal Vibration of a Rod and Young's Modulus.....	496
6.7	Classical Phonons in Diatomic Linear Crystal.....	498
6.7.1	The Dispersion Curves	498
6.7.2	Approximation for Small Wave Vector	500
6.7.3	Discussion.....	500
6.8	Phonons and Modes	502
6.8.1	Modes in Monatomic 1-D Finite Crystal with 1-D Motion and Fixed-Endpoint Boundary Conditions.....	502
6.8.2	Periodic Boundary Conditions	505
6.8.3	Modes for 2-D and 3-D Waves on Linear Monatomic Array	507
6.8.4	Modes for the 2-D and 3-D Crystal	508
6.8.5	Amplitude and Phonons	509
6.9	The Phonon Density of States	510
6.9.1	Introductory Discussion.....	510
6.9.2	The Density of States in $\vec{k}$ -Space	512
6.9.3	Density of States for 2-D Crystal Near $k = 0$ for the Acoustic Branch	514
6.9.4	Summary of Technique	515
6.9.5	3-D Crystal in Long-Wavelength Limit.....	516
6.10	Comments on Phonon Crystal Momentum	517
6.10.1	Anticipations for Momentum	517
6.10.2	Conservation of Momentum in Crystals	518
6.11	The Phonon Bose–Einstein Probability Distribution	519
6.11.1	Discussion of Reservoirs and Equilibrium.....	519
6.11.2	Equilibrium Requires Equal Temperatures	521
6.11.3	Discussion of Boltzmann Factor	522
6.11.4	Bose–Einstein Probability Distribution for Phonons	523
6.11.5	Statistical Moments for Phonon Bose–Einstein Distribution.....	524
6.12	Introduction to Specific Heat.....	526
6.12.1	Discussion of Specific Heat	526
6.12.2	Einstein Model for Specific Heat	528
6.12.3	Debye Model for Specific Heat.....	528

6.13	Quantum Mechanical Development of Phonon Fields	530
6.13.1	Basis States for Fourier Series with Periodic Boundary Conditions.....	531
6.13.2	Lagrangian for Line of Atoms	532
6.13.3	Classical Hamiltonian.....	535
6.13.4	Introduction to Quantizing Phonon Field and Hamiltonian	536
6.13.5	Introduction to Phonon Fock States.....	538
6.14	Phonons and Continuous Media.....	539
6.14.1	Wave Equation and Speed	540
6.14.2	Hamiltonian for One-Dimensional Wave Motion.....	542
	Review Exercises	543
	References and Further Readings.....	548
Chapter 7	Solid-State: Conduction, States, and Bands.....	551
7.1	Equation of Continuity	551
7.1.1	Classical DC Conduction	551
7.1.2	Collisions and Drift Mobility	553
7.1.3	Classical Equation of Continuity.....	555
7.1.4	Equation of Continuity for Quantum Particles	557
7.2	Scattering Matrices	560
7.2.1	Introduction to Scattering Theory	560
7.2.2	Amplitudes	562
7.2.3	Reflectivity and Transmissivity	563
7.2.4	Modifications for Heterostructure	567
7.2.5	Reflectance and Transmittance.....	568
7.2.6	Current-Density Amplitudes.....	569
7.3	The Transfer Matrix.....	570
7.3.1	Simple Interface.....	572
7.3.2	Simple Electronic Waveguide	573
7.3.3	Transfer Matrix for Electron-Resonant Device.....	574
7.3.4	Resonance Conditions for Electron Resonance Device	575
7.3.5	Quantum Tunneling.....	579
7.3.6	Tunneling and Electrical Contacts	580
7.4	Introduction to Free and Nearly Free Quantum Models	581
7.4.1	Potential in Cubic Monatomic Crystal.....	582
7.4.2	Free Electron Model.....	582
7.4.3	Nearly Free Electron Model.....	584
7.4.4	Bragg Diffraction and Group Velocity	587
7.4.5	Brief Discussion of Electron Density and Bandgaps.....	588
7.5	Bloch Function	589
7.5.1	Introduction to Bloch Wave Function.....	589
7.5.2	Proof of Bloch Wave Function	592
7.5.3	Orthonormality Relation for Bloch Wave Functions	594
7.6	Introduction to Effective Mass and Band Current	596
7.6.1	Mass, Momentum, and Newton's Second Law	596
7.6.2	Electron and Hole Current	599
7.7	3-D Band Diagrams and Tensor Effective Mass	602
7.7.1	E - k Diagrams for 3-D Crystals	602
7.7.2	Effective Mass for Three-Dimensional Band Structure	604
7.7.3	Introduction to Band-Edge Diagrams	609

7.8	The Kronig–Penney Model for Nearly Free Electrons	611
7.8.1	Model.....	611
7.8.2	Bands.....	614
7.8.3	Bandwidth and Periodic Potential.....	616
7.9	Tight Binding Approximation	617
7.9.1	Introduction	617
7.9.2	Bloch Wave Functions	619
7.9.3	Dispersion Relation and Bands	620
7.10	Introduction to Effective Mass Equation.....	623
7.10.1	Thesis.....	623
7.10.2	Discussion of the Single-Band Effective-Mass Equation	625
7.10.3	Envelope Approximation.....	628
7.10.4	Diagonal Matrix Elements of V_E	629
7.10.5	Summary.....	630
7.11	Introduction to $\vec{k} \cdot \vec{p}$ Band Theory	632
7.11.1	Brief Reminder on Bloch Wave Function	632
7.11.2	$\vec{k} \cdot \vec{p}$ Equation for Periodic Bloch Function.....	633
7.11.3	Nondegenerate Bands.....	634
7.11.4	$\vec{k} \cdot \vec{p}$ Theory for Two Nondegenerate Bands.....	637
7.12	Introduction to $\vec{k} \cdot \vec{p}$ Theory for Degenerate Bands.....	638
7.12.1	Summary of Concepts and Procedure	638
7.12.2	Hamiltonian for Kane’s Model.....	640
7.12.3	Eigenequation for Periodic Bloch States.....	641
7.12.4	Initial Basis Set.....	642
7.12.5	Matrix of Hamiltonian.....	643
7.12.6	Eigenvalues.....	646
7.12.7	Effective Mass	647
7.12.8	Wave Functions.....	648
7.13	Introduction to Density of States.....	649
7.13.1	Introduction to Localized and Extended States.....	649
7.13.2	Definition of Density of States.....	650
7.13.3	Relation between Density of Extended States and Boundary Conditions.....	653
7.13.4	Fixed-Endpoint Boundary Conditions.....	654
7.13.5	Periodic Boundary Condition.....	655
7.13.6	Density of k -States.....	657
7.13.7	Electron Density of Energy States for Two-Dimensional Crystal.....	659
7.13.8	Electron Density of Energy States for Three-Dimensional Crystal.....	661
7.13.9	General Relation between k and E Mode Density.....	662
7.13.10	Tensor Effective Mass and Density of States	663
7.13.11	Overlapping Bands	665
7.13.12	Density of States from Periodic and Fixed-Endpoint Boundary Conditions.....	667
7.13.13	Changing Summations to Integrals	668
7.13.14	Comment on Probability	669
7.14	Infinitely Deep Quantum Well in a Semiconductor.....	671
7.14.1	Envelope Function Approximation for Infinitely Deep Well.....	672

7.14.2	Solutions for Infinitely Deep Quantum Well in 3-D Crystal	673
7.14.3	Introduction to the Density of States	676
7.15	Density of States for Reduced Dimensional Structures	677
7.15.1	Envelope Function Approximation	678
7.15.2	Density of Energy States for Quantum Well	680
7.15.3	Density of Energy States for Quantum Wire	685
7.16	Review Exercises.....	689
	References and Further Readings.....	694
Chapter 8	Statistical Mechanics	695
8.1	Introduction to Reservoirs	695
8.1.1	Definition of Reservoir.....	696
8.1.2	Example of the Fluctuation-Dissipation Theorem	697
8.1.3	Reservoirs for Optical Emitter	698
8.1.4	Comment	698
8.2	Statistical Ensembles and Introduction to Statistical Mechanics	699
8.2.1	Microcanonical Ensemble, Entropy, and States.....	699
8.2.2	Canonical Ensemble	702
8.2.3	Grand Canonical Ensemble	704
8.3	The Boltzmann Distribution	704
8.3.1	Preliminary Discussion of States and Probability	704
8.3.2	Derivation of Boltzmann Distribution Using a Thermal Reservoir	707
8.3.3	Derivation of Boltzmann Distribution Using an Ensemble.....	708
8.3.4	Counting Degenerate States	711
8.3.5	Boltzmann Distribution for Distinguishable Boson-Like Particles.....	712
8.3.6	Independent, Distinguishable Subsystems	717
8.4	Introduction to Fermi–Dirac Distribution.....	718
8.4.1	Fermi–Dirac Distribution.....	719
8.4.2	Density of Carriers	720
8.4.3	Comments.....	722
8.5	Derivation of Fermi–Dirac Distribution	722
8.5.1	Pauli Exclusion Principle	722
8.5.2	Brief Review of Maxwell–Boltzmann Distribution	724
8.5.3	Fermi–Dirac and Bose–Einstein Distributions	725
8.6	Effective Density of States, Doping, and Mass Action	729
8.6.1	Carrier Concentrations.....	730
8.6.2	Law of Mass Action	732
8.6.3	Electric Fields	732
8.6.4	Some Comments.....	734
8.7	Dopant Ionization Statistics.....	734
8.7.1	Dopant Fermi Function	734
8.7.2	Derivation	735

8.8	pn Junction at Equilibrium	736
8.8.1	Introductory Concepts	736
8.8.2	Quick Calculation of Built-in Voltage of pn Junction	739
8.8.3	Junction Fields	741
8.9	Review Exercises	743
	References and Further Readings	745
Appendix A	Growth and Fabrication Methods	747
Appendix B	Dirac Delta Function	763
Appendix C	Fourier Transform from the Fourier Series	775
Appendix D	Brief Review of Probability	779
Appendix E	Review of Integrating Factors	787
Appendix F	Group Velocity	789
Appendix G	Note on Combinatorials	797
Appendix H	Lagrange Multipliers	799
Appendix I	Comments on System Return to Equilibrium	805
Appendix J	Bose–Einstein Distribution	809
Appendix K	Density Operator and the Boltzmann Distribution	811
Appendix L	Coordinate Representations of Schrödinger Wave Equation	813
Index		815

Preface

Commercialization has brought rapid change to technology using well-established physical principles such as infrastructure. Separating the physical principles from their device applications leads to a convenient division in a book such as this one since physical principles, concepts, and mathematical theory require only moderate revision over many years whereas the devices and processes inherent to new technology require more rapid and extensive change. However, the reader should not adopt the position that meaningful experimental work cannot be performed without first exhaustively modeling a new device. In fact, either appropriate models or the relevant parameters for existing models might not be available, and therefore the researcher would need to be guided by “informed intuition” gleaned from formal courses and experiment in the laboratory.

Optoelectronics and photonics implement and apply various forms of the “matter–light” interaction. This book primarily introduces the solid-state and quantum theory for “matter” but postpones a discussion of “light” and its interaction with matter to the companion volume *Physics of Optoelectronics*. The present book covers in some detail many of the transitional topics from the intermediate/elementary to advanced levels. Chapter 1 structures the general conceptual framework for the book regarding bonding, bands and devices. However, the concepts of some topical areas will be accessible to the reader only after digesting later chapters. Chapters 2 and 3 cover the mathematics of Hilbert spaces with the philosophy of providing conceptual pictures and an operational basis for computation without overburdening the reader with the “definition–theorem–proof” format often expected in mathematics texts. These mathematical foundations focus on the abstract form of the linear algebra for vectors and operators, and supply the “pictures” that are often lacking in studies of the quantum theory that would otherwise make the subject more intuitive. A picture does not always accurately represent the mathematics of a concept but does help in conveying the meaning or “way of thinking” about the concept.

This book provides several lead-ins to the quantum theory including a brief review of Lagrange and Hamilton’s approach to classical mechanics, a discussion of the link with Hilbert space, and an introduction to the Feynman path integral. Chapter 4 summarizes the Hamiltonian and Lagrangian formalism necessary for the proper development of the quantum theory. However, Chapter 5 provides the more fundamental connection between the Hilbert space and quantum theory as well as demonstrating the Schrödinger wave equation from the Feynman path integer. Chapter 5 discusses standard topics such as the quantum well, harmonic oscillator, representations, perturbation theory, and spin and expands into the density operator and applications to quantum computing and teleportation. Chapter 6 provides an introduction to the solid state with an emphasis on the crystalline form of matter and its implications for phonon and electronic properties required for a follow-on course in optoelectronics. Chapter 7 introduces effective mass (scalar and tensor), three different band theories (Kronig-Penney, Tight Binding, and k-p), and density of states for bulk and reduced dimensional structures. Chapter 8 provides the concepts for ensembles and microstates in detail with an emphasis on the derivation of particle population distributions across energy levels. These derivations start with entropy and incorporate indistinguishability and spin (Boson, Fermion) properties while providing clear pictures to illustrate the development.

The material has been taught for seven years in various formats to graduate research students and to undergraduates. The students come from a variety of departments but primarily from electrical and computer engineering, physics, and materials science. Beginning graduate students and advanced undergraduates can cover significant portions of this book in about 26–28 classes with 1.4 h of lecture per class. The number of classes devoted to the various topics often needs some adjustment depending on the pace of the course and the background of the students. The course devotes at least six or seven classes to the Hilbert spaces (discrete and continuous basis vectors,

projection operators, orthonormal expansions, commutators, Hermitian and unitary operators, eigenvectors, and eigenvalues), at least six or seven classes to the introductory quantum theory (quantum wells, harmonic oscillator, time-independent perturbation theory, density operator), approximately four or five classes to phonons (direct and reciprocal lattices, dispersion curves and group velocity, and density of states), five or six classes to conduction and bands (quantum equation of continuity, effective mass, band diagrams, density of states, and, most importantly, the Bloch theorem), and at least four or five classes covering statistical mechanics and its application to carrier concentration (Lagrange multipliers, Boltzmann and Fermi distributions, Fermi functions, and diodes). More advanced classes cover all of the mathematics, the classical mechanics, quantum mechanical spin and angular momentum, propagators and the Feynman path integral, tensor mass, tight-binding, and k-p band theory. However, these additional topics are not necessary to read *Physics of Optoelectronics* as a follow-on course for semiconductor emitters and detectors, and as an introduction to quantum optics. The undergraduate reader (junior–senior) will find the Hilbert space and matrices accessible along with select sections on the quantum theory including the quantum well material, the electron spin, the harmonic oscillator, and the time-independent perturbation theory, as well as all of the material on phonons. The average undergraduate will be able to handle the conduction processes, the scalar effective mass, the Kronig–Penney model, and the electron density of states.

A comment regarding the end-of-chapter review exercises should be made. The problems help one to understand and internalize the material contained in the chapter. The reader should make an effort to work through some of them. None of the problems are very difficult. However, some of the information or starting assumptions for a few of the problems have been omitted. As a result, the reader will need to understand the problem, develop a solution if possible, and then determine the range/conditions of validity.

The programs at Cornell University, Rutgers University, Syracuse University, and Rome Laboratory (AFRL) along with many publications have help mold the views presented within the text. A number of people deserve mention for assistance in various capacities over the years: Eun-Hyeong Yi, P.D. Swanson, C.L. Tang, and E.A. Schiff for research, publications, and advice; S. Thai, D.G. Daut, and R.J. Michalak for assistance with programs, committees, and funding; Z. Gajic, R.L. Liboff, J. Scafidi, M. Sussman, D. Parker, and P. Kornreich for their advice and helpful discussions; and Y. Lu, S. McAfee, P. Panayotatos, M.F. Caggiano, and J. Zhao for committee participation and discussion. Special recognition goes to the staff at Taylor & Francis for their advice and efforts to bring the text to publication while providing a sufficiently flexible schedule.

I am especially grateful to my wife Carol for her constant support, encouragement, and suggestions on various aspects of the book, and career advice. She has grown accustomed to the ever-present travel computer on many trips as well as the stacks of papers and books, reams of notes and calculation, and the long hours devoted to research and laboratory issues. I am also thankful to my students who have attended the courses and have applied the material to their research while posing challenging questions, interesting solutions, and helpful suggestions.

Michael A. Parker

Author

Dr. Michael A. Parker has developed optoelectronic theory and devices for the past several decades, taught graduate and undergraduate classes in physics and engineering at leading universities, served as a technical advisor and research scientist at a government laboratory, and founded a local firm for consulting, research, and development. He earned a PhD in physics for research in condensed matter physics with foundational work in the theory of particle physics and mathematics. He was especially interested in the quantum vacuum rich in ‘hidden’ intrinsic mechanisms with noise as the ‘rule’ rather than the ‘exception’. His post doctoral work branched into optical/photonic experiment, theory and fabrication.

Dr. Parker’s research includes applications of quantum optics (a close relative of quantum electrodynamics) in the area of noise as a conveyor of information, along with the associated areas of fabrication, experiment and theory for semiconductor emitters and novel optical logic components, optically controlled molecular processes for photodissolution, and optical processes in semiconductors and amorphous materials. Dr. Parker has publications ranging from high-impact journals to general-interest reading, patents and disclosures, conferences, and software.

1 Introduction to the Solid State

Matter, fields, and their interactions produce the world we know. Matter takes on various forms including gasses, liquids, and solids although the study of “solid state” traditionally focuses on solids and often specifically crystals. The present chapter overviews and summarizes important topics in the study of the solid state such as the origin of bands and the nature of transitions between bands. The discussion shows the transition of devices from tubes to bipolar junction transistors (BJTs) and field-effect transistors (FETs) to nanodevices.

1.1 BRIEF PREVIEW

The invention and development of new devices requires not only a clear understanding of present engineering and science practice, but also sufficient theoretical background to understand new discoveries in a variety of fields. For these reasons, we develop quantum theory from the start and then apply it to areas such as energy band theory and electrical transport. Our study concentrates on the electronic properties of solids (as opposed to gases and liquids). Modern technology primarily relies on the crystalline materials and secondarily on amorphous materials and polymers.

The present chapter introduces the various forms of matter including solids, liquids, and gases. The earliest studies of the solid state have focused on homostructures consisting of identical molecules arranged in a periodic array; these materials can be doped to enhance the electrical conduction. In contrast, heterostructures have layers of dissimilar materials. In all cases of crystalline solids, the atoms and molecules form a periodic array. The periodic structure is described by the lattice as a mathematical object consisting of a periodic array of points. The crystal is formed by adding a “cluster of atoms” (a.k.a, an atomic basis) to each lattice point—the cluster can have as few as one atom. The crystal structure has importance for the conduction properties of the material as well as many of the physical material properties such as “material hardness” and mass density, and for semiconductor processing such as for the possible cleave and etching planes. Every lattice has a reciprocal lattice that represents the k -vectors in spatial Fourier transforms. The reciprocal lattice vectors provide zone boundaries for phonon and carrier band diagrams.

The operation of the vast majority of modern electronic components can only be explained through band theory. The crystalline material structure immediately leads to the electron and hole bands. The relation between bands and crystalline structure can most easily be demonstrated by the Kronig–Penney model. This model makes explicit use of the wave nature of electrons and shows how bands arise from a one-dimensional (1-D) array of atoms. On the other hand, the K–P theory (as distinct from the Kronig–Penney model) provides a more predictive model for band structure and effective mass. The band structure produces an effective mass for the electron and hole, which can be many orders of magnitude smaller than a free-electron mass. The effective mass can most simply be calculated from the curvature of the conduction or valence band. Evidently, the effective mass has very important consequences for electrical conduction and the high-frequency performance of many devices. The bands themselves consist of very closely spaced discrete states usually termed extended states because they correspond to traveling plane waves. Purely crystalline materials do not have states in the energy bandgap. However, defects and doping result in localized states within the gap that can trap the electrons and holes in a specific region of the material.

The band structure of conventional electronic devices can only be fully described by resorting to the quantum theory, which is the study of the wave nature of material particles. Nanoscale and optoelectronic devices make extensive use of the quantum theory. Nanoscale devices have

dimensions on the order of the electron wavelength; the nanoscale ranges from 100 nm to the atomic scale. In fact, nanodevices hold special fascination for scientists and engineers in that only recently have they become possible to fabricate and engineer and they operate in the quantum regime with its myriad teases to common-sense reality. Optoelectronic devices use the interaction between light and matter, which can only be accurately described by the quantum theory. The quantum theory often describes the interaction using Fermi's golden rule, which originates in the time-dependent perturbation theory and describes how an electron can make an optical transition from one energy level to another under the action of a small perturbing electromagnetic (EM) field. A significant portion of this book introduces the quantum mechanics using the modern point of view based on abstract linear algebra and Hilbert spaces. In addition, it contains a visual approach to quantum mechanical spin and multiparticle systems.

Any description of electronic and optoelectronic devices must necessarily focus on equilibrium and nonequilibrium processes in semiconductors. Equilibrium statistics for carrier occupation numbers describes the number of carriers (e.g., in band states) for materials and devices without carrier injection (i.e., no light, no current). Applying light or voltage necessarily upsets the equilibrium conditions and changes the carrier occupation numbers. Therefore the probability that an electron occupies a given state must change and the new distribution must be described by nonequilibrium statistics. We will study the equilibrium statistics and focus on the Fermi function, carrier density, carrier recombination, and generation. We expect electrical conduction and photoconduction to involve nonequilibrium statistics to some extent. We introduce drift and diffusion currents, mobility, carrier scattering mechanisms, photoconduction, and the quasi-Fermi level.

Perhaps the majority of this book can best be summarized by the workings of the diode. The pn junction might arguably occur more often than any other electronic component in modern technology. As is well-known, the pn junction forms a diode (i.e., rectifier) that allows electrical current to flow in only one direction in the ideal case. There are many derivatives of the diode beside the pn junction diode including the Schottky diode, PIN photodetector, semiconductor laser and light emitting diode (LED), and solar cell. Some devices such as the bipolar transistor might have several pn junctions. Some components such as the Ohmic contact have diode-like junctions only by accident. Regardless of the exact device, the rectifying junctions use similar operating principles.

Needless to mention, much of the progress in technology has been through improved growth and fabrication. Crystals can now be grown one monolayer at a time with high uniformity and high purity using molecular beam epitaxy (MBE). Recent techniques permit single atoms to be positioned on a surface while lithography can pattern lateral dimensions to less than 100 Å. These techniques make it possible to engineer and directly explore the quantum world.

The study of solid state includes the transition from conventional devices and systems to those incorporating new quantum technologies. Cutting-edge state-of-the-art nanodevices using picosignals might 1 day appear in quantum computers and communication systems. Quantum technology spans a variety of devices and systems and operating principles. The Aharonov-Bohm (AB) device uses a classical electromagnetic (EM) vector potential to influence the phase of the electron wave function to produce interference effects. The single-electron transistor (SET) makes interesting use of the (resonant) tunneling effect. Small devices that produce small EM waves (RF or light) must be described by the quantum theory of EM fields. These EM waves satisfy Maxwell's equations but have amplitudes described by coherent, Fock, squeezed, or thermal optical states (or a combination). New system applications include the quantum computer, which defines a new computation class that can in principle solve classically intractable problems such as factoring large numbers for breaking Rivest-Shamir-Adleman (RSA) codes. A number of devices including the two-electron quantum dot have been investigated to make logic gates and nanowires. Integrated circuits can benefit by using nanoscale optical interconnects with their nanoscale power requirements. Communications systems potentially benefit from low-noise devices and those providing secure communications such as the entangle photon schemes.

The brief introduction in the present chapter shows the great diversity of study and applications for the solid-state and quantum theory. However, modern technology is founded on matter,

fields, and their interactions. The present course of study examines matter and the interaction with particles such as electrons and phonons. The companion volume on the physics of optoelectronics completes the story by examining the EM fields and their interaction with matter.

1.2 INTRODUCTION TO MATTER AND BONDS

Perhaps the earliest classification of matter originated with Aristotle with his terms of air, water, and earth (and fire) whereas today we examine gasses, liquids, and solids. Electronic and optical devices can use any of these forms of matter to provide functionality. The solid form of matter can be further classified according to the bonding order within the material which includes crystalline, polycrystalline, and amorphous. The present section reviews basic concepts.

1.2.1 GASSES AND LIQUIDS

Gases have atoms or molecules that do not bond to one another for a range of pressure, temperature, and volume (Figure 1.1). Argon consists of single atoms whereas hydrogen usually appears as H_2 . These molecules have not any particular order and freely move within a container.

Similar to gases, liquids also have not any atomic/molecular order and they assume the shape of the containers. Applying low levels of thermal energy can easily break existing weak bonds.

Liquid crystals have mobile molecules but a type of long-range order can exist. Figure 1.2 shows molecules having a permanent electric dipole. Applying an electric field rotates the dipoles and establishes order within the collection of molecules.

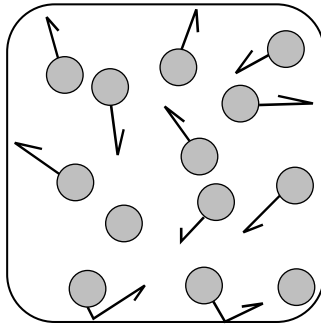


FIGURE 1.1 Gas molecules do not bind to one another.

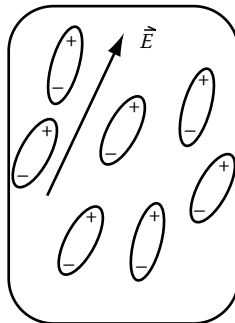


FIGURE 1.2 An electric field can rotate molecules with a permanent dipole to create order.

1.2.2 SOLIDS

Solids consist of atoms or molecules executing thermal motion about an equilibrium position fixed at a point in space. Solids can take the form of crystalline, polycrystalline, or amorphous materials. Solids (at a given temperature, pressure, and volume) have stronger bonds between molecules and atoms than do the liquids. Solids require greater amounts of energy to break the bonds.

Crystals have long-range order as indicated in Figure 1.3. Each lattice point in space has an identical cluster of atoms (atomic basis). Later chapters show how this order affects conduction and other properties. Silicon provides an example of a face-centered cubic (FCC) crystal with a two-atom basis set.

Polycrystalline materials consist of domains where the molecular/atomic order can vary from one domain to the next. Polycrystalline silicon has great technological uses for microelectrical-machines (MEMs). In general, the polycrystalline materials have medium range order that can extend over several or tens of microns. Figure 1.4 shows two domains with different atomic order. The interstitial material between the two domains has very little order, many unsatisfied bonds (dangling bonds), and regions of large voids. The growth process for polycrystalline materials can be imagined as follows. Consider a blank substrate placed inside a growth chamber. Crystals begin to grow at random locations with random orientation. Eventually the clusters meet somewhere on the substrate. Because the clusters have differing crystal orientations, the region where they meet cannot completely bond together. This results in the interstitial region.

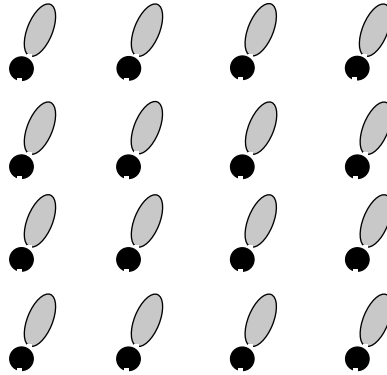


FIGURE 1.3 Crystals have identical clusters of atoms attached to lattice points in space.

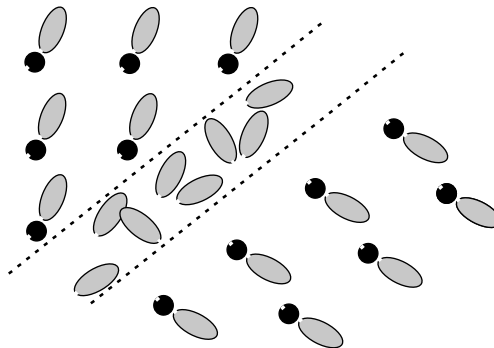


FIGURE 1.4 A polycrystalline material showing two crystal phases separated by interstitial material.

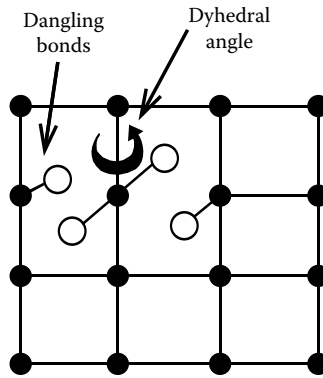


FIGURE 1.5 A rotation about the dihedral angle produces dangling bonds.

Amorphous materials do not have any long-range order but they have varying degrees of short-range order. Examples of amorphous materials include amorphous silicon, glasses, and plastics. Amorphous silicon provides the prototypical amorphous material for semiconductors. It has wide ranging and unique properties for use in solar cells and thin-film transistors. The material can be grown by a number of methods including sputtering and plasma-enhanced chemical vapor deposition (PECVD). The order of the atoms determines the quality of the material for conduction and the order depends on the growth conditions. Generally higher growth temperatures improve the quality.

In the amorphous state, the long-range order does not exist. The bonds for amorphous silicon all have essentially the same length but the dihedral angles can differ. A change in the dihedral angle occurs when two bonded atoms rotate with respect to each other about the bond axis as indicated by Figure 1.5. A cluster of fully coordinated silicon atoms produces local order but the distribution of dihedral angles yields variation in the spatial orientation of the clusters. Furthermore, some of the atoms have less than fourfold coordination and therefore have unsatisfied bonds. Under the proper preparation conditions, these dangling bonds terminate in hydrogen atoms to produce hydrogenated amorphous silicon (a-Si:H).

1.2.3 BONDING AND THE PERIODIC TABLE

Semiconductor materials generally fall in columns III through VI in the periodic table. Figure 1.6 shows a periodic table of elements. Spectroscopic notation uses the letters S, P, D, F... to denote the bonding levels. The first two columns of the periodic table correspond to the S-orbital, which requires two electrons to be stable. For example, hydrogen has only one valence electron that occupies the spherically symmetric S-orbital. Helium has two valence electrons in the S-orbital. As an exception, helium appears in the last column of the periodic table to designate it as a stable noble gas. Columns III-A through VI-A (labeled at the top of the column) plus column O represent the P-orbitals, which require six electrons for stability.

The column labeled “periods” represents the principal quantum number and the columns across correspond to electrons in shells. As will be discussed in more detail later in the book, the s-orbital refers to an electron orbital angular momentum of $\ell=0$ which has a z -component of $m=0$. The s-orbital therefore supports only the two different electron spin states of ± 1 which corresponds to hydrogen (H) (one electron in either spin state) and helium (He) (an electron in each spin state). Figure 1.7 shows the electron wave function for the s-orbital. The p-orbitals correspond to an electron orbital angular momentum of $\ell=1$ which has three possible z -components of $m=0, \pm 1$. The p-orbitals have a lobe along each axis x , y , and z which gives the name to the orbitals as p_x , p_y , and p_z , respectively (Figure 1.8). Each p-orbital can support two spin states so that the total number

Periods I A										VII A 0												
1	1.0079 H[1]	II A												1.0079 H[1]	4.00260 He[2]							
2	6.941 Li[3]	9.01218 Be[4]													10.81 B[5]	12.011 C[6]	14.0067 N[7]	15.9994 O[8]	18.9984 F[9]	20.179 Ne[10]		
3	22.9898 Na[11]	24.305 Mg[12]									VIII				III A		IV A		V A		VI A	
			III B	IV B	V B	VI B	VII B			I B	II B											
4	39.098 K[19]	40.08 Ca[20]	44.9559 Sc[21]	47.90 Ti[22]	50.9414 V[23]	51.996 Cr[24]	54.9380 Mn[25]	55.847 Fe[26]	58.9332 Co[27]	58.71 Ni[28]	63.546 Cu[29]	65.38 Zn[30]	69.72 Ga[31]	72.59 Ge[32]	74.9216 As[33]	78.96 Se[34]	79.904 Br[35]	83.80 Kr[36]				
5	85.4678 Rb[37]	87.62 Sr[38]	88.9059 Y[39]	91.22 Zr[40]	92.9064 Nb[41]	95.94 Mo[42]	98.9062 Tc[43]	101.07 Ru[44]	102.9055 Rh[45]	106.4 Pd[46]	107.868 Ag[47]	112.40 Cd[48]	114.82 In[49]	118.69 Sn[50]	121.75 Sb[51]	127.60 Te[52]	126.9045 I[53]	131.30 Xe[54]				
6	132.9054 Cs[55]	137.34 Ba[56]	[57-71]	178.49 Hf[72]	180.9479 Ta[73]	183.85 W[74]	186.2 Re[75]	190.2 Os[76]	192.22 Ir[77]	195.09 Pt[78]	196.9665 Au[79]	200.59 Hg[80]	204.37 Tl[81]	207.2 Pb[82]	208.9804 Bi[83]	(210) Po[84]	(210) At[85]	(222) Rn[86]				
7	(223) Fr[87]	226.0254 Ra[88]	[89-103]	[104]	[105]	[106]	[107]	[109]														

FIGURE 1.6 The periodic table.

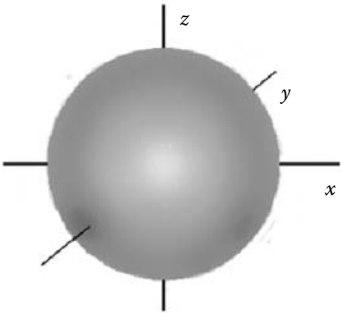


FIGURE 1.7 The wavefunction for the s-orbital is spherically symmetric.

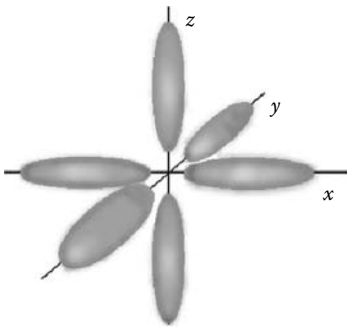


FIGURE 1.8 The p-orbitals.

of electrons in the p-orbitals comes to 6. The electronic structure of an element has the conventional notation

Element = ∏ (period)(orbital)^(number electrons)

(1.1)

where the large Pi represents a type of product (concatenation).

Example 1.1

Hydrogen needs a second electron for the S-orbital to be filled. The electronic structure of hydrogen can be written as $H = 1S^1$. We therefore expect to see hydrogen molecules as H_2 since the atoms can “share” two electrons and thereby fill their valence shells.

Example 1.2

Helium can be written as $He = 1S^2$. The outer shell is filled and the atom does not normally bond with other atoms.

Example 1.3

Silicon in column IV-A requires 4 extra electrons to fill the P level. The electronic structure has the form $Si = 1S^2 2S^2 2P^6 3S^2 3P^2$. Given the 4 electrons in 3S and 3P, we therefore expect one silicon atom to covalently bond to four other silicon atoms. Covalent bonds share valence electrons rather than completely transferring the electrons to neighboring atoms (as for ionic bonding).

Example 1.4

Silicon represents a prototypical material for electronic devices. Similarly, amorphous silicon represents a prototypical material for amorphous semiconductors. Gallium arsenide (GaAs) represents a prototypical direct bandgap material for optoelectronic components. Aluminum and gallium occur in the same column of the periodic table. We therefore expect to find compounds where an atom of aluminum can replace an atom of gallium. Such compounds can be designated by $Al_xGa_{1-x}As$ with x the mole fraction ranging from 0 to 1.

Atoms (e.g., silicon atoms) bond by virtue of electromagnetic (EM) forces and the associated EM energy. An excellent reference for the physics and chemistry of bonding can be found in the book titled *Valence* and written by Coulson. Consider two silicon atoms bonded together and sharing two electrons in the single bond. The atoms attract each other since each nucleus attracts the electrons. The situation is similar to two people each pulling on a shared object (such as a basket ball). The force on the electrons tends to pull the nuclei together. If one removes the electrons from the bonds, then the atoms no longer attract and they do not remain bonded. In fact, the net charge on the atoms would cause repulsion. In a semiconductor, adding holes to the material must therefore weaken the bonds.

The most stable atomic bonds release the greatest amount of energy during the bonding process. Figure 1.9 shows the potential energy between two atoms as a function of the distance between them. The separation distance labeled as a_0 yields a minimum in the energy. Moving the atoms

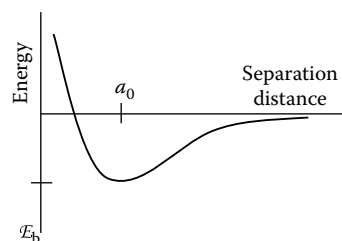


FIGURE 1.9 Total energy of two atoms as a function of their separation distance.

closer than this distance increases the energy as does moving them further apart. The binding energy $\mathcal{E}_b$ represents the approximate energy required to separate the two atoms once bonding occurs.

The atoms bond through the valence electrons, which for silicon comprises 3S and 3P. If only the 3P levels of each atom were involved with bonding, then one might expect the atoms to form a rectangular array similar to an xyz -coordinate system with an angle of 90° between bonds. In such a case, it is not clear how this bonding arrangement would give the necessary six additional electrons for each silicon atom.

Silicon (and GaAs for example) form hybrid orbits consisting of linear combinations of the 3S- and 3P-orbitals. These hybridized orbitals no longer form the rectangular array but instead have approximately 110° between bonds (as shown in Figure 1.10). In such a case, the bonding between atoms forms the tetrahedrons shown in Figure 1.11. As will be seen in Chapter 6, silicon has a FCC lattice with two atoms per lattice point (i.e., an atomic basis containing two silicon atoms).

1.2.4 DOPANT ATOMS

Adding impurity atoms can affect the electronic and optical properties of a material. Doping can be used to control the conductivity of a host crystal. n-Type dopants have one extra valence electron than the material itself. For example, we might expect phosphorus to be an n-type dopant for silicon (see Figure 1.12). Not all of the phosphorous valence electrons participate in bonding and the additional (unbonded) electrons can freely move about the crystal. p-Type dopants have one less electron in the valence shell than do the atoms in the host material. For example, boron is a p-type dopant for silicon.

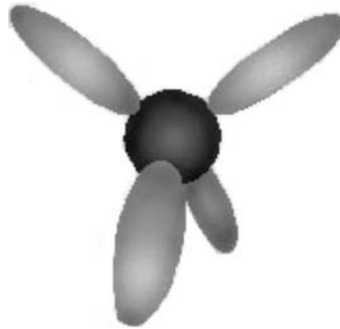


FIGURE 1.10 The hybridized s-p-orbitals have approximately 110° between the bonding states.

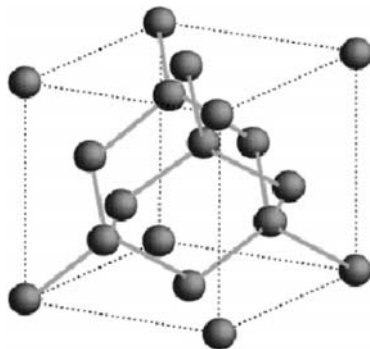


FIGURE 1.11 The s-p hybrid bonds give rise to tetrahedral bonding between the atoms. The bonding produces an FCC lattice with a atomic basis of two identical atoms (From Kittel, C., *Introduction to Solid State Physics*, 5th edn., John Wiley & Sons, New York, 1976. With permission.)

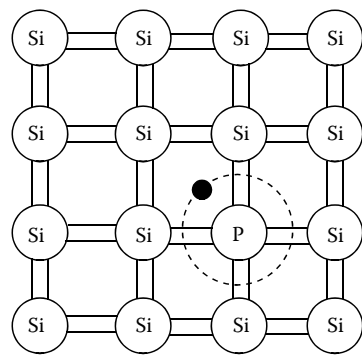


FIGURE 1.12 An n-type dopant atom embedded in a silicon host crystal. The electron is loosely bound to the dopant atom and free to roam about the crystal at room temperature.

The effects of doping on conduction can be easily seen for the n-type dopant in silicon. The “extra” fifth electron orbits the phosphorus nucleus similar to a hydrogen atom. However, the radius of the orbit must be much larger than the radius of a similar hydrogen orbit. Unlike the orbit shown in the figure, the electron orbit actually encloses many silicon atoms. The silicon atoms within the orbit can become polarized and screen the electrostatic force between the orbiting electron and the phosphorus ion. As a result, the electrons remain only weakly bonded to the phosphorus nucleus at low temperatures. These electrons break their bonds at room temperature and freely move about the crystal and thereby increase the conductivity of the crystal. For GaAs, zinc and silicon provide a p-type and n-type dopant, respectively.

1.3 INTRODUCTION TO BANDS AND TRANSITIONS

Semiconductor devices most often use the crystalline form of matter. The conduction and optical characteristics for emitters and detectors primarily depend on the band structure. The present section introduces the bands and the electronic transitions.

1.3.1 INTUITIVE ORIGIN OF BANDS

As previously discussed, a silicon atom can covalently bond to four other silicon atoms since it has four valence electrons. Figure 1.13 shows a cartoon representation (at 0 K) of the crystal and indicates adjacent atoms sharing two electrons. Adding energy to the crystal (Figure 1.14) frees electrons from the bonds so that they can move about the crystal lattice. This means that free electrons have larger energy than those electrons in the bonds. The bandgap energy represents the minimum energy required to liberate an electron. An electron that possesses this minimum amount of energy must have a potential energy equal to the gap energy. If the electron acquires more

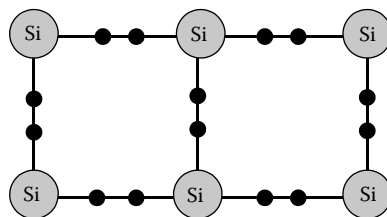


FIGURE 1.13 Cartoon representation of silicon crystal at 0 K.

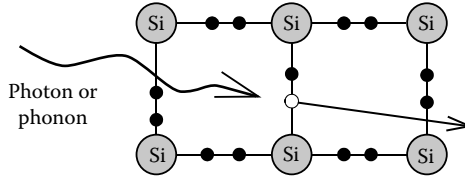


FIGURE 1.14 Cartoon representation of transition from valence band (vb) to conduction band (cb).

than the minimum, then it has not only the potential energy but also kinetic energy. The conduction band represents the energy of the free electrons (also known as conduction electrons). The vacancies left behind are “holes” in the bonding. The holes appear to move when electrons in neighboring bonds transfer to fill the vacancy. The transferred electron leaves behind another hole. The hole therefore appears to move from one location to the next. The hole acts like a positive charge; however, the neighboring atoms have net positive charge because of the missing electron in the bond.

The total energy of a conduction electron can be written as

$$E = \text{PE} + \text{KE} = E_g + \frac{1}{2}m_e v^2 \quad (1.2)$$

where the potential energy equals the gap energy E_g . Using the momentum $p = m_e v$ we can rewrite the relation as

$$E = E_g + \frac{p^2}{2m_e} \quad (1.3)$$

where m_e denotes an effective mass for the electron. Therefore, as shown in Figure 1.15, the plot of the energy E versus momentum p has a parabolic shape for the purposes of this conceptual explanation. If the electron receives just enough energy to surmount the bandgap, then it does not have enough energy to be moving and the momentum must be $p = 0$. We refer to these energy diagrams as band diagrams or dispersion curves.

The promoted electron (conduction electron in the conduction band (cb)) leaves behind a hole at the Si–Si bond. Neighboring bonded electrons can tunnel into the empty state. The holes therefore move from one site to the next. This means that the holes appear to have kinetic energy. A plot of the kinetic energy versus momentum p or wave vector k also has a parabolic shape for the holes

$$E = \frac{p^2}{2m_h} \quad (1.4)$$

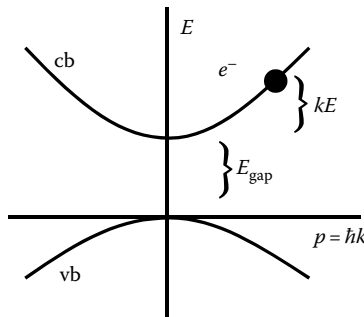


FIGURE 1.15 Band diagram showing a direct bandgap for materials such as GaAs.

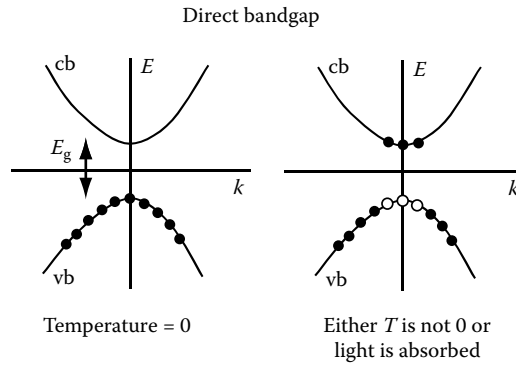


FIGURE 1.16 Electrons (solid dots) occupy states in a direct-bandgap semiconductor. The empty states (open dots) represent the empty states (holes) in the valence band (vb).

where m_h denotes the effective mass of the hole. The free holes live in the valence band and can participate in electrical conduction. The valence band has a parabolic shape similar to the conduction band. The holes behave similar to positively charged particles under the action of an electric field; however, only particles can have the property of charge. The hole has charge by virtue of the fact that when a bond loses an electron; the net charge in a small volume (encompassing neighboring atoms) centered on the bond then has a net positive charge carried by the neighboring nuclei (i.e., nuclei charge minus remaining electrons).

Some of the features of the bands require a quantum mechanical analysis. When atoms come close together to form a crystal, the energy levels for bonding split into many different energy levels. All of these split-levels from all of the atoms in the crystal produce the bands. “Bands” actually consist of a collection of “closely spaced” energy levels (see the circles in Figure 1.16). For example, the cb energies are very closely spaced and form a parabola. Sometimes people refer to these closely spaced states as “extended states” because the wave vector k indicates that electrons in these states are described by traveling plane waves.

The conduction and valence bands comprise the E versus k dispersion curve where k denotes the electron (or hole) wave vector. We imagine that the electrons (and holes) behave as waves with wavelength $\lambda = 2\pi/k$. Using the momentum $p = \hbar k$, the band diagrams can be relabeled as in Figure 1.16. The band diagram provides the energy of the electrons (and holes) as a function of the wave vector (or momentum). The stationary particles have $k = 0$ and those moving have nonzero wave vector. The E versus k diagrams are similar to the frequency ω versus k diagrams used for optics (where ω is the angular frequency related to the frequency ν by $\omega = 2\pi\nu$).

For recombination, an electron must give up excess energy for an electron to ‘drop’ into a hole which thereby eliminates both entities. Electrons and holes recombine when they collide with each other and shed extra energy by emitting photons and phonons. Regardless of the process, the total energy given up must equal or exceed the bandgap energy. The recombination of electrons and holes in direct bandgap materials produce photons (i.e., the electron loses energy and drops to the vb). These electron–hole pairs (sometime called excitons) are “emission centers” that can form the gain medium for a laser.

1.3.2 INDIRECT BANDS AND LIGHT- AND HEAVY-HOLE BANDS

The material represented by Figure 1.16 has a direct bandgap. A semiconductor has a direct bandgap when the conduction band (cb) minimum lines up with the valence band (vb) maximum (for example, GaAs). A material has an indirect bandgap (Figure 1.17) when the minimum and

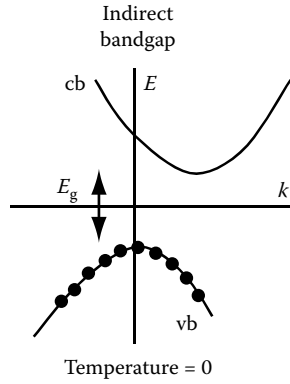


FIGURE 1.17 A semiconductor at 0 K with an indirect bandgap.

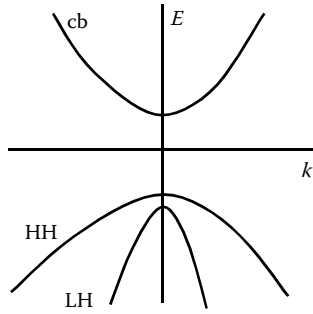


FIGURE 1.18 GaAs has a light-hole (LH) and heavy-hole (HH) band.

maximum do not have the same value for the wave vector k (silicon for example). For both direct and indirect bandgaps, the difference in energy between the minimum of the cb and the maximum of the vb equals the bandgap energy.

GaAs has light-hole (LH) and heavy-hole (HH) valence bands (see Figure 1.18). The effective mass of an electron or hole in one of the bands is proportional to the reciprocal of the band curvature according to

$$\frac{1}{m_{\text{eff}}} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k^2} \quad (1.5)$$

The HH band has holes with larger mass than the LH band. The light holes are a couple of orders of magnitude smaller than the free mass of an electron for GaAs. The effective mass m_e of a particle gives rise to the momentum according to $p = \hbar k = m_e v$. Both valence bands can contribute to the absorption and emission of light. For GaAs, the maximum of the two vb's have approximately the same energy.

Adding indium to the GaAs strains the lattice of gallium and arsenic atoms which forces them away from their normal equilibrium position in the lattice. Strain eliminates the degeneracy between the two valence bands at $k = 0$ (separates them in energy). Strain also tends to increase the curvature of the HH band, which reduces the mass of the holes in that band, and therefore increases the speed of GaAs devices. It increases the gain for lasers. It also changes the bandgap slightly and therefore also the emission wavelength of the laser.

1.3.3 INTRODUCTION TO TRANSITIONS

Consider two methods of adding energy to transition electrons from the valence band to the conduction band. First, atoms with vb electrons can absorb phonons. The phonon is the quantum of vibration of a collection of atoms about their equilibrium position. Second, atoms with electrons in the valence band can absorb a photon of light.

Figure 1.17 shows a full valence band at a temperature of $T = 0$ K. If the semiconductor absorbs light or the temperature increases, some electrons receive sufficient energy to make a transition from the valence to the conduction band. Those electrons in the conduction band (cb) and holes in the valence band (vb) are free to move and participate in electrical conduction. Each value of “ k ” labels an available electron state in either the conduction or valence band. Notice that for nonzero temperatures, the electrons reside near the bottom of the conduction band and the holes occupy the top of the valence band. Carriers tend to occupy the lowest energy states because if they had higher energy, they would lose it through collisions.

Optical transitions between the valence and conduction bands require photons with energy larger than the bandgap energy. A photon has energy $E_\gamma = \hbar\omega_\gamma$ and momentum $p_\gamma = \hbar k_\gamma$ where the wavelength is $\lambda_\gamma = 2\pi/k_\gamma$ and the speed of the photon is $v = \omega_\gamma/k_\gamma$. We expect momentum and energy to be conserved when a semiconductor absorbs (or emits) a photon. The change in the electron energy and momentum must be $\Delta E = \hbar\omega_\gamma$ and $\Delta p = \hbar k_\gamma$, respectively. However, the momentum of the photon $p_\gamma = \hbar k_\gamma$ is small (but not the energy) and so $\Delta p \cong 0$. This means that $0 = \Delta p = \hbar\Delta k$ and, as a result $\Delta k = 0$, and so the transitions occur “vertically” in the band diagram.

Figure 1.19 shows an atom absorbing energy and thereby promoting an electron to the cb. The absorbed photon has energy larger than the bandgap and the electron has nonzero wave vector k . Initially, the electron in the valence band had nonzero wave vector k (it was moving to the right). Now, the electron in the conduction band has nonzero wave vector (it also moves to the right with the same momentum as it had in the valence band). However, now the electron has more energy than the minimum of the conduction band. The electron collides with the atoms (etc.) to produce phonons and drops to the minimum of the conduction band. The produced particles must be phonons because the settling process (a.k.a., thermalization) requires a large change in wave vector and therefore a large change in momentum. Phonons have small energy but large momentum whereas photons have large energy but small momentum. Any process that involves the phonon leads to a change in the electron wave vector; this explains why phonons are involved in transitions across indirect bandgaps. As a side issue, notice the satellite valley on the conduction band in Figure 1.19 (i.e., the small dip on the right-hand side). Fast moving electrons (large k) can scatter into these valleys (intervalley scattering) which constitutes an undesirable process in most cases.

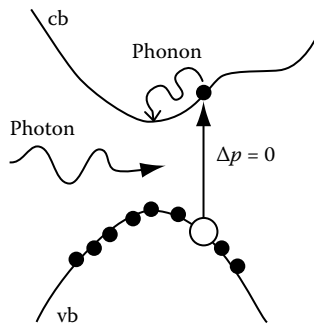


FIGURE 1.19 Optical transitions are “vertical” in the band diagram because the photon momentum is small. The electron can lose energy by phonon emission.

1.3.4 INTRODUCTION TO BAND-EDGE DIAGRAMS

Often times, we describe the workings of devices using band-edge diagrams. These diagrams plot energy versus position for the carriers inside a semiconductor. Section 1.4 uses this concept to explain the workings of the pervasive pn junction.

The band-edge diagrams (spatial diagrams) can be found from the normal E - k band diagrams (dispersion curves). Recall that a dispersion curve has axes of E versus k but does not provide any information on how the energy depends on the position variable x . In fact, there must exist one dispersion curve for each value of x (we assume just one spatial dimension) in the material. We group the states near the bottom of the E - k conduction band together to form the conduction band c for the band-edge diagram (see Figure 1.20). Similarly, we group the topmost hole states in the E - k valence band to produce the valence band for the band-edge diagram. Later chapters show the width of the levels c and v are approximately 25 meV which is much smaller than the bandgap. This is why the conduction and valence states in Figure 1.20 can be represented by thin lines labeled c and v and treated similar to distinct single states in an effective density of states approach.

Now consider the band-bending effect. Imagine a semiconductor material embedded between two electrodes attached to a battery as shown in Figure 1.21. The electric field points from right to left inside the material. An electron placed inside the material would move toward the right under the action of the electric field. We must add energy to move an electron closer to the left-hand electrode (since it is negatively charged and naturally repels electrons). This means that all electrons have higher energy near the left-hand electrode and lower energy near the right-hand electrode.

For the situation depicted in Figure 1.21, all of the electrons have higher energy near the left-hand electrode. The term “all electrons” refers to conduction and valence band electrons. This

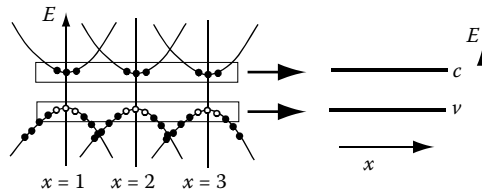


FIGURE 1.20 The states within an energy kT of the bottom of the conduction band or the top of the valence band form the levels in the band-edge diagram.

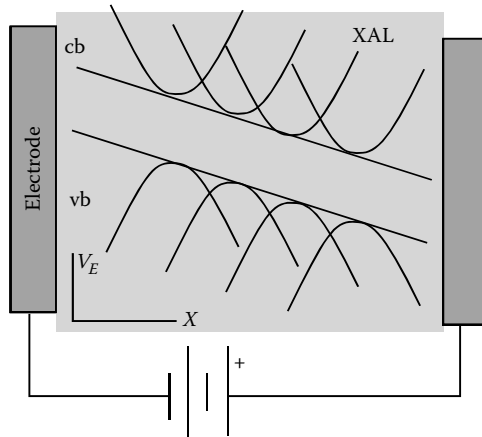


FIGURE 1.21 Band bending between parallel plates connected to a battery.

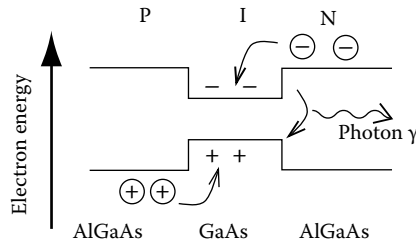


FIGURE 1.22 Band-edge diagram for heterostructure with a single quantum well.

means that near the left electrode, the E - P diagrams (i.e., E - k diagrams or dispersion curves) must shift upward to higher energy values. Once again grouping the states at the bottom of the conduction bands across the regions, we find a band edge. Similarly, we group the tops of the valence bands. When we say that the conduction band (cb) (for example) bends, we are actually saying that the dispersion curves are displaced in energy for each adjacent point x . Now we see that the electric field between the plates causes the electron energy to be larger on the left and smaller on the right. An electron placed in the crystal moves to the right to achieve the lowest possible energy. Stated equivalently, the electron moves opposite to the electric field toward the right-hand plate.

Band-edge diagrams can be used to understand a large number of optoelectronic components such as PIN photodetectors and semiconductor lasers. In fact, Figure 1.22 shows an example of a GaAs quantum well for a laser or LED having a PIN heterostructure. The doping does not extend up to the well, but remains at least 500 nm away. The bands appear approximately flat under forward bias of approximately 1.7 V. The bandgap in $\text{Al}_x\text{Ga}_{1-x}\text{As}$ is slightly larger than that for GaAs as can be seen from the approximate relation $E_g = 1.424 + 1.247x$ (eV) for $x < 0.5$. The semiconductor $\text{Al}_x\text{Ga}_{1-x}\text{As}$ has a direct bandgap for $x < 0.5$ and becomes indirect for $x > 0.5$. Barrier layers (the layers right next to the quantum well) with $x = 0.6$ provides an approximate bandgap of 1.9 eV compared with 1.3 eV for GaAs. Applying a bias voltage (positive on the left and negative on the right) to the structure causes carriers to be injected into the undoped GaAs region (well region) from the “p” and “n” regions. Electrons drop into the conduction band (cb) well and holes drop into the valence band (vb) well. The wells confine the carriers (holes and electrons) to a small region of space, which enhances the radiative recombination process and produces photons γ .

1.3.5 BANDGAP STATES AND DEFECTS

For perfect crystals, electrons can only occupy states in the valence and conduction bands (a similar statement holds for holes). The situation changes for doping and defects.

Consider the case for doping first. For simplicity, we specialize to n-type dopants such as phosphorus in silicon (refer to the discussion in connection with Figure 1.12). The electrons in Si-Si bonds require on the order of 1 eV of energy to break them free and promote them to the conduction band. Therefore, we know that the bonding electrons live in a band diagram with a bandgap on the order of 1 eV (see the band-edge diagram in Figure 1.23). However, recall that a phosphorus dopant atom has 5 valence electrons but only needs 4 of them for bonding in the silicon crystal. The 5th electron remains only weakly bonded to the phosphorus nucleus at low temperatures. Small amounts

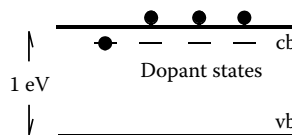


FIGURE 1.23 The n-type dopant states are very close to the conduction band.

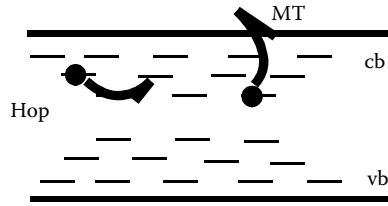


FIGURE 1.24 Amorphous materials have many bandgap states spread across a wide range of energy. Electrical conduction can occur by hopping (Hop) and multiple trapping (MT).

of energy can ionize the dopant and promote the electron to the conduction band. Therefore, the dopant states must be very close to the conduction band as shown in the figure. At very low temperatures (below 70 K), we might expect all of the Si–Si bonding electrons to be in the valence band and most of the dopant electrons to be in the shallow dopant states. As the temperature increases, more of the dopant states empty their electrons into the conduction band and the electrical conductivity must increase. By the way, the dopant states are localized states because electrons in the dopant states cannot freely move about the crystal; they orbit a nucleus in a fixed region of space.

The amorphous materials provide good examples for bandgap states arising from defects. Amorphous materials do not have perfect crystal structure. The material has many dangling bonds with 0, 1, or 2 electrons. The dangling bonds with 1 or 2 electrons require different amounts of energy to liberate an electron. For simplicity, consider dangling bonds with a single electron. These dangling bonds exist in a variety of conditions so that the electrons require a range of energy to be promoted to the conduction band (actually, for amorphous materials, the conduction band edge becomes the “mobility edge”). The dangling bonds have very high density (i.e., the number of bonds per unit volume) and occupy a wide range of energy as shown in the band-edge diagram (Figure 1.24).

Electrical conduction can proceed by two mechanisms in the amorphous materials. Hopping conduction can take place between spatially and energetically close bandgap states. The electron can quantum mechanically tunnel from one state to the next to produce current. Multiple trapping conduction takes place when conduction electrons repeatedly become trapped in the bandgap localized states and repeatedly absorb enough energy to become free again. Those electrons trapped closest to the center of the bandgap require the greatest amount of energy to be freed. At room temperature, most phonons have an energy of approximately 25 meV. Few phonons have larger energy. Therefore, those electrons in the deeper traps must wait a longer amount of time to be released to the conduction band (i.e., above the mobility edge). We therefore see that the traps lower the average mobility of the carriers by “freezing” them out for a period of time. With a little thought, you can see that the electrons tend to accumulate in the lower states. Also, these lower states near midgap tend to act as recombination centers. The electrons stay in the mid-gap traps so long, that nearby holes almost certainly collide with them and recombine.

We therefore see another facet of the bandgap states. Some act purely as temporary traps and others as recombination centers. The function of the gap states depends on their depth in the gap.

1.4 INTRODUCTION TO THE PN JUNCTION

Many modern devices use a pn junction of one form or another. For example, the semiconductor laser, LED, and detector have electronic structures very similar to a semiconductor diode. The emitter and detector use adjacent layers of p- and n-type material or p, n, and i (intrinsic or undoped) material. For the case of emitters, applying forward bias voltage controls a high concentration of holes and electrons near the junction and produces efficient carrier recombination for photon production. For the case of detectors, reverse bias voltages increase the electric field at the junction,

which efficiently sweeps out (removes) any hole-electron pairs created by absorbing incident photons. The emitting and detecting devices operate only by virtue of the matter properties and the imposed electronic junction structure. The majority of the technology preview in the present section especially that concerning Fermi levels, bands, doping, and junction behavior will become more accessible after reading later chapters.

1.4.1 JUNCTION TECHNOLOGY

The semiconductor pn junction (diode) has a special place in technology and forms an integral part of many devices. The diode has “p” and “n” type regions as shown in Figure 1.25. Gallium arsenide (GaAs) serves as a prototypical material for light emitting devices. The p-type GaAs can be made using beryllium (Be) or zinc (Zn) as dopants whereas the n-type GaAs uses silicon Si. The diode structure allows current to flow in only one direction and it exhibits a “turn-on” voltage which essentially gives the forward bias voltage that initiates conduction in the structure. In the laboratory, the turn-on voltage can be estimated using a curve tracer. One can see turn-on voltage of approximately 0.7 for Si, 0.5 for Ge, and 1.4 for GaAs. Often, the light emitters have the p-type materials on the topside of the wafer where all of the fabrication takes place.

Forward or reverse bias voltages can be applied to the diode structure. The forward bias applies an electric field parallel to the direction of the triangle (Figure 1.25). In the case of GaAs, electrons and holes move into the active region where they recombine and emit light. Reverse bias voltages can be applied to the semiconductor diode, laser, and LED to use them as photodetectors. In reverse bias, photocurrent can dominate the small amount of leakage current. Not all semiconductor junctions produce light under forward bias. Only the direct bandgap materials such as GaAs or InP efficiently emit light (a photon dominated process). The indirect bandgap materials like silicon support carrier recombination through processes involving phonons (lattice vibrations). Although indirect bandgap materials can emit some photons, the number of photons will be many orders of magnitude smaller than for the direct bandgap materials.

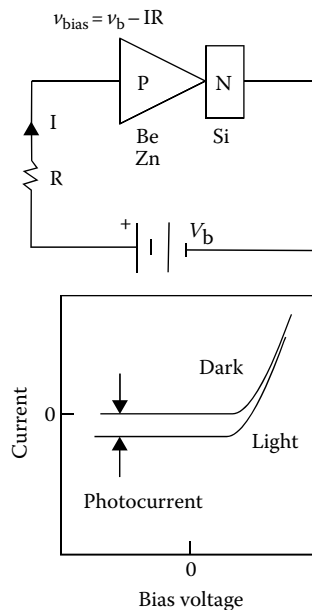


FIGURE 1.25 Forward biasing a diode (top). The I - V characteristics (bottom) show the photocurrent when the diode is reversed biased.

Semiconductor devices can be classified as homojunction or heterojunction depending on whether the device consists of a single material or two (or more) distinct materials. For the emitter, the heterojunction provides better carrier and optical confinement at the active region of the device than does the homojunction. Better confinement implies higher net gain and greater efficiency. Section 1.4.2 discusses the formation and operation of the pn homojunction. Equilibrium statistics describe the carrier distributions in a diode without an applied voltage whereas nonequilibrium statistics describe the carrier distributions for forward bias.

1.4.2 BAND-EDGE DIAGRAMS AND THE PN JUNCTION

The doping and characteristics of the material determine the properties of the pn junction. The pn diode consists of n- and p-type semiconductor layers. For the n-type material, the dopant atoms produce shallow donor states. The material should not have electrically active defects. Similar comments apply to the p-type material. Naturally, the doped crystalline materials most easily satisfy these requirements. However, it is possible to form pn junctions in amorphous materials under the appropriate conditions. In general, the doping process “grows” mobile holes and electrons into the material. Applying an electric field causes the electrons in the cb to move from negative to positive (opposite to the direction of the applied field); holes move parallel to the applied field.

A cartoon representation of the conduction and valence bands versus distance into a material appears in Figure 1.26. The position of the Fermi level in the bandgap indicates the predominant type of carrier. For p-type, the Fermi level E_F has a position closer to the valence band and the material has a larger number of free holes than free electrons. Similarly, a Fermi level E_F closer to the conduction band implies a larger number of conduction electrons. When the n- and p-type materials are isolated from each other, “excess” electrons in the n-type and holes in the p-type cannot come into equilibrium with each other and hence the Fermi levels (that represent statistical equilibrium) do not necessarily line up with each other.

Figure 1.26 shows an initial configuration for spatially separated and electrically isolated p- and n-type materials. Bringing the p- and n-type materials into contact forms a diode junction and forces the two Fermi energy levels to line up while approximately maintaining the their position relative to each band except in the junction region. The final band diagram requires the conduction and valence

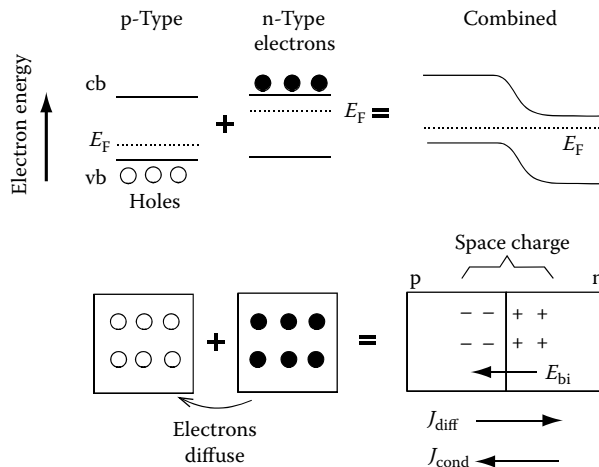


FIGURE 1.26 Combining two initially isolated doped semiconductors produces a pn junction with a built-in voltage (top). The built-in voltage is associated with a space charge region produced by drift and diffusion currents.

bands to “bend” in the region of the junction. The “band” represents the energy of electrons or holes. So, to bend the band, energy must be added or subtracted in regions of space. We know from electrostatics that electric fields can change the energy.

Why do the two Fermi levels come into coincidence? It would perhaps be easiest to imagine a fictional material with many states between the conduction band edge c and the valence band edge v . Assume two instances of the material, denoted A and B, have different Fermi levels such as $E_{FA} < E_{FB}$. The Fermi level represents the states (with that energy) that are 50% likely to have an electron. In this fictional case, the states in B with electrons will have larger energy than those states in A with electrons. Then the system minimizes the total energy for the electron distribution (more accurately, maximizes the entropy), the higher energy electrons in B will move to the vacant lower energy states in A. The increased number of electrons in A at lower energy necessarily moves the Fermi level E_{FA} to higher energy while decreasing the Fermi level E_{FB} . The process continues until $E_{FA} = E_{FB}$ since then, the electron flow from A to B will match that from B to A. This mechanism produces a built-in field that modifies the energy levels. This can be equivalently stated that at a given energy (either in A or B) for a state, the probability that an electron occupies the state must be the same. If the probabilities were not the same, then electrons would move until the probabilities equalize (for equilibrium).

What causes the electric field? When the two pieces of material come into contact, the electrons can easily diffuse from the n-type material to the p-type material; similarly, holes diffuse from “p” to “n.” This flow of charge maximizes the entropy and establishes equilibrium for the combined system. For example, the diffusion process might be pictured similar to the process occurring when a single blue drop and a single red drop of dye are spatially separated in a glass of water; each drop spreads out and eventually intermixes by diffusion. Unlike the dye drops, the holes and electrons carry charge and set up an electric field at the junction as they move across the interface. The diffusing electrons attach themselves to the p-dopants on the p-side (i.e., recombine with holes) but they leave behind positively charged cores. The separated charge forms a dipole layer (i.e., opposite charge separated through a distance). The direction of the built-in electric field prevents the diffusion process from indefinitely continuing. We define the diffusion current J_d to be the flow of positive charge due to diffusion alone (the figure shows positive charge diffusing to the right across the junction). We define the conduction current J_c to be the flow of positive charge in response to an electric field alone. Figure 1.26 shows that positive charge would flow from left to right under the action of the built-in field. Equilibrium occurs when $J_c = J_d$. The particles stop diffusing because of the established built-in field; an electrostatic barrier forms at the junction. Electrons on the n-side of the junction would be required to surmount the barrier to reach the p-side by diffusion; for this to occur, energy would need to be added to the electrons. Diffusion causes the two Fermi levels to line-up and become flat. The Fermi energy E_F is really related to the probability that an electron will occupy a given energy level.

1.4.3 NONEQUILIBRIUM STATISTICS

Section 1.4.2 discusses how n- and p-type semiconductors brought into contact establish statistical equilibrium for the junction. Applying forward bias to the diode produces a current and interrupts the equilibrium carrier population. Basically, any time the carrier population departs from that predicted by the Fermi–Dirac distribution, the device must be described by nonequilibrium statistics.

How should nonequilibrium situations be described? To induce current flow, we need to apply an electric field to reduce the electrostatic barrier at the junction so that diffusion again proceeds as shown in Figure 1.27. The built-in electric field E_{bi} (for the equilibrium case) points from “n” to “p” and so we must apply an electric field E_{appl} that points from “p” to “n” to reduce the total field and the barrier. This requires us to connect the p-side of the diode to the positive terminal of a battery and the n-side to the negative terminal. The figure shows how the applied voltage V reduces

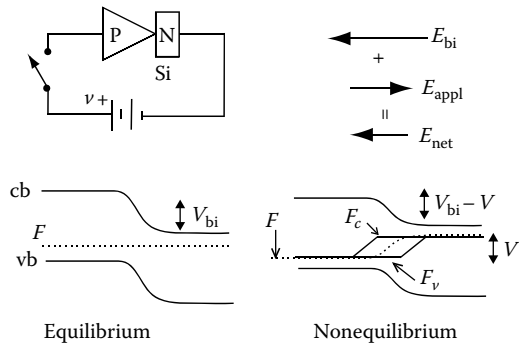


FIGURE 1.27 Band-edge diagrams for a PN diode in thermal equilibrium (no bias voltage) and one not in equilibrium (switch closed). The Fermi-level is flat for the case of equilibrium. However for the nonequilibrium case, the single Fermi level splits into two quasi-Fermi levels. The dotted line on the right hand side shows the position dependent Fermi level.

the built-in barrier and allows diffusion current to surmount the barrier. Notice also that the Fermi level is no longer flat in the junction region. The applied field is proportional to the gradient of the Fermi energy E_F . The hole and electron density in the “n” and “p” regions are described by the quasi-Fermi energy levels F_v and F_c , respectively. The quasi-Fermi levels describe nonequilibrium situations. The separation between the two quasi-Fermi levels can be related to the applied voltage. Studies of semiconductor optical sources use the quasi-Fermi levels to indicate a population inversion in a semiconductor to produce lasing.

The absorption of light by a semiconductor (without any bias voltage) shows the reason for using quasi-Fermi levels. Consider Figure 1.28. The semiconductor absorbs photons with energy larger than the bandgap $E_g = E_c - E_v$ by promoting an electron from the valence band to the conduction band. Therefore, shining light on the material produces more electrons in the conduction band and more holes in the valence band. For the intrinsic semiconductor, the number of holes and electrons remain equal. However, if we insist on describing the situation with a single Fermi level (F), then moving it closer to one of the bands increases the number of carriers in that band but reduces the number in the other. Therefore the single Fermi level must split into two in order to increase the number of carriers in both bands. The energy difference between the electron quasi-Fermi energy levels and the conduction band provides the density of electrons in the conduction band (a similar statement holds for holes and the valence band).

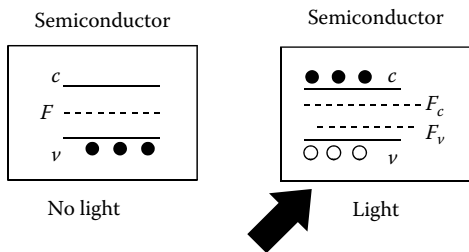


FIGURE 1.28 Light shining on a semiconductor produces two quasi-Fermi levels. The position of the quasi-Fermi levels indicate more electrons in the conduction band and more holes in the valence band than predicted by thermal equilibrium statistics.

1.5 DEVICE TRENDS

Developing low power, small, lightweight, optoelectronic components and subsystems comprises a primary trend for improving the performance of technological systems. Significant research focuses on physical systems having a small number of particles as well as those producing small or “fragile” signals. The small size naturally leads to higher speed by reducing signal propagation and interaction times. In addition, these devices will need to (and do) dissipate lower power than the conventional devices in order to keep operating temperatures low despite the higher integration density. Those devices producing small signals can have poor signal-to-noise ratios (SNRs) as well as low dynamic range that dramatically affect the performance of analog and digital devices.

1.5.1 MONOLITHIC INTEGRATION OF DEVICE TYPES

Signal processing systems perhaps impose the greatest demand on modern semiconductor technology. There exists a great need for higher performance signal processors that incorporate improved interconnects/links, greater storage capacity, better components, and miniaturization technologies. Realistic programs aim to design and implement high-performance RF signal processors with a tenfold improvement in the size, weight, and power requirements over presently available processors. The signal processor could consist of a variety of technologies including revolutionary nanoscale components, optoelectronic components, optical interconnects (rather than electrical connections between boards or chips), memory, and micromachines all of which are monolithically integrated on a chip. The medium that transports the optical signals could be free space, fiber, or monolithically integrated optical or electronic waveguides. At present, optical interconnects are important for long-haul transmission (on the order of kilometers) between global systems. The highest possible speed, using nanometer-scale components would be approximately 100 THz with an ultimate packing density of approximately 10 Tbit/cm². This ultimate speed and packing density are based on the speed of light between components that have atomic dimensions. Fewer atoms imply smaller signals, lower power dissipation and higher speeds.

A large amount of research focuses on small, low power, integrated devices for RF digital receivers, signal processors, and communications equipment. The trend continues toward circuits with the optics and electronics monolithically integrated on a single wafer and away from large power hungry multichip modules. Chip manufacturers agree on the need to further decrease size and power. These requirements pose significant problems for both the design and fabrication of the components.

Present trends reduce large-scale systems such as optical spectrum analyzers or blood pathogen analyzers to integrated form by incorporating micro-optical-electric machines (MOEMs). These integrate moveable devices include small motors and mirrors and diffraction gratings with sizes ranging smaller than a millimeter down to microns and smaller for proposed nanomachines. The micromachines are fast, rugged and use negligible power to function as switches, focusing elements and actuators. One can imagine integrating micro- and nanoelectronics with the MOEMs to incorporate a microprocessor for control of the system based on collected data.

1.5.2 YEAR 2000 BENCHMARKS

The progression to more highly integrated circuits and systems continues to be the trend. Present day electronics began in the early 1900s with the vacuum tube which gave way to the transistor in the late 1940s and then to integrated circuit soon afterwards in the 1960s. Along with the change in size from tens of centimeters on a side to tens of nanometers, the power requirements also transitioned from Watts to nano-Watts by the 1990s and early 2000s.

As a benchmark, the commercial components in the year 2000 have minimum sizes on the order of 200 nm for DRAM, $3 \times 30 \mu\text{m}^2$ for in-plane (edge-emitting) lasers, $10 \times 10 \mu\text{m}^2$ for VCSELS (with thresholds as low as 0.2 mA), 200 nm gate lengths for FETs, and 1000 nm pixel sizes for CD

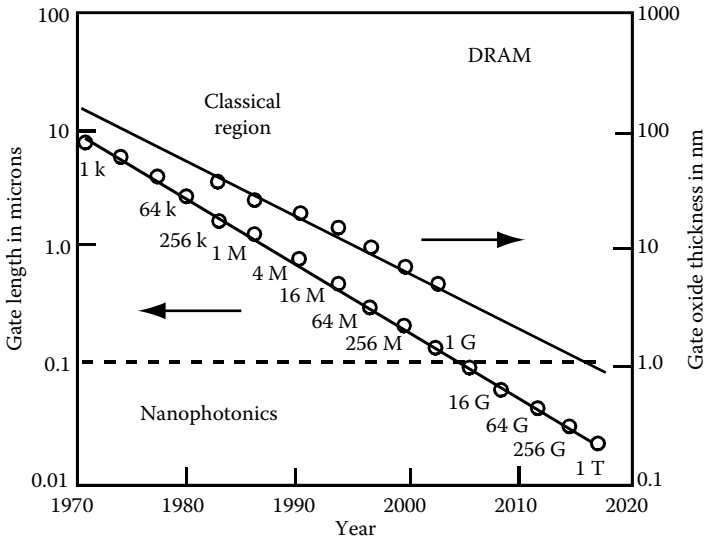


FIGURE 1.29 Device trends. (After Ando, T. et al. (Eds.), Mesoscopic Physics and Electronics, Springer, Berlin, Germany, 1998. With permission.)

ROM. Nanophotonic (i.e., nano-optoelectronic) components have features smaller than 100 nm, which corresponds to roughly 10^6 atoms or less. This trend appears in Figure 1.29.

1.5.3 SMALL OPTICAL SIGNALS

Significant research focuses on physical systems having a small number of particles ($<10^6$). Here, the term “system” can refer to either a device or to a signal for which the “particles” can be either the atoms in a device or for example, photons in an optical signal. Small optical signals can have poor signal-to-noise ratio (SNR) and low dynamic range that dramatically affect the performance of analog and digital devices. The statistics can be improved by increasing the number of particles in the system. However, such systems can involve large sizes and high power.

Devices with fewer atoms tend to produce smaller signals, lower power dissipation, and higher speeds. Low particle-count physical systems require nonclassical descriptions since (1) small volumes tend to emphasize quantization effects (electron wavefunctions in first quantization) and (2) the collective properties of a small number of atoms or photons are not adequately described by ensemble averages. The departure from the ensemble average (i.e., noise), as represented by a probability distribution, becomes significant for small physical systems that are best described by a form of quantum electrodynamics (QED) embodied by the study of quantum optics.

Smaller sized optical emitters for example might be expected to produce smaller optical signals and therefore have smaller SNRs. For example, a single atom spontaneously emitting photons at a maximum rate of $R_\gamma = 10^9$ (per second) with an external modulator operating at $R_b = 10^9$ (bits per second) produces an optical signal with an average of only 1 photon per bit, which leads to very poor SNR. Similar reasoning can be applied to an optical emitter with N atoms having lattice constant “a” and operating at the Poisson noise limit (shot-noise limit) whereby the variance in the photon number σ_n^2 equals the average number $\bar{n}$ (in a given period of time). As a rough calculation, the size of the device required to produce an SNR of $S \sim \bar{n}/\sigma_n$ would be on the order of

$$L = N^{1/3}a = a \left[\frac{S^2 R_b}{R_\gamma} \right]^{1/3}$$

For such a case, an SNR of 100 requires a minimum of 21 atoms per side or about 70–140 Å per side depending on the material. Reasonable limits on the quantum noise for a predetermined optical state can therefore place limits on the minimum size of the device. Smaller sized devices can be realized and modulated to provide useful signals so long as the noise can be further reduced or alternate methods of modulation can be developed. One such method employs squeezed optical signals as discussed in the companion volume on the physics of optoelectronics. These signals do not require exotic devices as signal generators, but as is well known, can be studied using LEDs and lasers.

1.5.4 FABRICATION CHALLENGES

Besides new applications of physical theory to engineering practice, the fabrication of nanometer scale circuits and devices poses technical challenges. It is relatively easy to grow layers as thin as one or two atoms into a material by using MBE. However, it is difficult to laterally pattern structures of similar sizes. One important challenge for the commercial sector is to develop reliable, low-cost lithography (or other means) capable of producing lateral feature sizes smaller than 100 nm. Sponsored programs continue to develop optical photoresist suitable for 80 nm feature sizes and smaller. Conventional electron lithography produces devices with feature sizes as small as 10 nm although the process remains expensive. Generally, smaller devices require material with smaller and fewer defects than for larger ones. For the fabrication, either the number of process steps must decrease or the reliability of each step must increase since the probability of at least one defective component on a wafer increases with the number of components. To minimize the number of wasted die, manufacturers of integrated circuits use extensive process development and calibration; they do not easily change even the smallest of details or process steps. As is evident, the development of new components places new requirements on design, materials, and fabrication. Appendix A summarizes and reviews typical microelectronic fabrication processes.

1.6 VACUUM TUBES AND TRANSISTORS

Technology has evolved from vacuum tube rectifiers and amplifiers to the solid-state diode and transistor. Reducing the size and power of devices generally requires new operating principles as evidenced by tracing the evolution of electronic components from the vacuum tube to the nanoscale device. However, it should be emphasized that “outdated devices” do not necessarily imply “outdated operating principles.” For example, electromagnetic (EM) fields still control charged particles even for nanodevices.

1.6.1 VACUUM TUBE

Some of the smallest vacuum tubes, the acorn tubes, have sizes on the order of several cubic centimeters. As with many tubes (Figure 1.30), they have an anode, cathode, and grid. The filament “boils off” electrons into a region of the tube having an electric field. The plate has a higher potential than the cathode and accelerates the electrons through the vacuum. A grid voltage controls the current flow by applying a repelling force to the electrons. The tube filaments dissipate on the order of 1 W of electrical power. For 5 mA of plate current, the power dissipation for 200 V between the cathode and the anode must also be on the order of 1 W. The vacuum tube device is considered to be voltage controlled since the current flow into the grid is negligible and therefore has large input impedance. As food for thought, it might be advantageous to resurrect the vacuum tube in a revised form for nano-devices because of the simple methods of electron control employed.

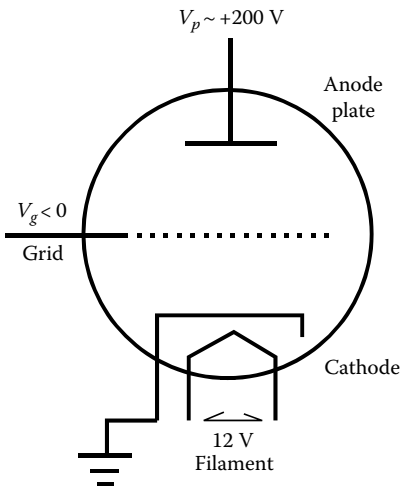


FIGURE 1.30 A typical triode vacuum tube.

1.6.2 BIPOLAR TRANSISTOR

The advent of the semiconductor devices such as the diode, bipolar junction transistor (BJT) and later, the field-effect transistor (FET), reduced the size and power requirements by several orders of magnitude.

As a homojunction device, the BJT uses a single type of semiconductor but in a sandwich configuration with a doping profile of either NPN or PNP. Figure 1.31 shows an NPN transistor. Forward biasing the base (p-type) with respect to the emitter (n-type) reduces the electron-barrier potential. Electrons then diffuse across the base-emitter (BE) junction from the emitter into the p-type base. The base has relatively small width making it easy for the injected electrons to diffuse across the base region and into the large base-collector (BC) junction field where they sweep out to the collector. The base bias voltage produces a small base current I_b but lowers the voltage across the entire BE

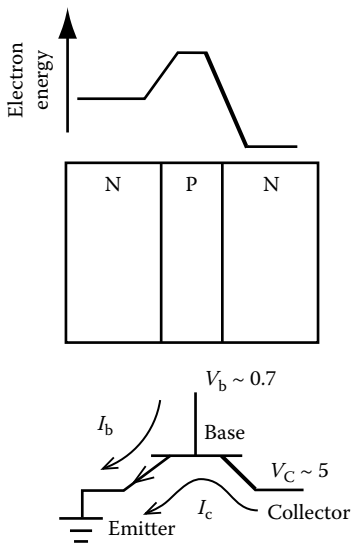


FIGURE 1.31 The NPN BJT. The top diagram shows the electron energy (band-edge diagram). The bottom diagram shows the typical biasing scheme for active region operation for a silicon device.

junction. The lower barrier voltage for all emitter electrons produces the large CE current flow I_c . The current gain can be written as $\beta = I_c/I_b$, which has a magnitude of approximately 150 for a typical small-signal transistor. The base bias voltage controls the current flow by lowering the barrier even though the BJT is considered to be a current-controlled device. The doping levels producing the built-in junction fields and the diffusion length of the electrons and holes limit the smallest size of the BJT. These transistors can have sizes on the order of tens of microns or less.

1.6.3 FIELD-EFFECT TRANSISTOR

Operationally, the field-effect transistor (FET) and vacuum tube are voltage-controlled devices. However, because these FET solid-state devices use a semiconductor platform, the operating principles must be quite different from the tube. As shown in Figure 1.32, the FET has a source, drain, and gate in analogy with the cathode, anode, and gate (i.e., screen) for the vacuum tube. The source (s) and drain (d) make ohmic contact with the channel.

The plot of electron energy versus depth into the device assumes an applied gate voltage (with respect to the source) of $V_g = 0$. In this case, the conduction band edge in the channel c has energy smaller than the Fermi energy E_F . The channel therefore has conduction electrons. A voltage impressed across the ds electrodes produces channel current. Applying a negative gate voltage (as shown in the bottom of the figure), shifts the conduction band edge to energy larger than E_F and exponentially reduces the conduction electrons. As a result, the FET turns-off and the current flow stops. In Figure 1.33, the barriers b confine the electrons to the channel c ; the top one also prevents

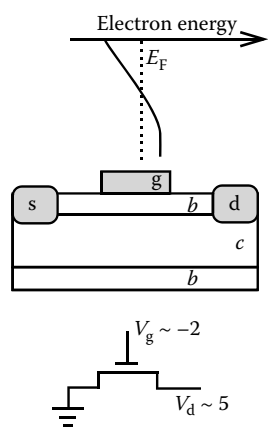


FIGURE 1.32 The n-channel FET. The top diagram shows the quiescent band-edge diagram. The bottom diagram shows the bias required for active region operation.

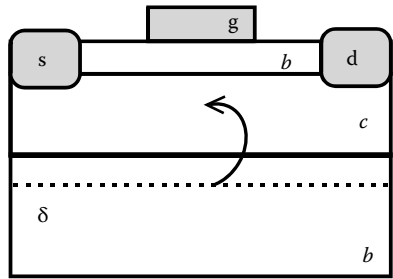


FIGURE 1.33 The HEMT structure. The delta doped layer outside the channel releases electrons to the channel.

current from flowing between the gate and the channel. Although not shown, the band edge position with respect to the Fermi level depends on the horizontal position within the channel. The field effect device, although perhaps conceptually simpler than the BJT, requires considerable effort to control the surface charges that otherwise affect the channel field and randomly shift the band edge. The smallest lithography feature size and the thickness of the barrier limit the smallest size of the FET. Thin barriers (along the vertical direction) can conduct tunneling current between the gate and the channel. Sufficiently small channels (along the horizontal direction) must be described by ballistic transport and the contact resistance dominates the channel resistance.

The gate width (along the horizontal direction) and the carrier mobility limit the switching speed. Reducing the device size and especially the gate must decrease the transition time and thereby increase the switching speed. The scattering mechanisms in the channel limit the carrier mobility. The high electron mobility transistor (HEMT) greatly reduces the electron-ion scattering by separating the dopant atoms from the channel. Figure 1.33 shows a heterostructure FET with epitaxially grown barriers and a channel. Outside the channel c , a very thin but highly doped layer (i.e., δ -doped layer) releases electrons to the channel well region. These electrons conduct current but do not suffer the electron-ion scattering since they are separated from the dopant ions.

1.7 BRIEF SUMMARY OF SOME EARLY NANOMETER-SCALE DEVICES

A number of nanometer-scale (quantum) devices have been developed by employing evolutionary or revolutionary approaches. An evolutionary approach scales down presently available devices making only minor modifications to the construction and operating principles. A revolutionary approach develops new devices with new operating principles; these new devices do not resemble former ones with respect to design and construction.

This section describes some of the well-known nanoscale devices. The resonant-tunneling diode (RTD) and resonant-tunneling transistor (RTT) both use a “quantum size” layer positioned between two tunneling barriers. With the proper source-to-drain or gate voltage, carriers can tunnel through the barriers to produce current. The single electron transistor (SET) is somewhat similar to the RTT except that only one electron can enter the gate region at one time (due to coulomb repulsion). The quantum dot (QD) has two free electrons within its volume; the configuration of these two electrons determines the state of the device. Aharanov–Bohm type devices use a magnetic field to change the phase of the electron wave function so that the wave function either constructively or destructively interferes with itself so as to produce “on” or “off” states, respectively.

1.7.1 RESONANT-TUNNEL DEVICE

A resonant-tunneling device (Figure 1.34) is similar in construction to a pn diode with two terminals. The device has a quantum well layer separated from two conducting regions by quantum barriers (on the order of 50–100 Å thick). The barriers can be AlGaAs, for example, while the well and conduction regions can be GaAs. Conduction through the device occurs by quantum tunneling through the barriers. The voltage must be adjusted across the device so that the energy levels in the well have the same energy as the conduction band on the electron injection side (left side of the figure).

1.7.2 RESONANT-TUNNELING TRANSISTOR

A resonant-tunneling transistor (RTT) is similar in construction to the resonant-tunneling device except, in addition to a source and drain, it has a gate electrode attached to the well region. Electrons are injected into the device through the source electrode. An applied gate voltage adjusts the energy of the well levels with respect to the energy range of the occupied conduction states in the source.

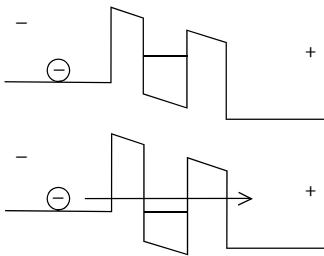


FIGURE 1.34 Band-edge diagram for electron transport in an RTD.

Carriers tunnel through the barrier into the well from the source when the energy of a well state falls within the energy range of the occupied conduction states in the source. With a voltage applied from the source to drain, the charge then flows into the drain.

1.7.2.1 Single-Electron Transistors

A single-electron transistor (SET) has a gate, drain, and source. With appropriate gate voltage, a single electron from the source can tunnel onto a center island (somewhat similar in structure to the RTT). The additional charge changes the energy levels of the island and coulomb repulsion prevents additional electrons from entering the island (for fixed gate voltage). Once the extra electron transfers to the drain, additional charge can enter the island and thereby contribute to the conduction process. The current can be quite large (tens of microamps).

1.7.2.2 Quantum Cellular Automation (QCA)

A basic cell spatially confines two charges. A length of end-to-end cells can be used to transfer a signal. Charges in neighboring cells assume the same configuration as the “input” (or control) cell, which is shown furthest to the left in the figure. If the charges in the control cell are rotated by 90° then the neighboring cell will assume the same state to minimize the coulomb interaction energy (c.f., the Turton book). Subsequent cells will also rotate their state by 90°. Figure 1.35 shows an inverter as an example. The cell furthest to the right assumes the opposite state from the others due to coulomb repulsion from the neighboring cells.

1.7.2.3 Aharonov–Bohm Effect Device

The Aharonov–Bohm effect device (ABED) in Figure 1.36 is based on the fact that the phase of an electron wave function can be influenced by the presence of an electric or magnetic field even when that field is completely isolated from the charge. The field is assumed to be confined to region A in the figure. Even though the field is isolated from the electron-guiding regions, adjusting the vector potential changes the phase of the wave function in the top branch versus that for the bottom branch. It is possible to have either constructive or destructive interference which in-turn controls whether current flows or not.

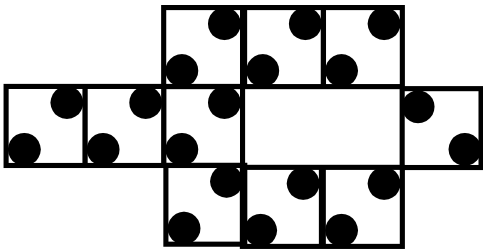


FIGURE 1.35 A QCA inverter. The signal propagates left to right.

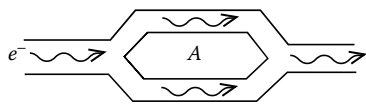


FIGURE 1.36 The Aharanov–Bohm effect device.

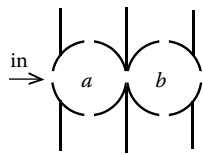


FIGURE 1.37 A representation of the quantum interference device.

1.7.2.4 Quantum Interference Devices

The quantum interference device (QUIT) uses the Aharanov–Bohm (AB) effect but in a slightly different manner than the ABED. Figure 1.37 shows two quantum island regions “a” and “b” defined by the surrounding electrodes. A magnetic field is applied vertically to the plane of the device. The magnetic field, according to classical physics, tends to force the resident electrons into circular orbits, which inhibits the electrons from following a trajectory from region “a” to region “b.” However, the magnetic field can be adjusted and the AB effect can be used to cancel out the circular motions and thereby enhance a trajectory passing between the two regions. These were developed in 1994 as a laboratory device.

1.7.2.5 Josephson Junction

A Josephson junction (JJ) is a superconducting device that uses a weak magnetic field to control the flow of current through a tunneling barrier. A thin-oxide layer that serves as the tunneling barrier separates two slabs of superconducting material. A slight voltage is applied across the two slabs so that current will flow through this junction. A small loop of wire functions as a control (or gate) electrode. A small current passing through the loop will produce a small magnetic field which is sufficient to destroy the pairing of charged particles (i.e., suspends superconduction) near the junction. As a result, the resistance of the junction increases and the devices enters the “off” state. JJ devices use up to a factor of 1000 less power than conventional semiconductor devices.

1.8 REVIEW EXERCISES

- 1.1 Design a simple voltage amplifier using a triode tube consisting of a 6 V filament, a plate, cathode, and a grid. The amplifier should be ac coupled. Assume you have the appropriate power sources, two capacitors and two resistors. Assume the tube has approximately constant transconductance of $g = 10 \text{ mA/V}$.
- 1.2 Design a simple AC voltage amplifier using a transistor, two capacitors, two resistors, and a battery. Explain how the circuit operates.
- 1.3 Draw the band-edge diagram for the case when PNP semiconductors are brought into contact to form a PNP arrangement. Include a diagram of the space charge regions. The diagrams should be similar to that shown in Figure P1.3.

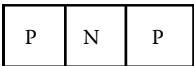


FIGURE P1.3 PNP device.

- 1.4 Semiconductors are placed together in the form: PNP. Assume the contact produces space charge regions in each semiconductor but none of them deplete. Draw a band-edge diagram corresponding to Figure P1.3. Repeat the problem for NPN. What, if anything, can be done to make the device conduct current from one P layer to the other P layer?
- 1.5 Explain how a PIN device would work for a photodetector. Draw a band-edge diagram and explain where light should be absorbed.
- 1.6 Explain how a triode vacuum tube could be used to amplify an input signal.
- 1.7 Perform a literature search for more examples of nanoscale devices. Report your findings.

REFERENCES AND FURTHER READINGS

The following list has references to interesting and informative reading material. Some of these books are considered classics in the field.

Construction and Fabrication

1. Moore J.H., Davis C.C., and Coplan M.A., *Building Scientific Apparatus, A Practical Guide to Design and Construction*, Addison-Wesley Publishing, London, U.K. (1983).
2. Williams R., *Modern GaAs Processing Methods*, Artech House, Boston, MA (1990).
3. Jager R.C., *Introduction to Microelectronic Fabrication*, 2nd ed., *Modular Series on Solid State Devices*, Vol. 5, Prentice Hall, Upper Saddle River, NJ (2002).
4. Campbell S.A., *The Science and Engineering of Microelectronic Fabrication*, 2nd ed., Oxford University Press, New York (2001).
5. Campbell S.A., *Fabrication Engineering at the Micro- and Nanoscale*, 3rd ed., Oxford University Press, New York (2008).

Electronic Circuits and Devices

6. Horowitz P. and Hill W., *The Art of Electronics*, Cambridge University Press, New York (1989).
7. Graf R.F. and Sheets W., *Encyclopedia of Electronic Circuits*, Vol. 7, McGraw-Hill/Tab Electronics, New York (1998).
8. Neudeck G.W., *The Bipolar Junction Transistor*, 2nd ed., *Modular Series on Solid State Devices*, Vol. 3, Addison-Wesley, Reading, MA (1989).
9. Pierret R.F., *Field Effect Devices*, 2nd ed., *Modular Series on Solid State Devices*, Vol. 4, Addison Wesley, Reading, MA (1990).
10. Fraser D.A., *The Physics of Semiconductor Devices*, 3rd ed., Oxford Physics Series, Clarendon Press, Oxford, U.K. (1985).

Amorphous Semiconductors

Average level

11. Brodsky M.H., Ed., *Amorphous Semiconductors*, 2nd ed., *Topics in Applied Physics*, Vol. 36, Springer-Verlag, New York (1985).
12. Redfield D. and Bube R.H., *Photoinduced Defects in Semiconductors*, *Cambridge Studies in Semiconductor Physics and Microelectronics Engineering*, Cambridge University Press, Cambridge, U.K. (1996).

Electrical Contacts

13. Henisch H.K., *Semiconductor Contacts: An Approach to Ideas and Models*, Clarendon Press, Oxford, U.K. (1984).
14. Roderick E.H., *Metal-Semiconductor Contacts*, Clarendon Press, Oxford, U.K. (1980).

NanoDevices

Easy reading level

15. Turton R., *The Quantum Dot—A Journey into the Future of Microelectronics*, Oxford University Press, New York (1996).
16. Milburn G.J., *Schrödinger's Machines*, W.H. Freeman and Company, New York (1997).
17. Crandall B.C., Ed., *Nano Technology—Molecular Speculations on Global Abundance*, MIT Press, Cambridge, MA (2000).
18. Gross M., *Travels to the Nanoworld—Miniature Machinery in Nature and Technology*, Perseus Publishing, Cambridge, MA (1999).
19. Fujimasa I., *Micromachines—A New Era in Mechanical Engineering*, Oxford University Press, Oxford, U.K. (1996).

Average level

20. Davies J.H. and Long A.R., Eds., *Physics of Nanostructures*, Proceedings of the 38th Scottish Universities Summer School In Physics, St. Andrews, July–August 1991, Published by SUSSP Publications (Edinburgh) and IOP Publishing Ltd. (London), 1992. ISBN:0-7503-0169-4 (pbk), 0-7503-0170-8(hbk), published 1992.
21. Timp G.L., Ed., *Nanotechnology*, Springer, New York (1999).
22. Drexler K.E., *Nanosystems—Molecular, Machinery, Manufacturing and Computation*, John Wiley and Sons, Inc., New York (1992).
23. Advanced but accessible after reading this book on *Solid State and Quantum Theory* Mahler G., Weberrub V.A., *Quantum Networks—Dynamics of Open Nanostructures*, Springer-Verlag, Berlin (1995).
24. An advanced book, Ando T., Anakawa Y., Furuya K., Komiyama S., and Nakashima, H., *Mesoscopic Physics and Electronics*, Springer, Berlin (1998).

Bonding and Solid State

Coulsons book is a classic in the field and can often be found for a couple of dollars on the used market. The book requires quantum mechanics.

25. Coulson C.A., *Valence*, Clarendon Press, New York (1956) “Valence” and written by Coulson.
26. Kittel C., *Introduction to Solid State Physics*, 5th edition, John Wiley & Sons, New York (1976 and later).

Misc. Projects

27. *The Amateur Scientist: The Complete Collection*, Version 2.0 (2004). Available from Bright Sciences, LLC, 5600 Post Rd. Suite 114–341, East Greenwich, RI 02818, www.brightscience.com

2 Vector and Hilbert Spaces

Any theory of the physical world employs mathematics as a means for the human mind to understand and reason with physical mechanisms. Mathematics and logic did not grow separately from the world but grew as a result of it. Mathematics might be imagined to have started with the counting whole numbers to represent individual items (such as sheep in a herd). One might imagine the rules of addition were developed to quickly track the total items in two separate groupings (sets) of items. The rules of logic can be linked to the physical world as well. Consider the operation of “and.” For example, an intelligent biped quickly learns to keep the left AND right foot on the ground for a stable stance. There exists evidence that innate mathematical ability is “hard wired” into the brain.

The primitive connection between mathematics and the physical world make use of everyday experience at the macroscale. With quantum phenomena, one encounters a regime far removed from ordinary experience. One relies on laboratory experiments to define the quantum world; the experimental process refines ordinary everyday experiments such as for counting and logic. One cannot expect to find results that appeal to “common sense.” Here, one must make a choice as to whether the new situation can be described by existing mathematical infrastructure or it requires the invention of new infrastructure. In either case, one must set up a mathematical system, which exists only in the human mind, in order to describe the system of nature. That is, “mathematics belongs to and exists in the human mind whereas nature belongs to God.” Mathematics and nature have separate but intertwined existences. As will become evident later, one must allow for the possibility that human observation (a physical process, not a mathematical one) can affect the state of the natural system.

Linear algebra is the natural mathematical language of quantum mechanics. The quantum mechanical wave functions live in Hilbert space essentially defined as a vector space with an inner product. For this reason, the chapter starts with finite and infinite dimensional vector and Hilbert spaces and uses the Dirac notation to help unify the concepts for these spaces. The chapter provides intuitive pictures to show how a function can be imagined as a vector in the Hilbert space. It discusses inner products, norms, closure and completeness, and dual spaces. Fourier, Cosine, and Sine series appear as examples of expansions in complete orthonormal sets of functions. The Minkowski space provides an example of a pseudo-inner product, and provides examples on the use of the metric matrix, and the tensor notation found in physical applications.

2.1 VECTOR AND HILBERT SPACES

Linear algebra comprises the natural language of quantum theory. The present section provides direction and motivation for the upcoming topical areas by first introducing the role of operators and vectors in the quantum theory and then defining vector and Hilbert spaces. The full view of the Hilbert space and its connection with the quantum theory will unfold over the next several chapters.

2.1.1 MOTIVATION FOR LINEAR ALGEBRA IN QUANTUM THEORY

In the case of quantum theory, it has become evident that the physical world can be represented by the ideas of linear algebra (more precisely, the Hilbert space). The abstract linear algebra focuses on operators and vectors but not necessarily on the matrix form. Thinking back to elementary studies of mechanics and electromagnetics, it becomes apparent that one can apply vectors and

operators (such as divergence) and expect to obtain reasonable description of some phenomena. However, the quantum theory uses the linear algebra more for the basis vectors, linear combinations, and eigenvalue equations.

The quantum theory represents physically “observable” quantities by (Hermitian) operators and the specific properties of a particle by a vector (wave function) in a vector space. In one sense, the mathematical object “operator” in the quantum theory represents the “act of observing.” The average of the operator represents the classically expected results of a measurement of the corresponding physical observable. For example, the momentum, energy, and position of a particle can all be observed and would be represented by operators. This assumption begins to set up the correspondence between the physical world and the mathematics.

Upon what does the operator act? We assume that abstract vectors (wave functions) describe all of the inherent properties and characteristics of particles. The theory must provide a method to calculate the results of measuring the value of an observable quantity. The act of observing the properties of a particle can be related to operating on the vector (wave function) with the corresponding operator. The vectors (wave functions) live in a Hilbert space, which consists of a vector space with an inner product. One can imagine the vector space as the ordinary Euclidean space one learns in high school. The vector space has basis vectors whereby every vector in the space can be written as a summation over the basis vectors (i.e., a linear combination of basis vectors). Here the basis vectors can be as simple as the unit vectors normally used to define a coordinate system. For certain conditions, the result of an operator acting on a basis vector directly produces the results of a measurement. For example, suppose $\hat{O}$ represents an operator corresponding to an observable such as, for example, energy or momentum. Suppose further that “ ϕ ” represents a basis vector (i.e., a fundamental wave or state function) giving, for example, the precise energy or momentum of the particle in question. An observation of observable O when the particle resides in state ϕ gives the following result.

$$\hat{O}\phi = o\phi \quad (2.1)$$

The real constant o represents the result of the observation. If for example, “ $\hat{O}$ ” represents the momentum operator, then o represents the value of the particle momentum when the particle occupies momentum state ϕ . Equation 2.1 shows the basis vector must also be an eigenfunction or eigenvector of the operator $\hat{O}$. The eigenfunction equation $\hat{O}\phi = o\phi$ produces the eigenvalue o . The result of every observation must always be an eigenvalue. We can always write an eigenfunction equation for every observable.

One might now entertain the following question. Suppose the operator $\hat{O}$ represents the act of measuring energy and o represent the possible observed values of energy. If the basis vectors correspond to unit vectors (defining the coordinate axes) in a finite dimensional space (finite number of basis vectors) so that the number of o values must be finite, then does this require the possible number of different energy values to be finite and distinct. The answer is yes. Notice that this does not correspond to our ordinary view of the world where energy can take on a continuous range of values (such as the kinetic energy of an automobile).

Equation 2.1 begins to show the special role played by the basis vectors. The general wave function Ψ (i.e., vector) can be written as a sum of basis functions (i.e., basis vectors) such as

$$\Psi = \sum_n A_n \phi_n \quad (2.2)$$

Equation 2.2 says that the properties of the particle modeled by the vector Ψ (i.e., wave function) become a mixture of those represented by the basis vectors ϕ_n . For example, the energy of a particle can be a combination of the various discrete energies as we will see later. Based on Equation 2.2, a measurement (i.e., observation) can produce only one value from the discrete set of values and not a continuous range. Further, we will see that the coefficient A_n of each individual term must be related

to the probability of finding the particle in the particular eignestate ϕ_n . Generally, the classically expected result of a measurement comes from calculating the expectation value of the operator using the general vector in Equation 2.2.

The quantum theory makes extensive use of a variety of operators besides the Hermitian operators that represent observables. For example, unitary operators rotate vectors in Hilbert space. They can map one basis set into another to make an operator (matrix) diagonal. Actually, the set of linear operators itself forms a vector space. A linear operator can be expanded in an operator basis set.

It should be clear at this point that we have a lot of learning and explaining to do in order to apply the linear algebra.

2.1.2 DEFINITION OF VECTOR SPACE

A vector space consists of a set F with a defined binary operation “+” and a scalar multiplication (SM) over the field of numbers $\mathcal{N}$ such that (assuming f, f_1, f_2 are in F and α, β are in $\mathcal{N}$),

Closure:	$f_1 + f_2$ is in F and αf is in F
Associative:	$(f_1 + f_2) + f_3 = f_1 + (f_2 + f_3)$
Commutative:	$f_1 + f_2 = f_2 + f_1$
Zero:	There exists a zero vector O such that $O + f = f$
Negatives:	For every f in F , there exists $(-f)$ in F such that $f + (-f) = O$
SM associative:	$(\alpha\beta)f = \alpha(\beta f)$
SM distributive:	$\alpha(f_1 + f_2) = \alpha f_1 + \alpha f_2$
SM distributive:	$(\alpha + \beta)f = \alpha f + \beta f$
SM unit:	$1f = f$

If “ F ” is a set of functions then F is sometimes called a function space. The vector space refers to both the set of objects V as well as the binary operation and will sometimes be denoted by $(V, +)$. Most often, the “+” operation is assumed and then the vector space is denoted by V .

We will generally refer to collections of vectors that represent direction and displacement as Euclidean vector spaces. We assume these spaces have a finite number of dimensions defined by unit vectors (i.e., basis vectors) such as those representing the x , y , and z axes as in $\{\hat{x}, \hat{y}, \hat{z}\}$.

Example 2.1

Let $\{\hat{x}, \hat{y}\}$ be the basis vectors (i.e., the “unit vectors”) typically defined for geometrical vectors in a Euclidean vector space. Let $\mathcal{N}$ be the set of real numbers $\mathcal{R}$. The set $V = \{\vec{v} = x\hat{x} + y\hat{y}: x, y \in \mathcal{R}\}$ forms a vector space. The set V represents the set of two-dimensional (2-D) vectors one normally draws on a sheet of paper.

Example 2.2

The set $\{f: \mathcal{R} \rightarrow \mathcal{R} \text{ such that } f \text{ is continuous and } \mathcal{R} = \text{reals}\}$ is the vector space of real, continuous functions.

Example 2.3

For complex functions F , the number field $\mathcal{N}$ must be the set of complex numbers $\mathcal{C}$ while, for real functions F , the number field $\mathcal{N}$ consists of the real numbers $\mathcal{R}$. For example, if F represents the set of real functions but the number field consists of complex numbers, then objects such as $c_1 f(x)$ (where c_1 is complex) cannot be in the original vector space because the function $g(x) = c_1 f(x)$ can have complex values.

Therefore, for this example, closure cannot be satisfied contrary to the requirements of the definition for the vector space.

2.1.3 HILBERT SPACE

We will refer to a Hilbert Space H as a vector space with an inner product defined on the space. The inner product between two elements f_1 and f_2 in H will be denoted by $\langle f_1 | f_2 \rangle$. Some books reserve the term “Hilbert space” for vector spaces of functions with an inner product; they sometimes denote the inner product by (f_1, f_2) . For function spaces, the functions must be square integrable in the sense that the following integral must exist for $f \in H$

$$\int_a^b dx |f(x)|^2$$

We will see below that the existence of the integral with a finite value is equivalent to requiring that the vector (i.e., function) have finite length. Sometimes the term “inner-product space” refers to a vector space (regardless of whether it is a Euclidean or function space) having a defined inner product. This book does not make any distinction between the function or Euclidean vector spaces and assumes all of the inner products exist (such as the previous integral).

Example 2.4

Consider $\mathcal{R}^2$ which is the set of (Euclidean) vectors in the x - y plane. Assume $\vec{r}_1$ and $\vec{r}_2$ are two vectors in $\mathcal{R}^2$ with $\vec{r}_1 = x_1\hat{x} + y_1\hat{y}$ and $\vec{r}_2 = x_2\hat{x} + y_2\hat{y}$. Simple vector analysis provides the relations

Inner product	$\langle \vec{r}_1 \vec{r}_2 \rangle = \vec{r}_1 \cdot \vec{r}_2 = x_1x_2 + y_1y_2$
Norm	$\ \vec{r}_1\ = \langle \vec{r}_1 \vec{r}_1 \rangle^{1/2} = (x_1^2 + y_1^2)^{1/2}$
Metric	$d(\vec{r}_1, \vec{r}_2) = \ \vec{r}_1 - \vec{r}_2\ = [(\vec{r}_1 - \vec{r}_2) \cdot (\vec{r}_1 - \vec{r}_2)]^{1/2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

The norm gives the length of a vector while the metric gives the distance between two vectors.

As shown in the next section, we do not normally include the “arrow” above the vector when using the bracket notation $\langle \bullet | \bullet \rangle$ (i.e., including the “arrow” above the vector abuses of the notation—but people abuse this notation all the time anyway). The typical scalar product for vectors appears in the above table. The norm provides the length of a vector. The metric provides the distance between two elements of the vector space. Notice that the definitions interrelate the inner product, norm, and metric. For example, notice that $\|\vec{r}_1\| = d(\vec{r}_1, 0)$. The following sections show how the formulas can be modified for complex vector spaces.

An inner product $\langle \bullet | \bullet \rangle$ in a (real or complex) vector space F is a scalar valued function that maps $F \times F \rightarrow C$ (where C is the set of complex numbers) with the properties

1. $\langle f | g \rangle = \langle g | f \rangle^*$ where f and g are elements in F and asterisk $*$ denotes complex conjugation.
2. $\langle \alpha f + \beta g | h \rangle = \alpha^* \langle f | h \rangle + \beta^* \langle g | h \rangle$ and $\langle h | \alpha f + \beta g \rangle = \alpha \langle h | f \rangle + \beta \langle h | g \rangle$, where f, g , and h are element of F and α, β are elements in the complex number field C .
3. $\langle f | f \rangle \geq 0$ for all vectors f . The inner product can be zero $\langle f | f \rangle = 0$ if and only if $f = 0$ (except at possibly a few isolated points for functions).

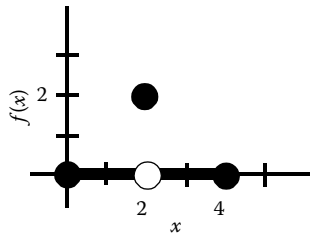


FIGURE 2.1 The inner product $\langle f|f \rangle = 0$ requires $f=0$ except at a few isolated points.

As a side note, consider the meaning of the symbol $F \times F \rightarrow C$. In the general setting, the notation $A \times B$ refers to the set comprised of “ordered pairs” with the first element from set A and the second element from set B ; that is, $A \times B = \{(a, b) \text{ such that } a \in A, b \in B\}$. In the present context of $F \times F$, one takes an element of F , say g , and places it in the first position of the inner product $\langle g|\bullet \rangle$, and then one takes a second element of F , say h , for the second position to form $\langle g|h \rangle$. The result represented by $\langle g|h \rangle$ (i.e., the mapping) is a complex number.

We have already stated the inner product for Euclidean vector spaces. The inner product (related to the norm and metric) can be defined for functions as

Inner product	$\langle f g \rangle = \int_a^b dx f(x)^* g(x)$
Norm	$\ f(x)\ = \langle f f \rangle^{1/2} = \int_a^b dx f(x)^* f(x) = \int_a^b dx f(x) ^2$

Several notes are in order. First, the inner product for a function space uses the Riemann integral over the domain of definition. Second, Property 3 in the definition of inner product holds exactly for continuous functions on $[a, b]$; however for piecewise continuous functions, the property holds except at possibly the points of discontinuity as shown in Figure 2.1. The exceptions arise for piecewise continuous functions because of the insensitivity of the Riemann integral to individual points. Third, real Euclidean vectors and functions do not need the complex conjugate.

The properties of the inner product are very similar to those of the metric. In fact, the metric $d(f, g)$ can be written as

$$d(f, g) = \langle f - g|f - g \rangle^{1/2}$$

Example 2.5

Find the length of $f(x) = x$ for $x \in [-1, 1]$

$$\|f\| = \langle f|f \rangle^{1/2} = \left[\int_{-1}^1 dx x^* x \right]^{1/2} = \left[\int_{-1}^1 dx x^2 \right]^{1/2} = \sqrt{\frac{2}{3}}$$

where we used the fact that $f(x) = x$ is real. If we were to divide the function by the norm and write $g(x) = f(x)/\|f\|$ then the length of $g(x)$ would be unity. In general, we normalize a function $f(x)$ to one by dividing by the norm of $f(x)$.

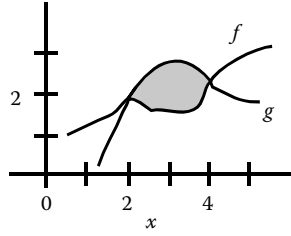


FIGURE 2.2 The area between the functions f and g on the interval $[2, 4]$ gives an interpretation of the “distance” between them.

A metric $d(f, g)$ is a relation between two elements f and g of a set F such that

1. $d(f, g) \geq 0$ and $d = 0$ only when $f = g$ (except at possibly a few points for $C_p[a, b]$, the set of piecewise continuous functions on $[a, b]$. Recall that two functions are equal only when $f(x) = g(x)$ for all “ x ” in the domain of definition)
2. $d(f, g) = d(g, f)$
3. $d(f, g) < d(f, h) + d(h, g)$ where h is any third element of F

The metric measures the distance between two elements of the space. Although there are better methods of “picturing” the “distance” between two functions, Figure 2.2 gives an “area” interpretation. The distance between f and g can be written as

$$d(f, g) = \langle f - g | f - g \rangle^{1/2} = \left[\int_2^4 dx (f - g)^2 \right]^{1/2}$$

where f and g are assumed real for the present illustration. Then $f - g$ calculates the difference, while the square ensures a positive result and the integral calculates an area.

2.1.4 COMMENT ON THE LENGTH OF A VECTOR FOR QUANTUM THEORY

For quantum theory, the vectors and functions representing the properties of physical particles have the general designation of state vectors or “wave functions” and they live in a Hilbert space. These wave functions being vectors therefore have length and direction. Obviously, not all vectors/functions have unit length; however, those corresponding to properties of physical objects and systems must have unit length (i.e., normalized to unity). The wave functions $\Psi(\vec{r}, t)$, also known as probability amplitudes, represent the probability of finding a particle in the vicinity of point $\vec{r}$ at time t . As we will see, an integral over the square of the probability amplitude represents the length (squared) of the vector

$$\|\Psi\|^2 = \int_{\text{all space}} dV \Psi^*(\vec{r}, t) \Psi(\vec{r}, t) = 1$$

This integral also states that the probability of finding the particle somewhere in space must be unity. The integrand comprises the probability density, namely $\Psi^* \Psi$. The allowed wave functions have unit length in the Hilbert space. These wave functions, being vectors, must be related to certain linear combinations of the basis vectors (which also have unit length) that produce a type of unit sphere. Elementary vector analysis suggests that these wave vectors must form an $(n - 1)$ -dimensional surface in an n -dimensional space (e.g., the surface of a sphere for

three-dimensional [3-D] space). We will see that these wave functions (once they have been normalized) do not themselves form a vector space, but reside on a surface within the space. The surface does not contain the zero vector and cannot therefore be a vector space.

2.1.5 LINEAR ISOMORPHISM

An isomorphism is a special function “ $f: V \rightarrow W$ ” that maps one set V into another W . For vector spaces V and W , this function is usually termed an operator (see Chapter 3). The function must be linear in the sense that if v_1 and v_2 are elements of V (i.e., $v_1 \in V$ and $v_2 \in V$), then $f(av_1 + bv_2) = af(v_1) + bf(v_2)$ where a and b are complex numbers. Further, the isomorphism requires the function to be (1) “onto” in the sense that every element $w \in W$ has a preimage $v \in V$ (i.e., $f(v) = w$) and (2) “1–1” in the sense that an element $w \in W$ has only one preimage in V . The reader will also recognize that such functions have inverses and these inverse functions will also be isomorphisms.

An isomorphism “ $f: V \rightarrow W$ ” between two sets V and W with binary operations “+” and “&” ensures that $(W, \&)$ is a vector space when $(V, +)$ is one. In this case, the property of linearity has the form $f(av_1 + bv_2) = af(v_1) \& bf(v_2)$ since the image of objects in V must be objects in W which can only be combined using the “&” operation. We can see $(W, \&)$ must be a vector space quite easily by showing the elements of W obey the properties in Section 2.1.2. For example, consider the associative property. Let w_1, w_2, w_3 be elements of W . Then there must exist elements v_1, v_2, v_3 in V such that $w_1 = f(v_1)$ and so on. Then the following list demonstrates the property.

Step	Reason
$[w_1 \& w_2] \& w_3 = [f(v_1) \& f(v_2)] \& f(v_3)$	“ f ” is onto
$[w_1 \& w_2] \& w_3 = f(v_1 + v_2) \& f(v_3)$	“ f ” is linear
$[w_1 \& w_2] \& w_3 = f[(v_1 + v_2) + v_3]$	“ f ” is linear
$[w_1 \& w_2] \& w_3 = f[v_1 + (v_2 + v_3)]$	Associative law holds in V
$[w_1 \& w_2] \& w_3 = f(v_1) \& [f(v_2) \& f(v_3)]$	“ f ” is linear
$[w_1 \& w_2] \& w_3 = w_1 \& [w_2 \& w_3]$	

The other properties can be proven similarly.

2.1.6 ANTILINEAR ISOMORPHISM

The antilinear isomorphism has a very important role with the inner product and the adjoint operator. It has all of the same properties as the linear isomorphism except a modified linearity property. If a and b are complex numbers and v_1 and v_2 are vectors in V then the antilinear function has the following property

$$f(av_1 + bv_2) = a^*f(v_1) + b^*f(v_2)$$

where “ $*$ ” represents the complex conjugate and the function f has the mapping $f: V \rightarrow W$ and it is assumed that the binary operation “+” applies to both vector spaces. If an antilinear isomorphism exist between two sets with one being a vector space, then it can be shown that the other satisfies vector space properties. The role of the antilinear isomorphism will become clear in relation to the dual space and the Dirac notation.

2.2 DIRAC NOTATION AND EUCLIDEAN VECTOR SPACES

The present section introduces a notation created by P.A.M. Dirac during the early twentieth century. Professor Dirac, a mathematician and physicist, was intimately familiar with linear algebra

and quantum theory. As one of many major accomplishments, he developed a partial differential equation that combined Einstein's special theory of relativity with quantum mechanics; this equation contained the first suggestion of the existence of antimatter. For our purposes, the Dirac notation helps to unify Euclidean and function spaces and those with discrete and continuous sets of basis functions. The notation is first applied to the vector space spanned by the basis set of unit vectors $\{\tilde{x}, \tilde{y}, \tilde{z}\}$. We then discuss the concepts of closure and completeness.

2.2.1 KETS, BRAS, AND BRACKETS FOR EUCLIDEAN SPACE

The basis vectors for 3-D Euclidean space $\{\tilde{x}, \tilde{y}, \tilde{z}\}$ can also be written as $\{\tilde{e}_1, \tilde{e}_2, \tilde{e}_3\}$. The basis set consists of the typical unit vectors that define the x -, y -, and z -axes. The “ $\tilde{e}$ ” unit vector notation can be simplified by omitting the redundant e and writing $\{\tilde{1}, \tilde{2}, \tilde{3}\}$. Dirac provides an alternate notation. Vectors $\vec{v}$ can also be written in “ket” $| \rangle$ notation as $|v\rangle$. The basis vectors become $\tilde{e}_n \leftrightarrow |n\rangle$ so that

$$\tilde{x} \leftrightarrow |1\rangle \quad \tilde{y} \leftrightarrow |2\rangle \quad \tilde{z} \leftrightarrow |3\rangle$$

For example, the vector $\vec{v} = 3\tilde{x} - 4\tilde{y} + 10\tilde{z}$ can be written as $|v\rangle = 3|1\rangle - 4|2\rangle + 10|3\rangle$. Sometimes the vector sum and scalar product are written as $|v_1\rangle + |v_2\rangle \equiv |v_1 + v_2\rangle$ and $|\alpha v\rangle = \alpha|v\rangle$, respectively.

We define a “bra” $\langle |$ to be a projection operator. The bras $\langle 1|, \langle 2|, \langle 3|$ represent operators that project a vector $\vec{v}$ onto the unit vectors $\tilde{x}, \tilde{y}, \tilde{z}$, respectively. For example, if $|v\rangle = 3|1\rangle + 4|2\rangle$ then the projection operators provide the components $\langle 1|\vec{v} = 3, \langle 2|\vec{v} = 4$ as shown in Figure 2.3. Here the bra $\langle 1|$, for example, operates on $\vec{v}$ to give the component of $\vec{v}$ along the $\tilde{x}$ -axis. We would do better to write the combination of projection operators and vectors as

$$\langle 1|\vec{v} = \langle 1|v\rangle$$

This combination of the “bra” + “ket” gives the “braket” (or bracket).

In general, $\langle w|$ represents the operator that projects an arbitrary vector onto the vector $\vec{w}$. The linear operator $\langle w|$ corresponds to “ $\vec{w} \bullet$ ” where the dot refers to the usual dot product

$$\langle w|v\rangle = (\vec{w} \bullet) \vec{v} = \vec{w} \bullet \vec{v}$$

We see that the bracket must be an inner product (the same inner product defined earlier). If “ n ” represents an integer corresponding to one of the basis vectors then $\langle n|v\rangle$ represents a component of the vector. The bras are linear operators and can be distributed across a sum.

$$\langle w| [|v_1\rangle + |v_2\rangle] = \langle w|v_1\rangle + \langle w|v_2\rangle$$

As a note, some books call the bras “projectors” and they call the more complicated objects $|\bullet\rangle\langle\bullet|$ as projection operators.

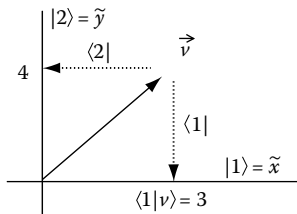


FIGURE 2.3 Dotted arrow shows action of the projection operator.

We typically think of vectors as objects that can be pictured similar to the example for $|v\rangle$ in Figure 2.3. However, we often think of projection operators as an “action” rather than an object. True, the projection operators form a vector space and so each operator is itself a vector, but they also perform a mapping. Figure 2.3 shows the act of projecting the vector $|v\rangle$ onto the basis vectors $|1\rangle$ and $|2\rangle$.

Example 2.6

Find the projection of $\vec{v} = 3\tilde{x} + 5\tilde{y} + 10\tilde{z}$ onto the three axes.

$$\langle 1|\vec{v}\rangle = (\tilde{x}\bullet)\vec{v} = \tilde{x}\bullet(3\tilde{x} + 5\tilde{y} + 10\tilde{z}) = 3$$

Similarly, the two other projections must be $\langle 2|v\rangle = \tilde{y}\bullet\vec{v} = 5$ and $\langle 3|v\rangle = 10$.

Example 2.7

The vector $\vec{v} = v_1\tilde{x} + v_2\tilde{y} + v_3\tilde{z}$ can be projected onto the x -axis by distributing “ $\tilde{x}\bullet$ ” across all of the terms

$$\tilde{x}\bullet\vec{v} = \tilde{x}\bullet(v_1\tilde{x} + v_2\tilde{y} + v_3\tilde{z}) = v_1\tilde{x}\bullet\tilde{x} + v_2\tilde{x}\bullet\tilde{y} + v_3\tilde{x}\bullet\tilde{z}$$

This can be equivalently written as

$$\langle 1|v\rangle = \langle 1|\{v_1|1\rangle + v_2|2\rangle + v_3|3\rangle\} = v_1\langle 1|1\rangle + v_2\langle 1|2\rangle + v_3\langle 1|3\rangle$$

The bras as linear operators distribute across the sum. Terms such as $\langle 1|2\rangle$ yield the results $\langle 1|2\rangle = \tilde{x}\bullet\tilde{y} = 0$ and so on.

2.2.2 BASIS AND COMPLETENESS FOR EUCLIDEAN SPACE

A basis set must be orthonormal and complete. Two vectors $|m\rangle, |n\rangle$ are orthonormal when

$$\langle m|n\rangle = \delta_{m,n} = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases} \quad (2.3)$$

The Kronecker delta function $\delta_{m,n}$ expresses orthonormality for countable (i.e., discrete) basis sets. Countable sets have elements in one-to-one correspondence with a subset of integers. Sets with an infinite number of elements can also be countable so long as each element of the set corresponds to an integer. As an example for Equation 2.3, the condition $m = n$ provides $\langle m|m\rangle = 1$ which means that the vector $|m\rangle$ is normalized to have unit length (this is similar to saying $\tilde{x}\bullet\tilde{x} = 1$ giving $\|\tilde{x}\| = 1$ for example). The case $m \neq n$ gives $\langle m|n\rangle = 0$ which indicates the unit vectors $|m\rangle, |n\rangle$ must be orthogonal. This is similar to saying for example that $\langle 1|2\rangle$ is the projection of $\tilde{y}$ onto $\tilde{x}$ which produces $\tilde{x}\bullet\tilde{y} = 0$. A set of vectors $B = \{|v_1\rangle, |v_2\rangle, \dots\}$ is orthonormal if for any two vectors $|v_m\rangle, |v_n\rangle$ in B , the inner product between them satisfies $\langle v_m|v_n\rangle = \delta_{m,n}$. For cases where a set of basis functions has a one-to-one relation with a continuous subset of the real numbers (i.e., continuous basis set), the Dirac delta function (i.e., the impulse function) $\delta(x - x')$ replaces the Kronecker delta function $\delta_{m,n}$. The “continuous” basis set has an uncountable number of elements.

A linear combination of N orthonormal vectors $B = \{|1\rangle, |2\rangle, \dots |N\rangle\}$ has the form

$$|v\rangle = \sum_{i=1}^N C_i|i\rangle \quad (2.4)$$

where $\{C_i\}$ can be complex numbers. The collection of all such vectors $V = \{|v\rangle\}$ forms a vector space and the set B is a basis set. The set B spans the vector space $V = Sp(B)$, which has dimension $\text{Dim}(V) = N$.

Notice that we started with the set B and generated a vector space V for Equation 2.4. Consider the reverse situation. Consider an arbitrary vector space W . A set of vectors B is complete in W if every vector in the space W can be written as a linear combination of the vectors in B . A complete set of orthonormal vectors forms a basis set (or sometimes to be redundant: “a complete orthonormal basis set”). As a note, the set of all possible linear combinations of the form (Equation 2.4) is complete by construction. This means that every vector in V can be found by a suitable choice of the C_i .

Example 2.8

The set $B_2 = \{|1\rangle, |2\rangle\}$ is not complete for the 3-D space spanned by $B_3 = \{|1\rangle, |2\rangle, |3\rangle\}$. For example, any vector of the form $\vec{v} = C_1 \hat{z}$ cannot be written in terms of the vectors in B_2 .

Example 2.9

The set $B'_3 = \{\tilde{x}, \tilde{y}, \tilde{z}\}$ is a basis set for R^3 (i.e., the same vector space as spanned by B_3 in Example 2.8) while $B_4 = \{\tilde{x}, \tilde{y}, \tilde{z}, \tilde{t}\}$ is a basis set for R^4 (i.e., a four-dimensional [4-D] vector space). B_3 spans a subspace of R^4 , namely R^3 . B_3 is complete for R^3 but not for R^4 .

2.2.3 CLOSURE RELATION FOR THE EUCLIDEAN VECTOR SPACE

We now turn attention to the closure relation also known as the completeness relation or resolution of unity. This relation is equivalent to the completeness property of the basis set.

We now show the form of the closure relation starting with the completeness property of a basis set. The components of the vector, namely C_i in Equation 2.4, can be written in terms of “brackets” by projecting the vector $|v\rangle$ onto each basis vector $|m\rangle$.

$$\langle m|v\rangle = \langle m|\sum_{i=1}^n C_i|i\rangle = \sum_{i=1}^n C_i\langle m|i\rangle = \sum_{i=1}^n C_i\delta_{i,m} = C_m \quad (2.5)$$

The results from Equation 2.5, equivalently written as $C_i = \langle i|v\rangle$, can be substituted into Equation 2.4 to obtain $|v\rangle = \sum_{i=1}^n C_i|i\rangle = \sum_{i=1}^n [\langle i|v\rangle]|i\rangle$ or, by switching the number $C_i = \langle i|v\rangle$ with the vector $|i\rangle$

$$|v\rangle = \sum_{i=1}^n |i\rangle\langle i|v\rangle \quad (2.6)$$

Equation 2.6 can be regrouped

$$|v\rangle = \left(\sum_{i=1}^n |i\rangle\langle i|\right)|v\rangle \quad (2.7)$$

In the last equation, consider the quantity in parenthesis to be an operator and realize that the equation must hold for all vectors $|v\rangle$ in the vector space V . Two operators are equal $\hat{A} = \hat{B}$ provided $\hat{A}|v\rangle = \hat{B}|v\rangle$ for every $|v\rangle \in V$. Consequently, Equation 2.7 shows that

$$\sum_{i=1}^n |i\rangle\langle i| = \hat{1} \quad (2.8)$$

for the vector space V spanned by the basis $B = \{|1\rangle, |2\rangle, \dots, |n\rangle\}$. The “ $\hat{1}$ ” that appears in Equation 2.8 represents an operator with the property $\hat{1}|w\rangle = |w\rangle$ for any vector $|w\rangle$ in V . For practical purposes at this time, treat the “ $\hat{1}$ ” as just the real number “1.” One should note that the exact form for the closure relation depends on the vector space. In particular, the number of terms in Equation 2.8 depends on the dimension of the space.

Example 2.10

The completeness relation for $\mathcal{R}^3$ using $\langle w| = \vec{w} \bullet$ is

$$1 = |1\rangle\langle 1| + |2\rangle\langle 2| + |3\rangle\langle 3| \quad \text{so} \quad 1 = \tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet.$$

Note that the unit vectors are written next to each other without an operator between them.

Example 2.11

To see that “ $\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet$ ” represents a unit operator, let $\vec{R} = x\tilde{x} + y\tilde{y} + z\tilde{z}$ and write

$$\begin{aligned} (\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet) \vec{R} &= (\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet)(x\tilde{x} + y\tilde{y} + z\tilde{z}) \\ &= (\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet)x\tilde{x} + (\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet)y\tilde{y} + (\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet)z\tilde{z} \\ &= x\tilde{x} + y\tilde{y} + z\tilde{z} \end{aligned}$$

So we get the vector $\vec{R}$ back again. The operator “ $(\tilde{x}\tilde{x} \bullet + \tilde{y}\tilde{y} \bullet + \tilde{z}\tilde{z} \bullet)$ ” just decomposes $\vec{R}$ into components and then reassembles it again.

First let’s prove that $\{|1\rangle, |2\rangle, \dots\}$ is complete if and only if $\sum_{m=1}^n |m\rangle\langle m| = 1$. First, show completeness leads to closure. The previous discussion shows that a complete orthonormal set leads to the closure relation since every $|v\rangle$ can be written as $|v\rangle = \sum_{m=1}^n c_m |m\rangle$ with expansion coefficients $c_m = \langle m|v\rangle$. Substituting for c_m leads to Equation 2.8 as required.

Second, show that the closure relation establishes the completeness of the orthonormal set. Assume to the contrary that the set is not complete. There must exist a vector $|v'\rangle \in V$ such that $|v'\rangle \neq \sum_m c_m |m\rangle$ for all constants c_m . Form the summation $\sum_m |m\rangle\langle m|v'\rangle$. It must be true that $\sum_m |m\rangle\langle m|v'\rangle \neq |v'\rangle$ because otherwise, there would be a set of constants c_m , namely $c_m = \langle m|v'\rangle$ such that $|v'\rangle = \sum_m c_m |m\rangle$ contrary to our starting assumption. Therefore there exists a vector $|v'\rangle$ such that $(\sum_m |m\rangle\langle m|)|v'\rangle \neq |v'\rangle$. Therefore $\sum_m |m\rangle\langle m| \neq 1$ which is contrary to the assumption of the theorem.

2.2.4 EUCLIDEAN DUAL VECTOR SPACE

The previous section shows that a bra $\langle w|$ projects an arbitrary vector onto the vector $\vec{w}$. The linear operator $\langle w|$ maps a vector space V into the complex numbers $\mathcal{C}$ (i.e., $\langle w|: V \rightarrow \mathcal{C}$). These projection operators form a vector space, denoted by V^+ or V^* , and termed the dual space (to the vector space V). The dual space consists of elements that are both vectors, since V^+ is a vector space, and linear operators, since $\langle \bullet|$ represents a projection operator. This means that the set of bras satisfies all of the properties listed in Section 2.1 for vector spaces. Chapter 3 will show that the dual space also has an inner product.

Example 2.12

Commutative law and additive inverse. The commutative law can be written as $\langle v_1| + \langle v_2| = \langle v_2| + \langle v_1|$ and there exists an additive inverse $-\langle v|$ such that $-\langle v| + \langle v| = \langle 0|$ where $\langle 0|$ is the

zero vector. Sometimes the notation for the zero ket (or bra) is confused with that for a vector like $|0\rangle$ where “0” is a label for a particular vector in a basis set; i.e., instead of writing a basis as $\{|1\rangle, |2\rangle, \dots\}$, it might be written as $\{|0\rangle, |1\rangle, \dots\}$ just by changing the index. One way to avoid this problem is just to write the zero vector as $|O\rangle$.

Example 2.13

Show the commutative law $\langle v_1| + \langle v_2| = \langle v_2| + \langle v_1|$ holds.

SOLUTION

Let $|\xi\rangle$ be an arbitrary vector in the vector space V . Then by definition of addition of projectors, we have

$$[(\langle v_1| + \langle v_2|)|\xi\rangle] = \langle v_1|\xi\rangle + \langle v_2|\xi\rangle = \langle v_2|\xi\rangle + \langle v_1|\xi\rangle$$

where the last step follows from the fact that complex numbers commute. Now use the definition of addition again to write

$$[(\langle v_1| + \langle v_2|)|\xi\rangle] = [(\langle v_2| + \langle v_1|)|\xi\rangle]$$

This last expression holds for every vector $|\xi\rangle$ in the space so that the definition of equal operators produces the desired results $\langle v_1| + \langle v_2| = \langle v_2| + \langle v_1|$.

The set V^+ consisting of all bra operators $\langle w|$ defines the “dual” of the vector space V . One of the most convenient methods for proving that V^* forms a vector space is to show there exists an isomorphic map between V and V^* . Such a map insures that the set V^* inherits the properties of V . We can start to see how such a map arises by noting for each ket $|w\rangle$, there exists a bra $\langle w|$ and vice versa. The original vector space V has a 1–1 relation with the “dual vector space V^+ .” Recall that a 1–1 relation (i.e., correspondence) between two sets A and B means that for each and every element of A , there corresponds exactly one element of B and vice versa. This 1–1 relation is a function. For the relation between kets and bras, it is a function that simply reverses the direction of the ket symbol; however, one should keep in mind that the bra has an operator character in addition to its vector nature. We will next describe this mapping (i.e., the 1–1 function) as the adjoint operator.

Mathematically, the two vector spaces $V = \{|v\rangle\}$ and $V^+ = \{\langle w|\}$ are related by an antilinear 1–1 (isomorphic) map denoted by the dagger superscript. The isomorphic map $+: V \leftrightarrow V^+$ is called the Hermitian conjugate (or adjoint operator) and performs the following operation

$$\langle \bullet | \quad \xleftrightarrow{+} \quad |\bullet\rangle \quad \text{or} \quad |w\rangle^+ = \langle w|$$

The linear property means the adjoint distributes across summations as

$$[\alpha|v\rangle + \beta|w\rangle]^+ = [\alpha|v\rangle]^+ + [\beta|w\rangle]^+$$

where α, β represent complex numbers. The “anti” part of the adjoint refers to how it handles the complex numbers. In particular the definition of adjoint includes the following relation

$$[\alpha|v\rangle]^+ = |v\rangle^+ \alpha^+ = |v\rangle^+ \alpha^* = \alpha^* |v\rangle^+ = \alpha^* \langle v|$$

The “anti” part of the definition changed the complex number into its conjugate. Also notice how the adjoint changed the order of the operands. The order change was not necessary for the complex

$$\begin{array}{c} \langle w | L_1 L_2 | v \rangle^* \\ \swarrow \quad \searrow \\ \langle v | L_1^+ L_2^+ | w \rangle \end{array}$$

FIGURE 2.4 Action of the adjoint operator.

number but it will be important for the adjoint of two operators. We will prove the relations given here using an alternative but equivalent definition for the adjoint in Chapter 3. In summary, if $\alpha, \beta \in \mathbb{C}$ (the complex numbers) then the adjoint produces the following result.

$$[\alpha|v\rangle + \beta|w\rangle]^+ = \alpha^* \langle v| + \beta^* \langle w|$$

Because the dual space satisfies the vector space properties in Section 2.1, you therefore already know a great deal on how to manipulate the projection operators.

How does the adjoint operator affect other operators besides the bra? For the sake of argument, let $\hat{L}, \hat{L}_1, \hat{L}_2$ be linear operators that act on the vector space V which has basis vectors $\{|1\rangle, |2\rangle, \dots, |n\rangle\}$. For example, consider $\hat{L} = \hat{L}_1 \hat{L}_2$. Note that $\hat{L}$ does not need to be a projection operator (there are other types of operators); however, as seen in a subsequent section, the linear operators can be written as combinations of bras and kets. As a prototype, consider $\langle w | \hat{L} | v \rangle$ where $|v\rangle, |w\rangle \in V$ and $\langle w| \in V^+$. We will see in Chapter 3 that the adjoint operator reverses the direction of all the objects and adds the “+” to each operator (see Figure 2.4).

$$\langle v | L_1 L_2 | w \rangle^+ = \langle w | L_2^+ L_1^+ | v \rangle.$$

The adjoint operator maps a basis set for V into a corresponding basis set for V^+ . If $\{|i\rangle : i = 1, \dots, n\}$ comprises a basis set for V then $\{\langle i| : i = 1, \dots, n\}$ will be a basis set for V^+ . Therefore the dual basis set consists of operators that project an arbitrary vector onto the set of basis vectors of the vector space V . The dual basis allows us to write an arbitrary bra as $\langle v| = \sum_n B_n \langle n|$ without reference to the ket $|v\rangle$. That is, we did not need to know $|v\rangle$ before writing $\langle v|$ since we only needed to know the basis vectors for the vector space V^+ . However, if we were given $|v\rangle = \sum_n A_n |n\rangle$ then we would be able to write $B_n = A_n^*$ when $\langle n| = |n\rangle^+$.

Finally, let us check a few more properties to show that the set of projection operators forms a vector space (the dual space). That is, we want to show that the set spanned by the elemental projection operators

$$V^+ = \text{Sp}\{|1\rangle\langle 1|, |2\rangle\langle 2|, \dots\}$$

is a vector space. We must show that all the properties for the definition of vector space hold. The proof proceeds by appealing to the properties of the space V . We will prove just two, namely, closure under multiplication by a constant and commutivity of vector addition. All of the other properties are shown in the same manner.

Closure under multiplication by a constant: Let $\alpha \in \mathbb{C}$ and $|v\rangle \in V$ so therefore $\alpha|v\rangle \in V$. Using the adjoint, we conclude $(\alpha|v\rangle)^+ \in V^+$ or equivalently $\alpha^* \langle v| \in V^+$ as required.

Commutivity of vector addition: V is a vector space and so $|v\rangle + |w\rangle = |w\rangle + |v\rangle$. Next use the adjoint to write $(|v\rangle + |w\rangle)^+ = (\langle w| + \langle v|)^+$ or equivalently as $\langle v| + \langle w| = \langle w| + \langle v|$.

Example 2.14

Find the vector dual to $|2\rangle = \hat{y}$.

The dual vector is $\langle 2| = \hat{y} \cdot$ which is an operator that projects an arbitrary vector $\vec{v}$ onto $\hat{y}$. We can explicitly represent the result of the projection as the y -component of $\vec{v} = v_x \hat{x} + v_y \hat{y}$:

$$\langle 2|v\rangle = \hat{y} \cdot \vec{v} = v_y$$

Example 2.15

Some relations can be demonstrated for $\vec{v} = |v\rangle = a|1\rangle + b|2\rangle$ where $\{|1\rangle, |2\rangle\}$ spans R^2 .

1. $\langle v| = |v\rangle^+ = [a|1\rangle + b|2\rangle]^+ = a^*\langle 1| + b^*\langle 2|$
2. $\langle v|1\rangle = [a^*\langle 1| + b^*\langle 2|]|1\rangle = a^*$ and $\langle 1|v\rangle = \langle 1|[a|1\rangle + b|2\rangle] = a$
3. $\langle 1|v\rangle = a = (a^*)^* = \langle v|1\rangle^*$. Note that $\langle v|1\rangle^+ = \langle 1|v\rangle = \langle v|1\rangle^*$.

2.2.5 INNER PRODUCT AND NORM

Now we will see two applications for the adjoint operator. Part of the reason for defining the antilinear map (as opposed to a linear map) for the adjoint is for the norm (i.e., length) of a complex vector. In the following calculation, note in particular that the inner product previously defined can now be viewed as two separate quantities. The inner product can be “pulled apart.” Also note the use of the adjoint to help with the calculation.

First we find the norm of a vector $|v\rangle$ defined on a 3-D vector space spanned by the basis set $\{|i\rangle: i = 1, 2, 3\}$; that is we assume $|v\rangle = \sum_{i=1}^3 v_i |i\rangle$. The norm (or length) of a vector is found by taking the square root of the inner product $\|\vec{v}\|^2 = \langle v|v\rangle$. The adjoint facilitates the calculation by allowing us to “pull apart” the inner product as

$$\|\vec{v}\|^2 = \langle v|v\rangle = \left(\sum_{i=1}^3 v_i |i\rangle \right)^+ \left(\sum_{j=1}^3 v_j |j\rangle \right)$$

Now use the properties of the adjoint to write

$$\|v\|^2 = \sum_{i=1}^3 \langle i|v_i^* \sum_{j=1}^3 v_j |j\rangle = \sum_{i,j=1}^3 \langle i|v_i^* v_j |j\rangle = \sum_{i,j=1}^3 v_i^* v_j \langle i|j\rangle$$

The last step follows since $v_i^* v_j$ is just a number and so it can be moved outside the brackets. Now use the orthonormality property for unit vectors to write

$$\|v\|^2 = \sum_{i,j=1}^3 v_i^* v_j \delta_{ij} = \sum_{i=1}^3 v_i^* v_i = \sum_{i=1}^3 |v_i|^2$$

where $|v_i|$ is the magnitude of the complex number. This is nothing but Pythagorean’s theorem.

We can show the same results using the closure relation. Starting with the definition of norm, then pulling apart the inner product and inserting a unit operator (an operator that maps a vector into itself) provides

$$\|v\|^2 = \langle v|v\rangle = \langle v|\hat{1}|v\rangle$$

Note the use of the karat to denote operators. Assuming the same 3-D space spanned by $\{|i\rangle: i = 1, 2, 3\}$, we replace the unit operator with the closure relation $\hat{1} = \sum_{i=1}^3 |i\rangle\langle i|$ to find

$$\|v\|^2 = \langle v | \sum_{i=1}^3 |i\rangle\langle i| v \rangle = \sum_{i=1}^3 \langle v | i \rangle \langle i | v \rangle$$

Using the expansion $|v\rangle = \sum_{i=1}^3 v_i |i\rangle$, one finds $\langle i | v \rangle = v_i$ and also

$$\langle v | i \rangle = \langle i | v \rangle^+ = \langle i | v \rangle^* = v_i^*$$

where the adjoint becomes the complex conjugate since $\langle i | v \rangle$ is a complex number. Putting these results together produces the desired results.

$$\|v\|^2 = \sum_{i=1}^3 v_i^* v_i = \sum_{i=1}^3 |v_i|^2$$

Example 2.16

Calculate $\|\vec{v}\|$ for $\vec{v} = 3\hat{x} - 2i\hat{y} + \hat{z}$ where $i = \sqrt{-1}$

$$\begin{aligned} \|\vec{v}\|^2 &= \langle v | v \rangle = (3|1\rangle - 2i|2\rangle + |3\rangle)^+ (3|1\rangle - 2i|2\rangle + |3\rangle) \\ &= (3\langle 1| + 2i\langle 2| + \langle 3|)(3|1\rangle - 2i|2\rangle + |3\rangle) \\ &= 9 + 4 + 1 = 14 \end{aligned}$$

The length of the vector is seen to be $\|v\| = \sqrt{14}$

Example 2.17

Use the definition of dot product to find the length $\|\alpha|i\rangle\|$.

$\|\alpha|i\rangle\|^2 = \|\alpha\tilde{e}_i\|^2 = (\alpha\tilde{e}_i)^* \cdot (\alpha\tilde{e}_i) = |\alpha|^2 \tilde{e}_i \cdot \tilde{e}_i = |\alpha|^2$ since $\tilde{e}_i$ are real. This can be equivalently written as $\|\alpha|i\rangle\|^2 = (\alpha|i\rangle)^+ \alpha|i\rangle = \langle i | \alpha^* \alpha | i \rangle = |\alpha|^2 \langle i | i \rangle = |\alpha|^2$.

2.3 INTRODUCTION TO COORDINATE AND VECTOR REPRESENTATION OF FUNCTIONS

The present section introduces the paradigm of a function as a vector in Hilbert space. Often one encounters what appears to be two separate views whereby either a function is projected into “coordinate space” or it is projected into function space (i.e., the function is a linear combination of a function basis set). We start with the “coordinate space” view and then indicate the link with the second view that functions can only be composed of other functions. The reader should consult Section 2.5.2 on coordinate space to complete the initial view contained in the present section.

The present section discusses a representation-independent view of functions reminiscent of denoting vectors by $|v\rangle$ rather than by the components $v_i = \langle i | v \rangle$ which require a specific coordinate system. The discussion on the coordinate representation of functions shows how projecting a function vector $|f\rangle$ into coordinate space, denoted by $\{|x\rangle\}$, provides the values of the function $f(x) = \langle x | f \rangle$. The vector $|f\rangle$ has a magnitude and direction in a vector space spanned by a basis of orthonormal functions. This section illustrates the coordinate representation of a function and clearly demonstrates the similarity between the function inner product $\langle f | g \rangle$ and the vector dot product.

The coordinate space representation of a function actually comes from properties of a Hilbert space of functions and generalized expansions in basis sets. Functions in the Hilbert space can be expanded in a basis set of functions. The basis set most commonly consists of other functions (such as sines and cosines). We will see that the view of a “function projected into coordinate space” and the view of the “function projected onto a function basis set” are really the same by defining the coordinate ket $|x\rangle$ to mean the function ket for the Dirac delta function $|\delta(x' - x)\rangle$. Although the two representations must be related, one sometimes thinks of projecting a function into coordinate space as being different (for convenience and efficiency) from projecting it into a Hilbert space with basis functions. Subsequent sections will make the concept clear.

2.3.1 INITIAL VIEW OF THE COORDINATE REPRESENTATION OF FUNCTIONS

First let us review how a Euclidean vector can be viewed as a function. A Euclidean vector $\vec{v}$ in a vector space has one component v_i for each value of the index i . The component v_i comes from projecting the vector $\vec{v}$ onto the basis vector $|i\rangle$ so that $v_i = \langle i|\vec{v}\rangle$. We can write the components v_i as a function $v(i) = v_i$ since the set of ordered pairs $\{(i, v_i)\}$ defines a function. The example in Figure 2.5 shows the vector $\vec{v} = 2.5\hat{x} + 0\hat{y} + 4\hat{z}$ as a function

$$v_i = \begin{cases} 2.5 & i = 1 \\ 0 & i = 2 \\ 4 & i = 3 \end{cases}$$

The inner product $\langle i|\vec{v}\rangle$ changes a “physical object” $|\vec{v}\rangle$ into a function $v(i) = v_i$ by projecting $\vec{v}$ onto the i -axis (i.e., projecting $\vec{v}$ onto the i -coordinates). That is, we think of i as a coordinate (i.e., a number) rather than related to the unit vector $|i\rangle$. The reader should carefully consider this shift in paradigm.

It should not be too surprising to find that functions such as $f(x)$ can be thought of as vectors $|f\rangle$ projected onto the x -axis. To set the stage, recall how functions such as in Figure 2.6 can be described as a collection of ordered pairs (x, f) . For some values of x , the function f might be discrete (as opposed to

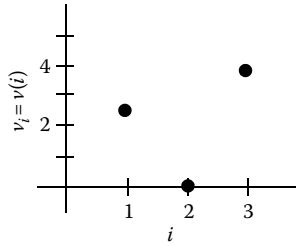


FIGURE 2.5 The components of the vector $\vec{v} = 2.5\hat{x} + 0\hat{y} + 4\hat{z}$ are functions of the variable i .

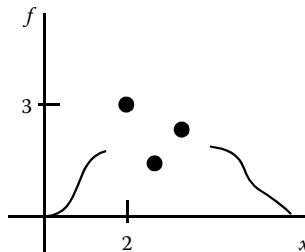


FIGURE 2.6 x is both a continuous and discrete index for the function f .

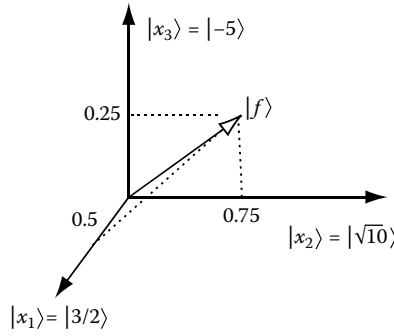


FIGURE 2.7 The function f projected onto several coordinates.

continuous) such as at the distinct points in Figure 2.6. Compare Figure 2.6 with Figure 2.5. The only significant difference is that $v(i)$ has a domain with a countable number of “ x -components” symbolized by i whereas f has regions of continuous values of x . We can consider x to be an index and write $f(x) = f_x$ where x takes on all values in the domain. Figure 2.6 show the example of $f(2) = f_2 = 3$.

Now, lets see how the function $f(x)$ can be pictured as a vector in coordinate space defined by $\{|x\rangle\}$. Imagine projecting a function f onto the coordinate x as $\langle x|f\rangle = f(x)$. Each real x is considered to be a basis vector as shown in Figure 2.7. Keep in mind that basis vectors have unit length which therefore has nothing to do with the magnitude of x . As a result, the different x corresponds to different directions. The figure shows just three values of x ; however, there can be an uncountably infinite number of “vectors” $|x\rangle$. For the figure, some of the values of the function $f(x)$ can be written as

$$\langle x|f\rangle = f(x) = \begin{cases} 0.5 & x = 3/2 \\ 0.75 & x = \sqrt{10} \\ 0.25 & x = -5 \end{cases}$$

or, in terms of “components” $\langle 3/2|f\rangle = 0.5$, $\langle \sqrt{10}|f\rangle = 0.75$, $\langle -5|f\rangle = 0.25$. Notice that the function is actually an abstract vector defined by $|f\rangle$. The figure shows three-axes but there must be as many axes as coordinates x . The value of the function at point x , namely $f(x)$, is actually the component of the vector along axis $|x\rangle$. Given the similarity to the Euclidean vectors, one expects to be able to write something of the form $|f\rangle \sim 0.5|x_1\rangle + 0.75|x_2\rangle + 0.25|x_3\rangle$. Such an expansion represents the correct “way of thinking” but the summation needs to be replaced with an integral. One should also note that quantities of the form $\langle f|x\rangle$ can now be defined using the adjoint operator as

$$\langle f|x\rangle = \langle x|f\rangle^+ = \langle x|f\rangle^* = [f(x)]^* = f^*(x) \quad (2.9)$$

2.3.2 COORDINATE BASIS SET

The previous section showed the interpretation of projecting the vector $|f\rangle$ onto the coordinate ket $|x\rangle$ produced the component of the vector, namely the value of the function $f(x) = \langle x|f\rangle$. However, the symbol $\langle \bullet|\bullet\rangle$ represents an inner product between vectors. In this case for functions, one expects the inner product to be an integral between two functions. However, the “ x ” appears to be a coordinate and not a function. Notice that we cannot define $g(x) = x$ in the inner product to use in an inner-product integral. Our previous discussion provided a “way to think about” the projection of an

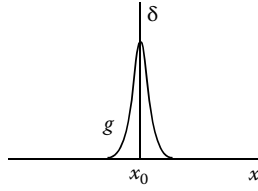


FIGURE 2.8 A Gaussian function $g(x)$ centered about the point x_0 . Taking the standard deviation of the function g to zero produces the Dirac delta function δ .

abstract function vector into “coordinate space.” Now we must show how this works. The reader should review the material in the first two sections of Appendix B covering the Dirac delta function.

When one projects a vector $|v\rangle$ onto basis vector $|n\rangle$ such as for Euclidean vectors, one is actually determining “how much” of vector $|v\rangle$ consists of vector $|n\rangle$. The same idea holds for projecting a function vector $|f\rangle$ onto the coordinate-function $|x_0\rangle$ (note the use of a specific coordinate x_0 in this case). The question becomes, what function best represents the coordinate x_0 ? Figure 2.8 shows an example of a Gaussian function $g(x)$ (in the calculus sense of the word “function” as opposed to the abstract function-vector used here) localized about the point x_0 . Taking the width of the function g to zero produces the Dirac delta function defined as $\delta(x - x_0)$ which is very sharply peaked at x_0 . That is, all of the weight of the delta function is at x_0 ; it represents the single point x_0 . For the function $|f\rangle$, one wants to know how much like $\delta(x - x_0)$ is $|f\rangle$ in the sense of the following questions. Is all of the weight of f at x_0 ? If not, how much of the weight of f is at x_0 ? This last question essentially asks what is the value of f at x_0 ? The reader should realize these are conceptual questions and one must rely on the mathematical definitions for their answers. We will see later, that to say a quantum particle is localized at a point x_0 is equivalent to saying it has a wave function similar to g or δ in Figure 2.8. Therefore, when we write $|x_0\rangle$, we really refer to the abstract function-vector consisting of the Dirac delta function.

Now to answer the question as to how much like $\delta(x - x_0)$ is $|f\rangle$, we just need to project f onto the function $\delta(x - x_0)$ which then forms the inner product $\langle \delta(x - x_0) | f(x) \rangle$. We find based on the properties of the Dirac delta function (also known as an impulse function)

$$\langle x_0 | f \rangle = \langle \delta(x - x_0) | f(x) \rangle = \int_{-\infty}^{\infty} dx \delta(x - x_0) f(x) = f(x_0) \quad (2.10)$$

where the Dirac delta function is real valued and we assume the domain of integration includes the point x_0 . The coordinate basis set is really a “continuous” set of Dirac delta functions. The bra $\langle x_0 |$ actually represents a projection operator that projects $|f\rangle$ onto the Dirac delta function $\delta(x - x_0)$. That is, coordinate space $\{|x_0\rangle\}$ really consists of the collection of basis vectors defined to be the Dirac delta functions $\{\delta(x - x_0)\}$.

So now we have another way to think of the decomposition of a vector $\vec{f} = |f\rangle$. We imagine f projected onto basis vectors of the form $\{\delta(x - x_0) : x_0 \in R\}$. There are an infinite number of such basis vectors. The procedure disassembles the function f into components. Adding up the components (with the basis vectors attached) will reproduce the original functions. We will see the procedure results in the closure relation for the coordinate basis set

$$\int |x\rangle dx \langle x| = \hat{1}$$

The actual demonstrations must wait until we discuss function spaces with an uncountable number of basis vectors (the continuous basis set).

Finally, the inner product between two coordinate kets can be demonstrated. Let $|\varepsilon\rangle$ and $|\eta\rangle$ be two kets with ε and η in the real interval (a, b) . The inner product can be written as

$$\langle\xi|\eta\rangle = \langle\delta(x-\xi)|\delta(x-\eta)\rangle = \int_a^b dx \delta(x-\xi)\delta(x-\eta) = \delta(\xi-\eta) \quad (2.11a)$$

where this equation uses the reality of the delta function $\delta^* = \delta$. Also note the delta function is even and the inner product can be written in either of two ways as

$$\langle\xi|\eta\rangle = \delta(\xi-\eta) = \delta(\eta-\xi) \quad (2.11b)$$

The inner product between coordinate kets has the form of the Dirac delta function rather than the Kronecker delta used for the Euclidean vectors.

2.3.3 INTRODUCTION TO THE INNER PRODUCT FOR FUNCTIONS

The notion of a function looking like a vector in coordinate space is related to the form of the inner product for function space. The notion will be substantiated in the next few sections by using the Dirac delta function. For now recall that an inner product like $\langle g|h\rangle$ is taken between vectors g and h . Recall for Euclidean vectors that

$$\langle g|h\rangle = \sum_i \langle g|i\rangle \langle i|h\rangle = \sum_i g_i^* h_i$$

where $\langle g|i\rangle = \langle i|g\rangle^+ = g_i^*$ since the inner product $\langle i|g\rangle$ is a complex number. Now suppose that g and h are functions so that the index i is replaced by the index x . The inner product might then be written as

$$\langle g|h\rangle \sim \sum_x \langle g|x\rangle \langle x|h\rangle \sim \sum_x g^*(x)h(x) \sim \sum_x g_x^* h_x \sim \int dx g_x^* h_x \sim \int dx g^*(x)h(x)$$

Therefore, for functions, the inner product

$$\langle g|h\rangle = \int dx g^*(x)h(x) \quad (2.12)$$

is viewed as a sum over components similar to the case for Euclidean vectors.

2.3.4 REPRESENTATIONS OF FUNCTIONS

An issue concerns why one should work with the abstract function vector $|f\rangle$ rather than $f(x)$. One answer involves the quantum theory. The abstract vectors provide a convenient (and common) means of visualizing many abstract concepts in the quantum theory. In fact, the new paradigm provides a framework for the interplay between nature and mathematics and simplifies many calculations.

Another point worth discussing concerns the notion of representations. Consider the function written as $|f\rangle$. The function can be represented in x -coordinate space as $f(x)$ or a Fourier transform space as $f(k)$ or in a Taylor series or by any number of trigonometric expansions (and so on). All of these representations of f bring out important characteristics unique to the representation.

For example, an expansion in sinusoids might be useful for circuit analysis and its important to have the frequency content of the function f . One cannot say that a given representation has more importance than another outside the context of the application. Any representation can be found from the abstract function $|f\rangle$ by a suitable choice of basis vectors. We have already seen that the Dirac delta function produces the x -coordinate representation. The Fourier representation (i.e., the Fourier transform) can be obtained by projecting onto the basis functions $\{e^{ikx}/\sqrt{2\pi}\}$. Many calculations can proceed without reference to any particular basis set. This can provide immense simplifications.

2.4 FUNCTION SPACE WITH DISCRETE BASIS SETS

One important type of Hilbert space consists of functions and has a basis set with a countable number of elements (such as sines and cosines). Functions in the Hilbert space can be expanded in this basis set of functions and give rise to the Fourier series type expansions. The reader should refer to Section 2.7 and the end-of-chapter review exercises for specific examples of basis sets functions.

The section first discusses the Hilbert space of functions. It then develops the notation for those Hilbert spaces that have a discrete basis set. The results quite straightforwardly generalize the concepts for the Euclidean vectors. In fact, if readers were not warned ahead of time, they might think they were reading about Euclidean vectors all over again. The section then explores alternate expressions for the inner product and norm for function spaces with a discrete basis using a coordinate basis set.

2.4.1 INTRODUCTION TO HILBERT SPACE

Hilbert space consists of a vector space having a defined inner product. For the present section, we describe the Hilbert space of functions. One views the function $|f\rangle$, (see Figure 2.9 for example), as a vector with magnitude and direction.

The vector picture has the most meaning when viewed with respect to a basis set such as $\{|\phi_1\rangle, |\phi_2\rangle, \dots\}$. Each basis vector (i.e., basis function) has unit length and each is orthogonal to every other such basis vector. The notions of length and orthogonality require the definition of an inner product. A vector space can have an infinite number of different basis sets; the choice of one particular such set depends on one's preferences for a particular situation. One views the basis vectors as fundamental. For example, the set of cosines and sines form a basis set for the vector space of periodic functions. In such a case, the summation

$$|f\rangle = \sum_{i=1}^{\infty} c_i |\phi_i\rangle \quad (2.13)$$

represents the Fourier series expansion of the function.

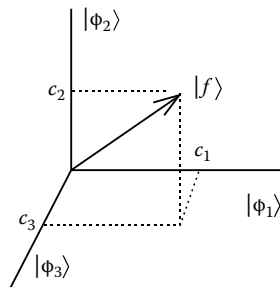


FIGURE 2.9 The function f projected onto the basis set of functions.

The direction of $|f\rangle$ determines the composition of the function with respect to the basis functions. Viewing $|f\rangle$ as a vector shows that c_i represents the component of $|f\rangle$ along the unit vector $|\phi_i\rangle$. Projecting $|f\rangle$ onto $|\phi_i\rangle$ provides the component $c_i = \langle\phi_i|f\rangle$ which symbolizes the inner product between the function ϕ_i and f . Larger values of the coefficient c_i in Equation 2.13 means that $|f\rangle$ and $|\phi_i\rangle$ are more similar. For Fourier series for example, the c_i represent the amount of a particular sinusoid in the function $|f\rangle$.

The magnitude of the function consists of the sum of the square of components of the function along each basis vector. This is similar to the Euclidean vectors and Pythagoreans theorem. A “larger” function has greater amounts of each basis vector; this also means that the function tends to have larger values in the usual sense of $f(x)$. This paradigm provides visual aid for solutions to partial differential equations where the solution function can change angles and length when viewed against the basis vectors (i.e., the components c_i can change with time in Figure 2.9). For quantum theory of closed systems, both the functions f representing physical systems and the basis vectors have unit length as we will see in later chapters.

2.4.2 HILBERT SPACE OF FUNCTIONS WITH DISCRETE BASIS VECTORS

A function $|f\rangle$ (Dirac notation), in a finite-dimensional vector space can be written as an orthonormal expansion

$$|f\rangle = \sum_{i=1}^{\infty} c_i |\phi_i\rangle \quad (2.14)$$

of basis functions

$$B = \{|\phi_0\rangle, |\phi_1\rangle, |\phi_2\rangle, \dots, |\phi_n\rangle\} \quad (2.15)$$

that span the space (Figure 2.9). Sometimes for convenience the basis vectors will be represented by their index so that the summation in Equation 2.14 and the basis set in Equation 2.15 have the alternate forms

$$|f\rangle = \sum_{i=1}^{\infty} c_i |i\rangle \quad \text{and} \quad B = \{|0\rangle, |1\rangle, |2\rangle, \dots, |n\rangle\} \quad (2.16)$$

The basis set B is “discrete” (i.e., the set B is countable) when there exists a 1–1 correspondence between the elements of B and a subset of the integers (or more exactly, a subset of the “whole” numbers) as is obviously the case for Equations 2.15 and 2.16. Projecting the function onto any of the “axes” defined by the unit vectors $|\phi_i\rangle$ yields the constants c_i . The projection of the function on the i th axis has the form of an inner product between the two complex functions ϕ_i and f over a range (a, b)

$$\langle\phi_i|f\rangle = \int_a^b dx \phi_i^*(x) f(x) \quad (2.17)$$

The inner product in Equation 2.17 comes about by applying the projection operator $\langle\phi_i|$ to the function $|f\rangle$. The projection operator (i.e., bra) for functions has the form

$$\langle\phi_i| = \int_a^b dx \phi_i^*(x) \circ \quad (2.18a)$$

where the small circle represents a place holder for the function f . This has a form similar to that for the Euclidean vectors from the previous section

$$\langle v| = \vec{v} \cdot = \sum_i v_i^* \circ \quad (2.18b)$$

so that the integral replaces the summation in the case of functions.

Functions in a set $F = \{\phi_0, \phi_1, \phi_2, \dots, \phi_n\}$ are linearly independent if for complex constants c_i ($i=0, \dots, n$), the sum

$$\sum_{i=0}^n c_i \phi_i(x) = 0$$

can only be true when all of the complex constants are zero $c_i=0$. Functions in the set $F = \{\phi_0, \phi_1, \phi_2, \dots, \phi_n\}$ are orthonormal if $\langle \phi_i | \phi_j \rangle = \delta_{ij}$ for every integer i, j in the set $\{0, 1, 2, \dots, n\}$.

Orthonormal functions must be linearly independent as can be seen by starting with a sum over the functions ϕ_i with complex coefficients c_i , specifically $\sum_i c_i \phi_i = 0$ or $\sum_i c_i |\phi_i\rangle = 0$. Next, operate on both sides with the bra operator $\langle \phi_m |$ to find $\sum_i c_i \langle \phi_m | \phi_i \rangle = 0$. Using the orthonormality $\delta_{mi} = \langle \phi_m | \phi_i \rangle$ shows that the complex coefficients must all be 0 (i.e., $c_i=0$ for all i).

A linearly independent set of functions $F = \{\phi_0, \phi_1, \phi_2, \dots, \phi_n\}$ is complete if every function $f(x)$ in the space can be written as

$$f(x) = \sum_{i=0}^n c_i \phi_i(x) \quad \text{or} \quad |f\rangle = \sum_{i=0}^n c_i |\phi_i\rangle \quad (2.19)$$

(except at possibly a few discrete points) for some choice of complex numbers c_i . If the set $\{\phi_i\}$ is “complete and orthonormal” then the functions ϕ_i can be chosen as basis functions (or basis vectors) to span the function space. A complete orthonormal set of functions $F = \{\phi_0, \phi_1, \phi_2, \dots, \phi_n\}$ form a basis for a Hilbert space $\mathcal{H}$. In some cases, there might be a countable infinite number of basis vectors in which case the infinite series

$$|f\rangle = \sum_{i=0}^{\infty} c_i |i\rangle \quad (2.20a)$$

must properly converge. Assume that the series has the appropriate convergence properties so that it can be integrated or differentiated as necessary. Notice the similarity between these formulas and those for the Euclidean space.

The components of the vector $|f\rangle$ (i.e., the expansion coefficients c_i) can be found from Equation 2.20a by operating with the bra $\langle j|$ as follows

$$\langle j|f\rangle = \langle \phi_j|f\rangle = \langle j| \sum_{i=0}^{\infty} c_i |i\rangle = \sum_{i=0}^{\infty} c_i \langle j|i\rangle = \sum_{i=0}^{\infty} c_i \delta_{ij} = c_j \quad (2.20b)$$

so, similar to Euclidean vectors, the vector components must be $c_j = \langle j|f\rangle$.

The norm of a function $\|f\|$ can be found from

$$\|f\|^2 = \langle f|f\rangle = \int dx f^*(x) f(x) \quad (2.21)$$

Any vector f can be normalized to have unit length by setting

$$f \rightarrow f/\|f\| \quad (2.22)$$

This can be seen by evaluating the inner product as follows

$$\left\langle \frac{f}{\|f\|} \middle| \frac{f}{\|f\|} \right\rangle = \frac{1}{\|f\|^2} \langle f|f \rangle = \frac{1}{\|f\|^2} \|f\|^2 = 1$$

Example 2.18

Is the set $\{1, x\}$ orthonormal on the interval $[-1, 1]$?

The functions are orthogonal on the interval as can be seen

$$\langle 1|x \rangle = \int_{-1}^1 dx \, 1 \cdot x = 0$$

Carefully note how the symbols 1 and x represent functions and not coordinates. We must distinguish between functions and coordinates since notation can be confusing. Neither function is normalized to unit length since

$$\|1\|^2 = \langle 1|1 \rangle = \int_{-1}^1 dx = 2 \quad \text{and} \quad \|x\|^2 = \langle x|x \rangle = \int_{-1}^1 dx \, x^2 = \frac{2}{3}$$

An orthonormal set can be formed by dividing each function by its length. The orthonormal set is

$$\left\{ \frac{1}{\sqrt{2}}, \sqrt{\frac{3}{2}} x \right\}$$

Example 2.19

Project $g(x) = \sin\left(\frac{2\pi x}{L}\right)$ onto $f(x) = \sin\left(\frac{\pi x}{L}\right)$ over the interval $[0, L]$

The projection operator is

$$\langle f| = \int dx \, f^*(x) \circ = \int_0^L dx \, \sin\left(\frac{\pi x}{L}\right) \circ$$

Therefore

$$\langle f|g \rangle = \int dx \, f^*(x)g(x) = \int_0^L dx \, \sin\left(\frac{\pi x}{L}\right) \sin\left(\frac{2\pi x}{L}\right) = 0$$

2.4.3 CLOSURE RELATION FOR FUNCTIONS WITH A DISCRETE BASIS

The closure relation provides a mathematical formula that expresses the completeness property of a set of orthonormal vectors $\{|0\rangle, |1\rangle, \dots\}$. For this reason, the closure relation can be alternatively

termed the completeness relation. To demonstrate the closure relation, let $|f\rangle$ be an arbitrary element of the Hilbert space and write it as a sum over the complete set of unit vectors

$$|f\rangle = \sum_{i=1}^{\infty} c_i |\phi_i\rangle$$

The definition of the vector components $c_i = \langle \phi_i | f \rangle = \langle i | f \rangle$ can be used to write

$$|f\rangle = \sum_{i=0}^{\infty} c_i |i\rangle = \sum_{i=0}^{\infty} \langle i | f \rangle |i\rangle = \sum_{i=0}^{\infty} |i\rangle \langle i | f \rangle = \left(\sum_{i=0}^{\infty} |i\rangle \langle i| \right) |f\rangle \quad (2.23a)$$

The vector $|f\rangle$ is an arbitrary member of the Hilbert space. Recall that two operators $\hat{A}, \hat{B}$ are equal if and only if $\hat{A}|v\rangle = \hat{B}|v\rangle$ for all vectors $|v\rangle$ in the Hilbert space. Notice the use of the karat to signify an operator. Therefore, by definition of equality between operators, Equation 2.23a yields

$$\sum_{i=0}^{\infty} |\phi_i\rangle \langle \phi_i| = \hat{1} \quad \text{or} \quad \sum_{i=0}^{\infty} |i\rangle \langle i| = \hat{1} \quad (2.23b)$$

Similar to Section 2.2.3, the closure relation ensures completeness of the basis set and vice versa.

The closure relation for functions such as in Equation 2.23b always has an alternate form involving actual functions rather than the abstract vectors in the basis set. Such expansions of functions are very useful for solving partial differential equations. The alternate form of the closure relation can be found by using the projection of one coordinate ket onto another as demonstrated in the previous section $\langle x' | x \rangle = \delta(x - x')$. Starting with the coordinate ket inner product and inserting the resolution of unity for the function space produces

$$\delta(x - x') = \langle x' | x \rangle = \langle x' | \hat{1} | x \rangle = \langle x' | \left\{ \sum_{i=0}^{\infty} |\phi_i\rangle \langle \phi_i| \right\} | x \rangle = \sum_{i=0}^{\infty} \phi_i^*(x) \phi_i(x') \quad (2.24)$$

As a note, the completeness relation represents a special case for a basis expansion of an operator as we will see in Chapter 3. An arbitrary linear operator that maps a vector space V into itself according to $\hat{T}: V \rightarrow V$ for example, has a basis expansion of the form

$$\hat{T} = \sum_{a,b} T_{ab} |a\rangle \langle b| \quad (2.25)$$

where T_{ab} is the matrix of the operator $\hat{T}$ in the basis of V . The kets $|a\rangle, |b\rangle$ represent basis vectors for V and interestingly, the collection of all $|a\rangle \langle b|$ forms a basis for the vectors space of linear operators $\hat{T}$. Equation 2.25 looks very similar to Equation 2.23b in that $\hat{T}$ can be written as a summation over all basis vectors. We will discuss this further in Chapter 3.

2.4.4 NORMS AND INNER PRODUCTS FOR FUNCTION SPACES WITH DISCRETE BASIS SETS

So far, we have introduced the discrete basis set of functions. We have also introduced coordinate space as an example of a continuous basis set. Now we apply the closure relation for the functions with a discrete basis to finding the norm of a function. We also compare the norm of a function with that for a Euclidean vector. We see that the index x takes the place of the index i , and the integral

replaces the summation. Similar comments apply to the inner product of two functions. As a reminder, some books denote the inner product for functions as (f, g) but then miss out on the convenient form for the projection operators and the closure relation.

Norm of a Function Compared with Norm of a Euclidean Vector

Euclidean Vectors

$$\begin{aligned}
 \|\vec{v}\|^2 &= \langle v|v \rangle \\
 &= \langle v|1|v \rangle \\
 &= \langle v| \left[\sum_{i=1}^n |i\rangle \langle i| \right] |v \rangle \\
 &= \sum_{i=1}^n \langle v|i\rangle \langle i|v \rangle \\
 &= \sum_{i=1}^n \langle i|v \rangle^+ \langle i|v \rangle \\
 &= \sum_{i=1}^n v_i^* v_i \\
 &= \sum_{i=1}^n |v_i|^2
 \end{aligned}$$

Functions

$$\begin{aligned}
 \|f\|^2 &= \langle f|f \rangle \\
 &= \langle f|1|f \rangle \\
 &= \langle f| \left[\int |x\rangle \langle x| dx \right] |f \rangle \\
 &= \int \langle f|x \rangle dx \langle x|f \rangle \\
 &= \int \langle x|f \rangle^+ dx \langle x|f \rangle \\
 &= \int f^*(x) f(x) dx \\
 &= \int |f(x)|^2
 \end{aligned}$$

Inner Product between Functions Compared with Euclidean Vectors

Euclidean Vectors

$$\begin{aligned}
 \langle v|w \rangle &= \langle v| \sum_i |i\rangle \langle i| |w \rangle \\
 &= \sum_i \langle v|i\rangle \langle i|w \rangle \\
 &= \sum_i v_i^* w_i
 \end{aligned}$$

Functions

$$\begin{aligned}
 \langle f|g \rangle &= \langle f| \left[\int |x\rangle \langle x| dx \right] |g \rangle \\
 &= \int dx \langle f|x \rangle \langle x|g \rangle \\
 &= \int dx f^*(x) g(x)
 \end{aligned}$$

2.4.5 DISCUSSION OF WEIGHT FUNCTIONS

Often eigenfunctions of some specific operator, say $\hat{L}$, are chosen as a basis set. Sometimes an independent variable such as θ is chosen rather than one, say x , that produces an orthonormal set. In such cases, one typically includes a weight function, denoted W , in the definition of inner product to enforce the orthonormality in the variable θ . Changing the variables in $\hat{L}$ changes the weight function $w(x)$. For example, if the operator $\hat{L}$ has “Legendre” polynomials as eigenfunctions, and the independent variable is the angle θ , then the “weight” function is proportional to $\sin(\theta)$. But if a new variable $x = \cos(\theta)$ is defined in the operator then the weight function $w(x) \rightarrow 1$. The spherical harmonics provide another example. The “moral of the story” is that sometimes a problem initially implements a set of coordinates which “hides” the orthonormality of the eigenfunctions X_m . The reader can skip the section without loss of continuity.

The change of variables influences the form of the inner product. Consider function $X_m(x)$ that satisfy the typical orthonormality relation

$$\delta_{mn} = \langle X_m(x) | X_n(x) \rangle = \int_a^b dx X_m^*(x) X_n(x) \quad (2.26)$$

If one changes the variable x to y through the relation $x = \eta(y)$ then

$$\frac{dx}{dy} = \eta'(y) \quad \text{and} \quad dx = \eta'(y) dy \quad (2.27)$$

where one assumes the inverse to η exists so that $y = \eta^{-1}(x)$. Define $w(y) = \eta'(y)$ and also change coordinates in X to find $X(x) = X(\eta(y))$. This is a composite function normally written as $X \circ \eta$. It is customary to write

$$X(y) = X(\eta(y)) \quad (2.28)$$

even though $X(y)$ is a new function and not merely substituting y for x in $X(x)$. For example, if $x = \cos \theta$ then $X(x) = X(\cos \theta)$ is actually a new function $\xi(\theta)$.

Substituting Equations 2.28 and 2.27 into the inner product in Equation 2.26 provides

$$\delta_{mn} = \int_{a'}^{b'} dy \, w(y) X_m^*(y) X_n(y) \quad (2.29)$$

Note the new limits on the integral. The new “basis” functions $X(y)$ satisfy a new orthonormality condition that involves a weight function $w(y)$. By the way, y is a dummy variable and can be changed to x (just to add to the confusion) and Equation 2.29 can be rewritten as

$$\delta_{mn} = \int_{a'}^{b'} dx \, w(x) X_m^*(x) X_n(x)$$

We will avoid this type of substitution for dummy variables. For the new variable $y = y(x)$, the inner product between any two functions in the Hilbert space has the form

$$\langle f|g \rangle = \int_{a'}^{b'} dy \, w(y) f^*(y) g(y)$$

Sometimes to be explicit, people write this new inner product as

$$\langle f|wg \rangle \quad \text{or as} \quad \langle f|g \rangle_w$$

just as a reminder that a weight function is involved. We use the notation

$$\langle f|wg \rangle = \int_{a'}^{b'} dy \, w(y) f^*(y) g(y)$$

In this manner, $\langle f|g \rangle$ retains its “old” meaning of

$$\langle f|g \rangle = \int_a^b dx \, f^*(x) g(x)$$

The reader should note that a change of variable $x = \eta(y)$ affects all of the functions in Hilbert space...essentially giving a new Hilbert space. If $g(x)$ is in the original Hilbert space and if $g: R^{(b)} \rightarrow R^{(c)}$ (i.e., $x \in R^{(b)} \rightarrow g(x) \in R^{(c)}$) and if $\eta: R^{(a)} \rightarrow R^{(b)}$ (i.e., $y \in R^{(a)} \rightarrow \eta(y) = x \in R^{(b)}$)

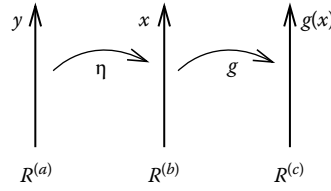


FIGURE 2.10 Relation between the various mappings.

where $R^{(a)}, R^{(b)}, R^{(c)}$ are copies of the set of real numbers (Figure 2.10). The function $g(x)$ can be written as

$$g(x) = g[\eta(y)] = (g \circ \eta)(y)$$

where $\circ$ indicates the composition of two functions. The functions in the Hilbert space are being replaced by new functions

$$g \circ \eta: R^{(a)} \rightarrow R^{(c)}$$

So essentially, the function $g(y)$, which conventionally replaces $g(x)$ for a change of variables, should really be written as $(g \circ \eta)(y)$.

Now how are orthonormal expansions handled when using the weight function? Let us use the same notation

$$x = \eta(y) \quad \text{and} \quad w(y) = \eta'(y) \quad (2.30)$$

For the coordinate x , the inner products are the same as always.

$$\langle f(x) | g(x) \rangle = \int_a^b dx f^*(x) g(x)$$

The orthonormality relation for basis functions $|\phi_i\rangle$ is

$$\langle \phi_m(x) | \phi_n(x) \rangle = \delta_{mn} = \int_a^b dx \phi_m^*(x) \phi_n(x)$$

Expansion are given as usual by

$$f(x) = \sum_n \beta_n \phi_n(x)$$

with

$$\beta_n = \langle \phi_n(x) | f(x) \rangle = \int_a^b dx \phi_n^*(x) f(x)$$

Now, what about the coordinate y ? The inner product becomes, using Equation 2.30,

$$\langle f(x)|g(x)\rangle = \int_a^b dx f^*(x)g(x) = \int_{a'}^{b'} dy w(y)f^*(y)g(y) = \langle f(y)|w(y)g(y)\rangle \quad (2.31)$$

where the integral is transformed by the indicated change of variables. Notice that the real effect is to include $w(y)$ in the integrand. However Equation 2.31 shows that either inner product gives the same answer. The orthonormality relation among basis vectors is therefore

$$\delta_{mn} = \langle \phi_m(x)|\phi_n(x)\rangle = \langle \phi_m(y)|w(y)\phi_n(y)\rangle = \int_{a'}^{b'} dy w(y)\phi_m^*(y)\phi_n(y)$$

Again it is clear that the change of variables places a weight in the integrand and changes the limits. Finally, the expansions can be written as

$$f(y) = \sum_n \beta_n \phi_n(y)$$

with

$$\beta_n = \langle w\phi_n(y)|f(y)\rangle = \langle \phi_n(y)|w f(y)\rangle$$

where w can be switched from side to side since it is real.

Example 2.20

Suppose $X_\lambda(x, y, z)$ is an eigenfunction of an operator with

$$\delta_{mn} = \int dV X_m^*(x, y, z) X_n(x, y, z)$$

where $dV = dx dy dz$ is an element of volume. In xyz -coordinates, the integral can be written as

$$\delta_{mn} = \int dx dy dz X_m^*(x, y, z) X_n(x, y, z)$$

so that $w(x) = 1$. But for spherical coordinates r, θ, ϕ , the integral becomes

$$\delta_{mn} = \int dr d\theta d\phi (r^2 \sin \theta) X_m^*(r, \theta, \phi) X_n(r, \theta, \phi)$$

so the weight function is $w \sim r^2 \sin \theta$.

2.4.6 SOME MISCELLANEOUS NOTES ON NOTATION

The norm of $f - g$ can be written as (real functions)

$$\|f - g\|^2 = \int_a^b dx (f(x) - g(x))^2$$

but the average of $(f - g)^2$, is written as

$$\overline{(f - g)^2} \equiv \langle (f - g)^2 \rangle = \frac{1}{b - a} \int_a^b dx (f - g)^2$$

Similar averages hold for complex functions. Note the use of “<” and “>” to indicate an average. As we will see later, quantities of the form $\langle i|L|i \rangle$ can be viewed as an “average” or as a “matrix element.” These brackets are used all over and with many meanings! So careful! Note the different uses of the < > to indicate inner products and averages in the following.

$$\int_a^b dx (f - g)^2 = (b - a) \overline{(f - g)^2}$$

so that

$$\langle f - g | f - g \rangle = \|f - g\|^2 = (b - a) \langle (f - g)^2 \rangle$$

2.5 FUNCTION SPACES WITH CONTINUOUS BASIS SETS

The Hilbert space with a continuous basis set has important applications to the quantum mechanics (especially for free-space propagation) and to transform theory. This type of Hilbert space has an uncountably infinite number of basis vectors. The basis set is in 1–1 correspondence with a continuous subset of the real numbers. We will encounter situations where the basis set consists of a range of both continuous and discrete basis vectors. Furthermore, the section demonstrates the $\vec{r}$ coordinate space and the Fourier transform coordinate space.

So far we have developed new notation to show the similarity between Euclidean space and function space with a discrete set of basis functions. For both cases, the inner product between two basis vectors uses the Kronecker delta function and a vector in the space can be written as a discrete summation over the basis set. The Euclidean inner product reduces to a discrete summation over the components whereas the function space uses the integral over the components. For the continuous basis set, we will see that the inner product between two basis vectors produces the Dirac delta function and a general vector can be written as the integral (rather than the discrete summation) over the basis set. For the continuous basis set, the inner product reduces to an integral over the spatial components of two functions.

2.5.1 CONTINUOUS BASIS SET OF FUNCTIONS

Now we discuss the continuous basis set of functions. Let $B = \{\phi_k\}$ (i.e., $B = \{|\phi_k\rangle\}$) be a set of basis vectors with one such vector for each real number k in some interval $[a, b]$, where generally one should expect to have $a = -\infty$ or $b = +\infty$. The basis set is termed continuous not because the functions are continuous but for the reason that given $\phi_a, \phi_b \in B$ there does not exist c such that $a < c < b$ without ϕ_c also being in B . For continuous basis sets, the orthonormality relation has the form (The reader should consult Section 2.7 for specific examples of the continuous basis set of functions such as for the Fourier transform.)

$$\langle \phi_K | \phi_k \rangle = \delta(k - K) \quad (2.32)$$

where the inner product between two general functions has the form

$$\langle f|g\rangle = \int_a^b dx f^*(x)g(x) \quad (2.33)$$

Notice the inner product has an integral over x and not k . For the Dirac delta normalization, the integral will generally have at least one integration limit of infinity. The k values serve as indices to distinguish the functions.

A general vector $|f\rangle$ can be written as a summation of basis functions. However, the expansion uses an integral rather than a discrete summation since there are more basis vectors in the continuous basis set than a conventional summation can handle.

$$|f\rangle = \int_a^b dk c_k |\phi_k\rangle \quad (2.34a)$$

The subscript on the coefficient c resembles the index used in the summation over discrete sets. As discussed later, the expansion coefficients c_k can be written as a function $c_k = c(k)$ and can be viewed as the components of the vector or as the transform of the function f with respect to the particular continuous basis (such as the Fourier transform). Figure 2.11 shows the function $|f\rangle$ projected onto two of the many basis vectors. If desired the coordinate projection operator $\langle x|$ can be applied to both sides of Equation 2.34a to obtain

$$f(x) = \int_a^b dk c_k \phi_k(x) \quad (2.34b)$$

The quantities c_k and ϕ_k can also be written in functional form as $c_k = c(k)$ and $\phi_k(x) = \phi(x, k)$.

Continuing to work with Equation 2.34a, the component c_K can be found by operating on the left with $\langle\phi_K|$ (note the index of capital K) and then using the orthonormality relation to get

$$\langle\phi_K|f\rangle = \int_a^b dk c_k \langle\phi_K|\phi_k\rangle = \int_a^b dk c_k \delta(k - K) = c_K \quad (2.35)$$

which assumes that $K \in (a, b)$. The operator $\langle\phi_K|$ was moved under the integral since the integral is over k and not K . Notice that when computing inner products such as $\langle\phi_K|\phi_k\rangle$, the integral runs over a spatial coordinate x and has the following form by definition of the inner product between functions.

$$\langle\phi_K|\phi_k\rangle = \int dx \phi_K^*(x) \phi_k(x) = \delta(k - K)$$

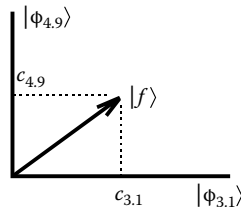


FIGURE 2.11 A function projected onto two of the many basis vectors.

This section will later show how the closure relation for coordinate space also produces this last result.

The closure relation can be found by using $c_k = \langle \phi_k | f \rangle$ as follows

$$|f\rangle = \int dk c_k |\phi_k\rangle = \int dk \langle \phi_k | f \rangle |\phi_k\rangle = \int dk |\phi_k\rangle \langle \phi_k | f \rangle$$

where $\langle \phi_k | f \rangle$ is just a complex number and can be moved behind the vector $|\phi_k\rangle$ without violating any rules. This last relation holds for arbitrary functions $|f\rangle$ in the Hilbert space so that

$$\int dk |\phi_k\rangle \langle \phi_k| = \hat{1} \quad (2.36)$$

by definition of operator equality. Two operators $\hat{A}, \hat{B}$ are equal if they map each vector $|v\rangle$ in the space in an identical manner, that is, $\hat{A}|v\rangle = \hat{B}|v\rangle$ for all $|v\rangle$ in V . Equation 2.36 provides the closure relation for a continuous set of basis vectors.

The closure relation is equivalent to a Dirac delta function. Operating on Equation 2.36 with $|x'\rangle$ and $\langle x|$ produces the desired relation.

$$\delta(x - x') = \langle x | x' \rangle = \langle x | \hat{1} | x' \rangle = \langle x | \left\{ \int dk |\phi_k\rangle \langle \phi_k| \right\} | x' \rangle = \int dk \phi_k^*(x') \phi_k(x)$$

2.5.2 COORDINATE SPACE

What does it mean to project a function f into coordinate space to find an inner product $\langle x | f \rangle$? We already know that functions $f \sim |f\rangle$ can be projected into function space (i.e., Hilbert space) to form inner products between functions such as $\langle f | g \rangle$. The coordinate basis set $\{|\xi\rangle\}$ really consists of a set of Dirac delta functions $\{|\xi\rangle \equiv |\delta(x - \xi)\rangle \sim \delta(x - \xi)\}$ as suggested by Figure 2.12. The coordinate ket $|x_0\rangle$ in the set $\{|\xi\rangle\}$ has the meaning of $|x_0\rangle \sim \delta(x - x_0)$ which essentially is a function with infinite “weight” at the single point x_0 . The bra $\langle x_0| \equiv \langle \delta(x - x_0)|$ is a projection operator that projects a function $|f\rangle$ onto the Dirac delta function $\delta(x - x_0)$. The projection of $f(x)$ onto the coordinate x_0 becomes

$$\langle x_0 | f \rangle = \langle \delta(x - x_0) | f(x) \rangle = \int_{-\infty}^{\infty} dx \delta(x - x_0) f(x) = f(x_0) \quad (2.37)$$

The bra $\langle x_0|$ essentially selects (i.e., projects out) the value of f at the particular single coordinate x_0 .

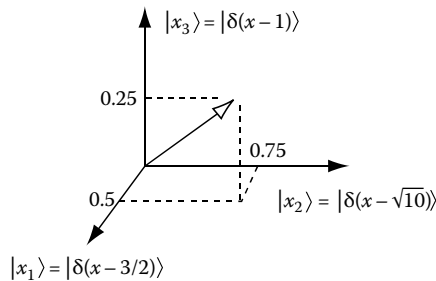


FIGURE 2.12 The coordinate space basis vectors are actually the Dirac delta functions.

We can demonstrate the orthonormality relation for the coordinate space. Let $|\xi\rangle$ and $|\eta\rangle$ be two of the uncountable many coordinate kets. Using Equation 2.33 for the inner product, we can write

$$\langle \xi | \eta \rangle = \langle \delta(x - \xi) | \delta(x - \eta) \rangle = \int_{-\infty}^{\infty} dx \delta(x - \xi) \delta(x - \eta) = \delta(\xi - \eta) \quad (2.38)$$

Therefore rather than have an orthonormality relation involving the Kronecker delta function as for Euclidean vectors, we see that the coordinate space uses the Dirac delta function.

Basis sets need to be complete in the sense that any function can be expanded in the set. Let f be an arbitrary element in the function space and consider its expansion in the coordinate basis set.

$$|f\rangle = \int dx' |x'\rangle g(x')$$

Here $g(x')$ appears as the component of a vector! If this represents a legitimate expansion of $f(x)$ then we should be able to show that $g(x)$ equals $f(x)$. To this end, operate on this last equation with $\langle x|$ to find

$$f(x) = \langle x | f \rangle = \int dx' \langle x | x' \rangle g(x') = \int dx' \delta(x' - x) g(x') = g(x)$$

So now we can think of the decomposition of a vector $\vec{f} = |f\rangle$ either in a function basis (Equations 2.34a and b) or a “coordinate” basis. Actually, both types of decomposition are in terms of functions except the “coordinate” basis uses Dirac delta functions.

Next, let us examine the closure relation for coordinate space. Table 2.1 shows how to replace the indices for the Euclidean vector and the summation by the coordinate x and integral, respectively.

$$\begin{aligned} \sum_{i=1}^n |i\rangle \langle i| &= 1 \quad \rightarrow \quad \int |x\rangle dx \langle x| = 1 \\ \langle m | n \rangle &= \delta_{mn} \quad \rightarrow \quad \langle x' | x \rangle = \delta(x - x') \\ m, n &\in \text{integers} \quad \quad \quad x, x' \in \mathbb{R} \end{aligned}$$

Note that the Dirac delta function replaces the Kronecker delta function for the continuous basis set $\{|x\rangle\}$. Also notice that an integral replaces the discrete summation for the continuous basis.

Let us demonstrate the closure relation for the coordinate basis set. First consider the inner product between any two elements of the Hilbert space using the basic definition of inner product from Section 2.1 as the first step.

$$\begin{aligned} \langle f | g \rangle &= \int dx f^*(x) g(x) = \int dx \langle x | f \rangle^+ \langle x | g \rangle = \int dx \langle f | x \rangle \langle x | g \rangle \\ &= \langle f | \left(\int |x\rangle dx \langle x| \right) | g \rangle \end{aligned} \quad (2.39a)$$

However, the unit operator $\hat{1}$ does not change the vector $|g\rangle$, that is $\hat{1}|g\rangle = |g\rangle$, so that the inner product can be also written as

$$\langle f | g \rangle = \langle f | \hat{1} | g \rangle \quad (2.39b)$$

TABLE 2.1
Summary of Results

	Euclidean Vectors	Functions—Discrete Basis	Functions—Continuous Basis
Basis	$\{ n\rangle: n = 1, 2, 3, \dots\} \sim \{\tilde{x}, \tilde{y}, \tilde{z}, \dots\}, n = \text{integer}$	$\{ n\rangle = u_n\rangle \sim u_n(x)\}, n = \text{integer}$	$\{ k\rangle = \phi_k\rangle \sim \phi_k(x)\}, k = \text{real}$
Projector	$\langle w = \tilde{w}^\dagger \circ$	$\langle f = \int dx f^*(x) \circ$	$\langle f = \int dx f^*(x) \circ$
Orthonormality	$\langle m n\rangle = \delta_{m,n}$	$\langle u_m u_n\rangle = \delta_{mn}$	$\langle \phi_k \phi_k\rangle = \delta(k - K)$
Complete	$ v\rangle = \sum_n c_n n\rangle$	$ f\rangle = \sum_n c_n u_n\rangle$	$ f\rangle = \int dk c_k \phi_k\rangle$
		$f(x) = \sum_n c_n u_n(x)$	$f(x) = \int dk c_k \phi_k(x)$
Components	$c_n = \langle n v\rangle$	$c_n = \langle u_n f\rangle$	$c_k = \langle \phi_k f\rangle$
Inner product	$\langle v w\rangle = \sum_n v_n^* w_n$	$\langle f g\rangle = \int dx f^*(x)g(x)$	$\langle f g\rangle = \int dx f^*(x)g(x)$
Closure	$\sum_n n\rangle \langle n = \hat{1}$	$\sum_n u_n\rangle \langle u_n = \hat{1}$	$\int dk \phi_k\rangle \langle \phi_k = \hat{1}$
		$\delta(x - x') = \sum_n u_n^*(x')u_n(x)$	$\delta(x - x') = \int dk \phi_k^*(x')\phi_k(x)$

Comparing the last two relations (Equations 2.39a and b) shows

$$\langle f | \hat{1} | g \rangle = \langle f | \left\{ \int |x\rangle dx \langle x| \right\} | g \rangle$$

This last relation must hold for all vectors $|f\rangle$ and $|g\rangle$ and therefore the operators on either side must be the same

$$\int |x\rangle dx \langle x| = \hat{1} \quad (2.40)$$

Example 2.21

Consistent notation

$$1 = \int |x\rangle dx \langle x|$$

Operate on the left with the bra $\langle x'|$ and on the right by a function $|f\rangle$ to get

$$\begin{aligned} \langle x' | f \rangle &= \langle x' | 1 | f \rangle = \langle x' | \int |x\rangle dx \langle x| | f \rangle \\ &= \int \langle x' | x \rangle dx \langle x | f \rangle = \int \delta(x - x') f(x) dx \\ &= f(x') \end{aligned}$$

which shows that the notation is consistent.

2.5.3 REPRESENTATIONS OF THE DIRAC DELTA USING BASIS VECTORS

Different sets of basis function lead to different representations of the Dirac delta function. First, consider a function space with a countable number of basis functions $\{\phi_i(x)\}$. Use the definition of inner product between coordinate kets and the definition of the unit operator to find

$$\delta(x - x') = \langle x | x' \rangle = \langle x | \hat{1} | x' \rangle$$

Next insert the closure relation in terms of the basis functions $\{\phi_i(x)\}$ and distribute the kets into the summation.

$$\delta(x - x') = \langle x | \left[\sum_{i=0}^{\infty} |\phi_i\rangle \langle \phi_i| \right] | x' \rangle = \sum_{i=0}^{\infty} \langle x | \phi_i \rangle \langle \phi_i | x' \rangle$$

Finally use the adjoint of the inner product $\langle \phi_i | x' \rangle = \langle x' | \phi_i \rangle^+ = \langle x' | \phi_i \rangle^* = \phi_i^*(x')$

$$\delta(x - x') = \sum_{i=0}^{\infty} \phi_i^*(x') \phi_i(x) \quad (2.41a)$$

The relation shows that any complete orthonormal set of functions gives a representation of the Dirac delta function. Therefore, different basis sets give different representations of the Dirac delta function. Section 2.7 shows that a basis set of sines produces a representation as does the basis set of Cosines. Different sets but the same Dirac delta function.

A similar set of manipulations hold for a continuous set of basis function $\{|\phi_k\rangle\}$

$$\delta(x - x') = \langle x|x'\rangle = \langle x|\hat{1}|x'\rangle = \delta(x - x') = \langle x|\left[\int |\phi_k\rangle d\langle\phi_k|\right]|x'\rangle$$

Distributing the kets under the integral then produces the desired results.

$$\delta(x - x') = \int \langle x|\phi_k\rangle d\langle\phi_k|x'\rangle = \int dk \phi_k^*(x') \phi_k(x) \quad (2.41b)$$

2.6 GRAHAM–SCHMIDT ORTHONORMALIZATION PROCEDURE

The Graham–Schmidt orthonormalization procedure transforms two or more independent vectors into two or more orthogonal vectors. The Graham–Schmidt procedure starts with a vector space and then develops a basis set. The opposite but usual approach starts with a basis set to determine the vector space (by taking all linear combinations of the basis elements). The present section uses the slightly more complicated case of functions and leaves the Euclidean vectors for the exercises.

2.6.1 SIMPLEST CASE OF TWO VECTORS

Let two functions be represented as vectors $|f\rangle$ and $|g\rangle$ in a Hilbert space H . The set of independent functions $\{|f\rangle, |g\rangle\}$ spans a 2-D subspace of the full space H . We wish to generate a basis set $\{|\phi_1\rangle, |\phi_2\rangle\}$ for this 2-D vector space.

The procedure starts by choosing the first basis vector to be parallel to either f or g . The choice does not matter so choose g for example. Then normalizing g provides

$$|\phi_1\rangle = |g\rangle / \|g\| \quad (2.42a)$$

A second basis vector $|\phi_2\rangle$ must exist since the set $\{|f\rangle, |g\rangle\}$ has two independent functions that necessarily span a 2-D subspace. Let $|h\rangle$ represent a function orthogonal to $|\phi_1\rangle$ or equivalently, orthogonal to $|g\rangle$ (see Figure 2.13), such that

$$|f\rangle = |h\rangle + c_1 |\phi_1\rangle \quad (2.42b)$$

Operating with $\langle\phi_1|$ on both sides of the equation for f , we find an expression for the component c_1

$$\langle\phi_1|f\rangle = \langle\phi_1|h\rangle + c_1 \langle\phi_1|\phi_1\rangle = c_1$$

where we have used the orthogonality of ϕ_1 and h , namely $\langle\phi_1|h\rangle = 0$, and the fact that ϕ_1 is normalized to 1. Now Equation 2.42b for f can be rewritten as

$$|h\rangle = |f\rangle - c_1 |\phi_1\rangle = |f\rangle - |\phi_1\rangle \langle\phi_1|f\rangle \quad (2.43a)$$

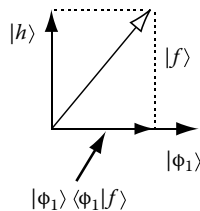


FIGURE 2.13 The relation between $|f\rangle, |\phi_1\rangle, |h\rangle$.

The usual form of the function $|h\rangle$, which is $h(x)$, can be recovered by operating on Equation 2.43a with $\langle x|$ to find

$$h(x) = f(x) - \phi_1(x)\langle\phi_1|f\rangle \quad (2.43b)$$

which can also be written as

$$h(x) = f(x) - \phi_1(x) \int_a^b dx \phi_1^*(x) f(x) \quad (2.43c)$$

We can easily prove that h and ϕ_1 are orthogonal by using Equation 2.43a and operating with $\langle\phi_1|$ as follows

$$\langle\phi_1|h\rangle = \langle\phi_1|\{f\} - |\phi_1\rangle\langle\phi_1|f\rangle\rangle = \langle\phi_1|f\rangle - \langle\phi_1|\phi_1\rangle\langle\phi_1|f\rangle = 0$$

as required. In order for the set $\{|h\rangle, |\phi_1\rangle\}$ to be orthonormal, we need to normalize the function $|h\rangle$. That is, we find the second basis vector $|\phi_2\rangle$ as

$$\phi_2(x) = \frac{h(x)}{\|h(x)\|} \quad (2.44)$$

The two functions ϕ_2 and h are similar much like $2\hat{x}$ and $\hat{x}$ are considered to be similar. We can see the function ϕ_2 has unit length by calculating the inner product

$$\langle\phi_2|\phi_2\rangle = \left\langle \frac{h}{\|h\|} \left| \frac{h}{\|h\|} \right. \right\rangle = \frac{1}{\|h\|^2} \langle h|h \rangle = \frac{\|h\|^2}{\|h\|^2} = 1$$

2.6.2 MORE THAN TWO VECTORS

We can easily include three or more vectors in the initial set. Consider the case of three vectors. Assume that the Graham–Schmidt procedure has been used to make two of the vectors ϕ_1, ϕ_2 orthonormal and that the third function f in the set $\{\phi_1, \phi_2, f\}$ Assume f to be independent of ϕ_1, ϕ_2 . There must be a third function $h(x)$ orthogonal to ϕ_1, ϕ_2 in order for the set $\{\phi_1, \phi_2, f\}$ to be independent. Therefore, set $|f\rangle = |h\rangle + c_1|\phi_1\rangle + c_2|\phi_2\rangle$. The constants c_1 and c_2 are found similar to above. We can write

$$|h\rangle = |f\rangle - |\phi_1\rangle\langle\phi_1|f\rangle - |\phi_2\rangle\langle\phi_2|f\rangle \quad (2.45)$$

Therefore the function $h(x)$ can be found by projecting this last equation into coordinate space. The function h must be normalized to unity in order to serve as a basis function.

$$\phi_3 = h/\|h\| \quad (2.46)$$

This procedure can be generalized to a set of arbitrarily many linearly independent functions from which we can find a basis set for the space.

2.7 FOURIER BASIS SETS

The Fourier series and Fourier transforms provide important applications of the generalized summations over basis vectors. The Fourier series uses a summation over a discrete collection of

basis functions consisting of sines and cosines for function space. This Hilbert space consists of bounded, piecewise continuous, and periodic functions. The sine portion of the series describes a subspace of odd functions while the cosine portion describes a subspace of even functions. Sections 2.7.1 and 2.7.2 describe the Fourier cosine and sine series as distinct from the full Fourier series.

The Fourier transform appears in many elementary studies in optics and electronics. The Fourier transform provides the decomposition of nonperiodic functions into a continuous basis set of complex exponentials.

2.7.1 FOURIER COSINE SERIES

The set of functions

$$B_c = \left\{ \frac{1}{\sqrt{L}}, \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi x}{L}\right), \dots \text{ for } n = 1, 2, 3, \dots \right\} = \{\phi_0, \phi_1, \dots\}$$

is orthonormal on the interval $x \in (0, L)$. The functions in B_c form a basis set for piecewise continuous functions on $(0, L)$. The function space can be enlarged to include functions that repeat every $2L$ along the entire x -axis; however, there are restrictions for the range $(L, 2L)$ (see below Section 2.8 and the chapter review problems). An arbitrary function $f \in Sp(B_c)$ can be written as a summation

$$|f\rangle = \sum_{n=0}^{\infty} c_n |\phi_n\rangle \quad (2.47a)$$

Operating on both sides with $\langle x|$ provides

$$f(x) = \frac{c_0}{\sqrt{L}} + \sum c_n \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi x}{L}\right) \quad (2.47b)$$

The normalization $\sqrt{2/L}$ depends on the interval endpoint L in $(0, L)$ and also upon the fact that the $n\pi x/L$ occurs as the argument of the cosine function with n being an integer.

The expansion coefficients $c_0, c_1, \dots$ (i.e., the components of the vector) in Equation 2.47 can be found from the inner product of f with each of the basis vectors $\cos(n\pi x/L)$

$$c_0 = \langle \phi_0 | f \rangle = \left\langle \frac{1}{\sqrt{L}} \middle| f(x) \right\rangle = \frac{1}{\sqrt{L}} \int_0^L dx f(x) \quad (2.48)$$

and

$$c_n = \langle \phi_n | f \rangle = \left\langle \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi x}{L}\right) \middle| f(x) \right\rangle = \sqrt{\frac{2}{L}} \int_0^L dx f(x) \cos\left(\frac{n\pi x}{L}\right) \quad (2.49)$$

where this expression for c_n holds for $n > 0$.

Example 2.22

Show that the cosine basis vectors are correctly normalized.

$$\chi_n(x) = \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi x}{L}\right)$$

Calculate the inner product

$$\|\phi_n\|^2 = \langle \phi_n | \phi_n \rangle = \int_0^L dx \phi_n(x) \phi_n(x) = \frac{2}{L} \int_0^L dx \cos^2 \left(\frac{n\pi x}{L} \right)$$

The last integral can be rewritten using the trigonometric identity

$$\cos^2 \theta = [\cos(2\theta) + 1]/2$$

so that

$$\|\phi_n\|^2 = \frac{1}{L} \int_0^L dx \left[\cos \left(\frac{2n\pi x}{L} \right) + 1 \right] = \frac{1}{L} \left[\frac{L}{2n\pi} \sin \left(\frac{2n\pi x}{L} \right) + x \right]_0^L = 1$$

2.7.2 FOURIER SINE SERIES

The sine functions provide another basis set for functions defined on the interval $x \in (0, L)$

$$B_s = \left\{ \sqrt{\frac{2}{L}} \sin \left(\frac{n\pi x}{L} \right) \quad n = 1, 2, 3, \dots \right\} = \{ \psi_n(x) : n = 1, 2, 3, \dots \}$$

The function space can be enlarged to include functions that repeat every $2L$ along the entire x -axis; however, there are restrictions for the range $(L, 2L)$ (refer to the chapter review problems). The normalization of $\sqrt{2/L}$ depends on the width of the interval L and on the fact that the sine function has $n\pi x/L$ in the argument (where n is an integer).

A function in the vector space spanned by B_s can be written as a summation over the basis vectors

$$|f\rangle = \sum_{m=1}^{\infty} c_m |\psi_m\rangle \quad \text{or} \quad f(x) = \sum_{n=1}^{\infty} c_n \sqrt{\frac{2}{L}} \sin \left(\frac{n\pi x}{L} \right) \quad (2.50)$$

The expansion coefficients are found by projecting the function onto the basis vectors

$$\langle \psi_n | f \rangle = \langle \psi_n | \left\{ \sum_m c_m |\psi_m\rangle \right\} = \sum_m c_m \langle \psi_n | \psi_m \rangle = c_n$$

These components can be evaluated

$$c_n = \langle \psi_n | f \rangle = \left\langle \sqrt{\frac{2}{L}} \sin \left(\frac{n\pi x}{L} \right) \middle| f(x) \right\rangle = \sqrt{\frac{2}{L}} \int_0^L dx f(x) \sin \left(\frac{n\pi x}{L} \right) \quad (2.51)$$

Example 2.23

Show that the set

$$B_s = \{ \psi_n(x) : n = 1, 2, 3, \dots \} = \left\{ \sqrt{\frac{2}{L}} \sin \left(\frac{n\pi x}{L} \right) \quad n = 1, 2, 3, \dots \right\}$$

is orthonormal on $0 < x < L$. The typical inner product looks like (changing variables to $y = \pi x/L$)

$$\langle \psi_n | \psi_m \rangle = \frac{2}{L} \int_0^L dx \sin\left(\frac{n\pi x}{L}\right) \sin\left(\frac{m\pi x}{L}\right) = \frac{2}{L} \frac{L}{\pi} \int_0^\pi dy \sin(ny) \sin(my) = \frac{2}{\pi} \int_0^\pi dy \sin(ny) \sin(my)$$

The integrals are easy to evaluate by recalling a couple of trigonometric identities

$$\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta \quad (2.52a)$$

$$\sin(\alpha + \beta) = \sin \alpha \cos \beta + \cos \alpha \sin \beta \quad (2.52b)$$

which can be combined to give some expressions useful to help demonstrate the orthonormality relations

$$\sin[(n+m)y] + \sin[(n-m)y] = 2 \sin(ny) \cos(my) \quad (2.53a)$$

$$\cos[(n+m)y] + \cos[(n-m)y] = 2 \cos(ny) \cos(my) \quad (2.53b)$$

$$\cos[(n-m)y] - \cos[(n+m)y] = 2 \sin(ny) \sin(my) \quad (2.53c)$$

The inner products are

$$\langle \psi_n | \psi_m \rangle = \frac{2}{\pi} \int_0^\pi dy \sin(ny) \sin(my) = \frac{1}{\pi} \int_0^\pi dy \{ \cos[(n-m)y] - \cos[(n+m)y] \}$$

The vectors are normalized to one as can be seen ($m=n$)

$$\|\psi_n\|^2 = \langle \psi_n | \psi_n \rangle = \frac{1}{\pi} \int_0^\pi dy \{ 1 - \cos(2ny) \} = \frac{1}{\pi} \left(y - \frac{\sin(2ny)}{2n} \right) \Big|_0^\pi = 1$$

Distinct vectors $n \neq m$ are orthogonal

$$\langle \psi_n | \psi_m \rangle = \frac{1}{\pi} \int_0^\pi dy \{ \cos[(n-m)y] - \cos[(n+m)y] \} = \frac{\sin[(n-m)y]}{(n-m)\pi} \Big|_0^\pi - \frac{\sin[(n+m)y]}{(n+m)\pi} \Big|_0^\pi = 0$$

2.7.3 FOURIER SERIES

The basis functions for this vector space are

$$B_F = \left\{ \frac{1}{\sqrt{2L}}, \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi x}{L}\right), \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi x}{L}\right) : n = 1, 2, 3, \dots \right\} = \{ |C_n\rangle, |S_n\rangle \}$$

where $x \in (-L, +L)$. The basis functions can be renamed in abbreviated form as

$$C_0(x) = \frac{1}{\sqrt{2L}} \quad C_n(x) = \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi x}{L}\right)$$

$$S_n(x) = \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi x}{L}\right) \quad n = 1, 2, \dots$$

The Fourier series for a function $|f\rangle$ is defined as

$$|f\rangle = \sum_{n=0}^{\infty} \alpha_n |C_n\rangle + \sum_{n=1}^{\infty} \beta_n |S_n\rangle \quad (2.54)$$

or equivalently, by operating with $\langle x|$

$$f(x) = \alpha_0 \frac{1}{\sqrt{2L}} + \sum_{n=1}^{\infty} \alpha_n \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} \beta_n \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi x}{L}\right) \quad (2.55)$$

Sometimes people write

$$\langle x|C_n\rangle = \cos\left(\frac{n\pi x}{L}\right) \quad \text{or} \quad |C_n\rangle = \left|\cos\left(\frac{n\pi x}{L}\right)\right\rangle$$

which abuses the Dirac notation (but it gets abused all the time anyway). The abused form $|C_n\rangle = |\cos(n\pi x/L)\rangle$ helps keep track of the variable x . Notice that the functions $f(x)$ will repeat every $2L$.

If we know the expansion coefficients α_n, β_n in Equation 2.48, then we know the function $f(x)$. However in most cases, we initially know the function $f(x)$ and we must determine the expansion coefficients. The expansion coefficients (i.e., components of the vector) α_n, β_n in Equations 2.54 and 2.55 can be determined using the basis set B_F . For the functions in B_F to be orthonormal, we must have

$$\langle C_n|C_m\rangle = \delta_{nm} \quad \langle S_n|S_m\rangle = \delta_{nm} \quad \langle C_n|S_m\rangle = 0$$

To find the expansion coefficients, start with Equation 2.54

$$|f\rangle = \sum_{n=0}^{\infty} \alpha_n |C_n\rangle + \sum_{n=1}^{\infty} \beta_n |S_n\rangle$$

Operating with $\langle C_m|$ yields

$$\langle C_m|f\rangle = \sum_{n=0}^{\infty} \alpha_n \langle C_m|C_n\rangle + \sum_{n=1}^{\infty} \beta_n \langle C_m|S_n\rangle = \sum_{n=0}^{\infty} \alpha_n \delta_{mn} = \alpha_m$$

Consequently, the expansion coefficients can be written as integrals

$$n = 0 \quad \alpha_0 = \langle C_0|f\rangle = \left\langle \frac{1}{\sqrt{2L}} \middle| f \right\rangle = \int_{-L}^L dx \frac{1}{\sqrt{2L}} f(x) \quad (2.56)$$

$$n > 0 \quad \alpha_n = \langle C_n|f\rangle = \left\langle \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi x}{L}\right) \middle| f \right\rangle = \int_{-L}^L dx \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi x}{L}\right) f(x) \quad (2.57)$$

Similarly, the β_n coefficients can be written as

$$n > 0 \quad \beta_n = \langle S_n|f\rangle = \left\langle \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi x}{L}\right) \middle| f \right\rangle = \int_{-L}^L dx \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi x}{L}\right) f(x) \quad (2.58)$$

2.7.4 ALTERNATE BASIS FOR THE FOURIER SERIES

For the Hilbert space of periodic, piecewise continuous functions on the interval $(-L, L)$, there exists an alternate set of basis functions as shown in the next paragraph.

$$B = \left\{ \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right) \quad n = 0, \pm 1, \pm 2, \dots \right\}$$

The orthonormality relation and the orthonormal expansion become

$$\left\langle \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right) \middle| \frac{1}{\sqrt{2L}} \exp\left(i \frac{m\pi x}{L}\right) \right\rangle = \delta_{nm}$$

and

$$f(x) = \sum_{n=-\infty}^{\infty} \frac{D_n}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right) \quad (2.59)$$

Notice how this expansion in terms of the complex exponential begins to look like a Fourier transform. The coefficients D_n can be complex.

The alternate basis set can be demonstrated by starting with Equation 2.55 and transforming it into Equation 2.59 as discussed in the Chapter 2 problems. The coefficients are related as follows.

$$D_n = \begin{cases} \alpha_0 & n = 0 \\ \frac{1}{\sqrt{2}}(\alpha_n - i\beta_n) & n = 1, 2, \dots \\ \frac{1}{\sqrt{2}}(\alpha_{-n} + i\beta_{-n}) & n = -1, -2, \dots \end{cases} \quad (2.60)$$

2.7.5 FOURIER TRANSFORM

The complete orthonormal basis set for a Hilbert space of bounded functions defined over the real x -axis is

$$\left\{ \frac{e^{ikx}}{\sqrt{2\pi}} \right\}.$$

Notice that the set can be indexed by either the continuous x or k variables. As a result, a generalized expansion can be made in either x or k such as

$$\int_{-\infty}^{\infty} dk \alpha(k) \frac{e^{ikx}}{\sqrt{2\pi}} \quad \text{or} \quad \int_{-\infty}^{\infty} dx \beta(x) \frac{e^{ikx}}{\sqrt{2\pi}}$$

The second integral is not a Fourier transform since a “minus” sign is missing from the exponent.

For α and β to be Fourier transform pairs, the x -integral must have a minus sign in the argument of the exponential as in e^{-ikx} . For this section, the generalized expansion will be defined as the integral over k .

$$f(x) = \int_{-\infty}^{\infty} dk \alpha(k) \frac{e^{ikx}}{\sqrt{2\pi}} \quad (2.61)$$

Define $\{|k\rangle\}$ to be the basis set

$$\left\{ |k\rangle = |\phi_k\rangle = \left| \frac{1}{\sqrt{2\pi}} e^{ikx} \right\rangle \rightarrow \phi_k(x) = \langle x|k\rangle = \frac{1}{\sqrt{2\pi}} \exp(ikx) \right\} \quad (2.62)$$

where k is real and “ $\circ$ ” provides a place for the variable x when the function is projected into coordinate space. We can demonstrate orthonormality for the basis set by substituting any two of the functions into the definition of the inner product.

$$\langle K|k\rangle = \int_{-\infty}^{\infty} dx \frac{e^{-iKx}}{\sqrt{2\pi}} \frac{e^{ikx}}{\sqrt{2\pi}} = \int_{-\infty}^{\infty} dx \frac{e^{i(k-K)x}}{2\pi} = \delta(k-K) \quad (2.63)$$

This expression for the Dirac delta can be found in Appendix B.

The closure relation

$$\hat{1} = \int_{-\infty}^{\infty} |k\rangle dk \langle k| \quad (2.64)$$

comes from the definition of completeness of the continuous basis set $\{|k\rangle = |\phi_k\rangle\}$. The projection of the closure relation into coordinate space and its dual produces a Dirac delta function. Operate on Equation 2.64 with $\langle x'|$ and $|x\rangle$ where x and x' represent spatial coordinates.

$$\langle x'|x\rangle = \langle x'| \left[\int dk |k\rangle \langle k| \right] |x\rangle = \int_{-\infty}^{\infty} \left\langle x' \left| \frac{1}{\sqrt{2\pi}} e^{iko} \right\rangle \left\langle \frac{1}{\sqrt{2\pi}} e^{iko} \right| x \right\rangle dk$$

This last expression can also be written as

$$\delta(x-x') = \int_{-\infty}^{\infty} dk \frac{e^{+ikx'}}{\sqrt{2\pi}} \frac{e^{-ikx}}{\sqrt{2\pi}} = \int_{-\infty}^{\infty} dk \frac{e^{-ik(x-x')}}{2\pi}$$

The Fourier series leads to the Fourier transform by starting with a function with period $2L$ and then allowing $L \rightarrow \infty$ (as shown in Appendix C). The generalized Fourier expansion of the function $f(x)$ must be written with an integral because of the continuous basis set

$$f(x) = \int_{-\infty}^{\infty} dk F(k) \frac{e^{ikx}}{\sqrt{2\pi}} \quad (2.65a)$$

Equation 2.65a is the “forward integral” or the “reverse transform.” The Fourier transform can be written as

$$F(k) = \int_{-\infty}^{\infty} dy \frac{e^{-iky}}{\sqrt{2\pi}} f(y) \quad (2.65b)$$

We discuss the basis set in the next subsection. People write $f(x)$ as the function and $f(k)$ as the Fourier transform. Notice that we use the same symbol f for both $f(x)$ and $f(k)$ since they are different representations of the same thing namely $|f\rangle$. Projecting $|f\rangle$ into coordinate space produces $\langle x|f\rangle = f(x)$. Projecting $|f\rangle$ into k -space produces the Fourier transform $\langle k|f\rangle = f(k)$.

Summary for Fourier Transform

Fourier Transform

$$\begin{aligned}
 f(k) &= \langle k|f\rangle = \langle k|1|f\rangle \\
 &= \langle k|\int dx|x\rangle\langle x|f\rangle \\
 &= \int dx\langle k|x\rangle f(x) \\
 &= \int dx\langle x|k\rangle^+ f(x) \\
 &= \int dx\left(\frac{e^{ikx}}{\sqrt{2\pi}}\right)^+ f(x) \\
 &= \int dx f(x) \frac{e^{-ikx}}{\sqrt{2\pi}}
 \end{aligned}$$

Inverse Transform

$$\begin{aligned}
 f(x) &= \langle x|f\rangle = \langle x|1|f\rangle \\
 &= \langle x|\int dk|k\rangle\langle k|f\rangle \\
 &= \int dk\langle x|k\rangle\langle k|f\rangle \\
 &= \int_{-\infty}^{\infty} dk f(k) \frac{e^{ikx}}{\sqrt{2\pi}}
 \end{aligned}$$

Example 2.24

Find the Fourier transform of

$$f(x) = \begin{cases} 1 & x \in [-L, L] \\ 0 & \text{elsewhere} \end{cases}$$

which represents an optical aperture.

The Fourier transform can be written as

$$f(k) = \int_{-\infty}^{\infty} dx f(x) \frac{e^{-ikx}}{\sqrt{2\pi}} = \frac{1}{ik\sqrt{2\pi}} [e^{ikL} - e^{-ikL}] = \sqrt{\frac{2}{\pi}} \frac{\sin kL}{k} \quad (2.66)$$

Notice that as the width of the aperture increases $L \rightarrow \infty$, the width of $f(k)$ decreases but its height increases. In fact, the representation of the Dirac delta function in Equation B.10 (Appendix B) has the form

$$\delta(x) = \lim_{\alpha \rightarrow \infty} \frac{\sin(\alpha x)}{\pi x}$$

Then Equation 2.66 gives

$$\lim_{L \rightarrow \infty} f(k) = \sqrt{2\pi} \lim_{L \rightarrow \infty} \frac{\sin(kL)}{\pi k} = \sqrt{2\pi} \delta(k)$$

So very wide optical apertures give Fourier transforms $f(k)$ that approximate a Dirac delta function.

2.8 CLOSURE RELATIONS, KRONECKER DELTA, AND DIRAC DELTA FUNCTIONS

Every basis set must span a vector space, must be complete and must give rise to a closure relation. Depending on whether the basis set is discrete or continuous, the closure relation produces either a

Kronecker delta or a Dirac delta function. Every function space produces a Dirac delta function and, in turn, every delta function can be expanded in any desired basis set. This fact becomes very useful for solving partial differential equations, for example, using the method of eigenfunction expansion. In such a case, the Green function can be easily found; the solutions for arbitrary forcing functions can be determined.

This section will demonstrate examples for Euclidean and function spaces. Special attention will be given for three types of Fourier series to illustrate how different basis sets produce delta functions and how the size of the domain affects the Dirac delta function.

2.8.1 ALTERNATE CLOSURE RELATIONS AND REPRESENTATIONS OF THE KRONECKER DELTA FUNCTION FOR EUCLIDEAN SPACE

Previous sections show that for a basis set

$$\{|1\rangle, |2\rangle, |3\rangle\} \quad (2.67)$$

a closure relation can be written

$$\hat{1} = \sum_{i=1}^3 |i\rangle\langle i| \quad (2.68)$$

Let $V_3 = Sp\{|1\rangle, |2\rangle, |3\rangle\}$ be the vector space spanned by the basis set. The closure relation (Equation 2.68) refers explicitly to this vector space. For example, if we add one more vector to the basis set in Equation 2.67

$$V_4 = Sp\{|1\rangle, |2\rangle, |3\rangle, |4\rangle\}$$

such that $V_3 \subseteq V_4$, then the closure relation in Equation 2.68 must be changed to include the new basis vector.

$$\hat{1} = \sum_{i=1}^4 |i\rangle\langle i|$$

Therefore, the definition of the unit operator in terms of a summation over basis vectors (i.e., basis vectors for the vector space and its dual) depends on the vector space. The exact meaning of the unit operator (i.e., expansion) depends on the particular vector space.

We can easily see that the representation of the Kronecker delta function depends on the particular vector space. In addition, given a particular vector space, we can see that changing basis within the space also affects the form of the Kronecker delta function. Figure 2.14 shows the

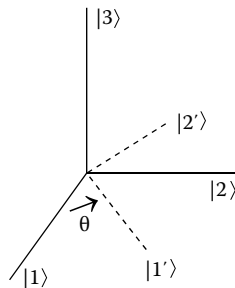


FIGURE 2.14 Rotated basis vectors.

vector space $V_3 = Sp\{|1\rangle, |2\rangle, |3\rangle\}$ with basis vectors rotated by an angle θ to produce $V_3 = Sp\{|1'\rangle, |2'\rangle, |3'\rangle\}$. Notice that the vector space does not change, but the basis vectors do. The new closure relation becomes

$$1 = \sum_{i=1}^3 |i'\rangle\langle i'|$$

where $|3'\rangle = |3\rangle$. Now operate with $\langle i|$ on the left and $|j\rangle$ on the right. The result can be written as

$$\delta_{ij} = \langle i|1|j\rangle = \langle i|1'\rangle\langle 1'|j\rangle + \langle i|2'\rangle\langle 2'|j\rangle + \langle i|3'\rangle\langle 3'|j\rangle \quad (2.69)$$

We could use the angle θ in Figure 2.14 to rewrite Equation 2.69 for specific i and j . The result gives a very common formula found in many texts but not so easily derived without the aid of the closure relation.

Example 2.25

If $i=j=1$ in Equation 2.69 then

$$\langle 1|1'\rangle = \cos(\theta) \quad \langle 1|2'\rangle = -\sin(\theta) \quad \langle 1|3'\rangle = 0$$

and so Equation 2.69 reduces to

$$1 = \cos^2 \theta + \sin^2 \theta$$

If $i=1$ and $j=2$ then using

$$\begin{aligned} \langle 1|1'\rangle &= \cos(\theta), & \langle 1|2'\rangle &= -\sin(\theta) \\ \langle 1'|2\rangle &= \sin(\theta), & \langle 2'|2\rangle &= \cos(\theta) \end{aligned}$$

Equation 2.69 becomes

$$0 = \delta_{12} = \cos(\theta) \sin(\theta) - \sin(\theta) \cos(\theta)$$

By including the third basis vector, more interesting relations can be determined (see Section 2.12).

2.8.2 COSINE BASIS FUNCTIONS

Consider the cosine basis functions defined in Section 2.7.1 with period $L = \pi$. The set of functions

$$B_c = \left\{ \frac{1}{\sqrt{\pi}}, \sqrt{\frac{2}{\pi}} \cos(nx), \dots \quad \text{for } n = 1, 2, 3, \dots \right\} = \{\phi_0, \phi_1, \dots\}$$

is orthonormal on the interval $x \in (0, \pi)$. The closure relation for the vector space $V = Sp B_c$ can be written as

$$1 = \sum_{n=0}^{\infty} |\phi_n\rangle\langle\phi_n| = \left| \frac{1}{\sqrt{\pi}} \right\rangle \left\langle \frac{1}{\sqrt{\pi}} \right| + \sum_{n=1}^{\infty} \left| \sqrt{\frac{2}{\pi}} \cos(n^\circ) \right\rangle \left\langle \sqrt{\frac{2}{\pi}} \cos(n^\circ) \right| \quad (2.70)$$

where the “o” reserves a location for the variable. The left side of this last equation produces the Dirac delta function

$$\delta(x - x') = \langle x | x' \rangle$$

for $x \in (0, \pi)$ by applying $\langle x |$ on the left side and $|x' \rangle$ on the right side of the unit operator 1. Therefore, Equation 2.70 produces the Dirac delta function

$$\delta(x - x') = \frac{1}{\pi} + \sum_{n=1}^{\infty} \frac{2}{\pi} \cos(nx) \cos(nx') \quad (2.71)$$

Or writing this as a limit

$$\delta(x - x') = \lim_{N \rightarrow \infty} \left[\frac{1}{\pi} + \sum_{n=1}^N \frac{2}{\pi} \cos(nx) \cos(nx') \right] \quad (2.72)$$

with the understanding that an integration operation must precede the limit operation. To check that the right hand side integrates to one, consider

$$\begin{aligned} \int_0^{\pi} dx \lim_{N \rightarrow \infty} \left[\frac{1}{\pi} + \sum_{n=1}^N \frac{2}{\pi} \cos(nx) \cos(nx') \right] &= \lim_{N \rightarrow \infty} \int_0^{\pi} dx \left[\frac{1}{\pi} + \sum_{n=1}^N \frac{2}{\pi} \cos(nx) \cos(nx') \right] \\ &= \lim_{N \rightarrow \infty} [1 + 0] = 1 \end{aligned}$$

where the integral $\int_0^{\pi} dx \cos(nx) = 0$ was used. Figure 2.15 shows two plots of Equation 2.72 corresponding to $N=10, 50$. Notice how the function sharpens for larger values of N . As an important note, the x -coordinates must be restricted to the range $(0, \pi)$ since the product of cosines in Equation 2.72 repeats every π . We would get multiple delta functions.

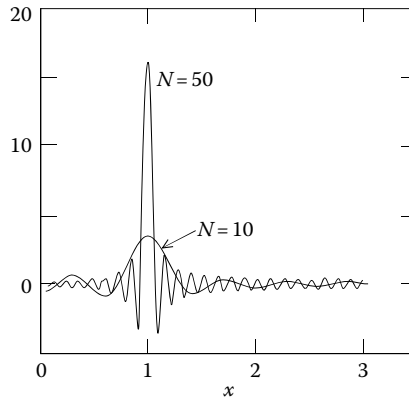


FIGURE 2.15 A representation of the Dirac delta function $\delta(x - 1)$ for the cosine basis vectors with x restricted to $(0, \pi)$. The plots are shown for two different values of N in Equation 2.71 and $x' = 1$.

2.8.3 SINE BASIS FUNCTIONS

The basis set $B_s = \left\{ \sqrt{\frac{2}{\pi}} \sin(nx) \mid n = 1, 2, 3, \dots \right\} = \{\psi_n(x) : n = 1, 2, 3, \dots\}$ is orthonormal on the interval $x \in (0, \pi)$. The closure relation for the vector space $V = \text{Sp } B_s$ can be written as

$$1 = \sum_{n=0}^{\infty} |\psi_n\rangle \langle \psi_n| = \sum_{n=1}^{\infty} \left| \sqrt{\frac{2}{\pi}} \sin(n^\circ) \right\rangle \left\langle \sqrt{\frac{2}{\pi}} \sin(n^\circ) \right| \quad (2.73)$$

where the “ $\circ$ ” reserves a location for the variable. The Dirac delta function

$$\delta(x - x') = \langle x | x' \rangle$$

for $x \in (0, \pi)$ comes from applying $\langle x |$ on the left side and $|x'\rangle$ on the right side of the unit operator in Equation 2.73.

$$\delta(x - x') = \sum_{n=1}^{\infty} \frac{2}{\pi} \sin(nx) \sin(nx') \quad (2.74)$$

Figure 2.16 shows a plots of Equation 2.74 corresponding to $N=20$. As an important note, the x -coordinates are restricted to the range $(0, \pi)$ since the product of sines in Equation 2.72 repeats every π .

2.8.4 FOURIER SERIES BASIS FUNCTIONS

Out of an infinite number of different basis sets, the Fourier series has two very popular ones. The first one for $x \in (0, \pi)$

$$B_F = \left\{ C_0 = \frac{1}{\sqrt{2\pi}}, \quad C_n = \frac{1}{\sqrt{\pi}} \cos(nx), \quad S_n = \frac{1}{\sqrt{\pi}} \sin(nx) : n = 1, 2, 3, \dots \right\}$$

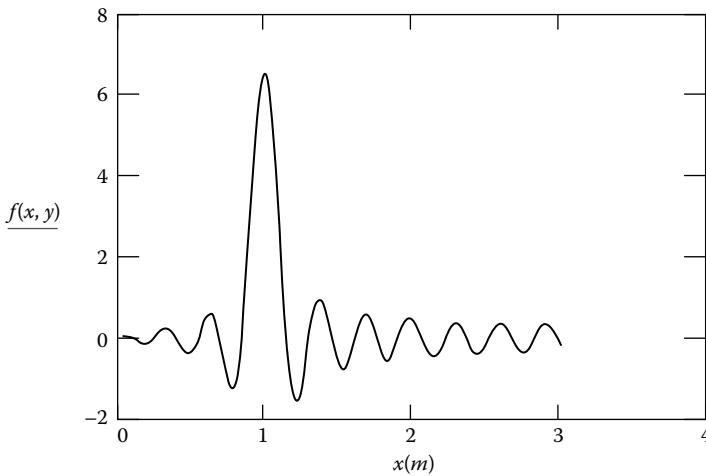


FIGURE 2.16 A sine representation ($N=20$) of the delta function $\delta(x - 1)$.

produces the closure relation

$$1 = \sum_{n=0}^{\infty} |C_n\rangle\langle C_n| + \sum_{n=1}^{\infty} |S_n\rangle\langle S_n| \quad (2.75)$$

The Dirac delta function

$$\delta(x - x') = \langle x|x' \rangle$$

for $x \in (0, \pi)$ comes from applying $\langle x|$ on the left side and $|x'\rangle$ on the right side of Equation 2.75. We find

$$\delta(x - x') = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{n=1}^{\infty} \cos(nx) \cos(ny) + \frac{1}{\pi} \sum_{n=1}^{\infty} \sin(nx) \sin(ny)$$

The alternate basis set in Section 2.7 is

$$B = \left\{ |\phi_n\rangle = \left| \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi^\circ}{L}\right) \right\rangle \quad n = 0, \pm 1, \pm 2, \dots \right\}$$

where again “ $\circ$ ” reserves a location for the variable. We can therefore write an alternate closure relation

$$1 = \frac{1}{2L} \left| \exp\left(i \frac{n\pi^\circ}{L}\right) \right\rangle \left\langle \exp\left(i \frac{n\pi^\circ}{L}\right) \right|$$

A representation of the Dirac delta function on $(-L, L)$ must be

$$\delta(x - x') = \frac{1}{2L} \sum_{n=-\infty}^{\infty} \exp\left[\frac{in\pi}{L}(x - x')\right]$$

where recall

$$\left\langle \exp\left(i \frac{n\pi^\circ}{L}\right) \middle| x' \right\rangle = \left\langle x' \middle| \exp\left(i \frac{n\pi^\circ}{L}\right) \right\rangle^+ = \left(\exp\left(i \frac{n\pi x'}{L}\right) \right)^+ = \exp\left(-i \frac{n\pi x'}{L}\right)$$

Appendix C shows that the new basis is essential for “generalizing” Fourier series to Fourier transforms.

2.8.5 SOME NOTES

1. Even discrete basis sets with Kronecker-delta orthonormalization can give Dirac delta functions when projecting the closure relation onto coordinate space. This occurs when the vector space consists of functions.
2. Dirac delta functions can provide some formulas. For example, we can show

$$1 = \sum_{n=1}^{\infty} \frac{2[1 - (-1)^n]}{n\pi} \sin \frac{n\pi x}{L} \quad \text{for } x \in (0, L)$$

The proof goes as follows.

$$\delta(x - \xi) = \sum_{n=1}^{\infty} \phi_n(x) \phi_n(\xi) = \sum_{n=1}^{\infty} \frac{2}{L} \sin \frac{n\pi x}{L} \sin \frac{n\pi \xi}{L}$$

since ϕ is real. Integrating this last equation over ξ from 0 to L provides

$$1 = - \sum_{n=1}^{\infty} \frac{2}{L} \frac{L}{n\pi} \sin \frac{n\pi x}{L} \sin \frac{n\pi \xi}{L} \Big|_{\xi=0}^L = \sum_{n=1}^{\infty} \frac{2}{n\pi} \sin \frac{n\pi x}{L} [1 - (-1)^n]$$

3. Dirac delta functions are important for solving partial differential equations with an impulse driving term

$$\hat{L}u(x, t) = \delta(t - t_0)$$

by the method of eigenfunction expansion. The Dirac delta function can be expanded in the basis set obtained from the boundary value problem with $\hat{L}u = 0$. It's fortunate that every function basis provides a Dirac delta function. Expand δ in the same basis set used to expand u . The rest of the eigenfunction expansion method is the same.

2.9 INTRODUCTION TO DIRECT PRODUCT SPACES

In quantum mechanics, one imagines that each particle inhabits its own vector space. For the translational coordinates, each particle would have a 3-D vectors space say V_1 . If one includes spin as a separate degree of freedom, then a single particle has mathematical representations in two vector spaces—call them V_1 and V_2 . A vector $|Y\rangle$ representing a single particle then consists of two parts placed side-by-side $|Y\rangle = |\phi\rangle|\psi\rangle = |\phi\psi\rangle$ where $|\phi\rangle$ and $|\psi\rangle$, respectively reside in V_1 and V_2 . The full vector $|Y\rangle$ necessarily decomposes into two parts since the vector $|Y\rangle$ represents the full particle having characteristics from two distinct spaces. The vector $|Y\rangle$ lives in the direct product space (sometimes also called a tensor product space). Similarly a vector representing two distinct particles will be represented as a direct product spaces (sometimes also termed tensor product spaces) product of vectors with one from the vector space for the first particle placed next to the one from the vector space for the second particle. One normally considers the vectors spaces to be separate independent spaces and represents the interaction between particles by an operator acting between vector spaces. Later chapters will clarify the dynamics involved. The direct product spaces (sometimes also termed tensor product spaces) product differs from the superposition. The superposition consists of vectors in the single space and represents the fact that a particle can simultaneously have characteristics corresponding to each vector in the summation. Direct product vectors can also be summed for the same reasons. However, the product occurs because a single vectors space representing some specific property of a particle (position for example) must be made larger to include other independent properties (such as particle spin).

2.9.1 OVERVIEW OF DIRECT PRODUCT SPACES

Mathematically, direct product spaces (sometimes also termed tensor product spaces) simply join two other spaces together. The two vector spaces can be quite dissimilar as would be the case, for example, with the Euclidean and function spaces. The procedure to produce direct product spaces will likely remind the reader of that for forming the Cartesian product using x -, y -coordinates. This section will cover many of the concepts familiar from our previous work on vector spaces while

subsequent section will develop an intuitive approach for functions using a multidimensional Fourier expansion.

Consider two vector spaces V and W . The direct product $\otimes$ of two vectors $|v\rangle \in V$ and $|w\rangle \in W$ is written as $|v\rangle \otimes |w\rangle$. Often for convenience, one omits the cross symbol to write

$$|v\rangle \otimes |w\rangle = |v\rangle |w\rangle = |vw\rangle \quad (2.76)$$

One must remember from which space each vector originates since an operator defined on W , such as $|w\rangle^+ = \langle w|$, never “sees” vectors in V . The collection of all direct product vectors forms the direct product space $V \otimes W$.

Suppose the two vector spaces V and W have respective (discrete) basis sets

$$B_v = \{|\phi_i\rangle\} \quad B_w = \{|\psi_j\rangle\} \quad (2.77)$$

The spaces V and W do not need to be the same size nor the same type. The product space has the basis set given by

$$\{|\phi_i\rangle \otimes |\psi_j\rangle = |\phi_i\rangle |\psi_j\rangle = |\phi_i, \psi_j\rangle\} \quad (2.78)$$

where obviously, the size of the direct product space $V \otimes W$ is given by

$$\dim[V \otimes W] = \dim(V) \dim(W)$$

One can picture the direct product space $V \otimes W$ as having axes (as usual) with each axis labeled by a different basis vector $|\phi_i, \psi_j\rangle$ (see for example, Figure 2.17). For simplicity, sometimes the basis vectors written as $|\phi_i, \psi_j\rangle = |ij\rangle$

A vector $|\gamma\rangle$ in the direct product space can be written as a summation over the basis set (Equation 2.78) as

$$|\gamma\rangle = \sum_{i,j} C_{i,j} |\phi_i \psi_j\rangle \quad (2.79)$$

For example, the vector $|\gamma\rangle$ in Figure 2.17 has the expansion $|\gamma\rangle = 1|\phi_1 \psi_1\rangle + 3|\phi_3 \psi_6\rangle + 4|\phi_5 \psi_1\rangle$, which represents a superposition of basis vectors. The reader should note that one can uniquely identify a vector $|\gamma\rangle = |v\rangle \otimes |w\rangle$ in direct product space if one knows the vectors $|v\rangle = \sum_i v_i |\phi_i\rangle$ and $|w\rangle = \sum_j w_j |\psi_j\rangle$

$$|\gamma\rangle = \sum_{i,j} \gamma_{i,j} |\phi_i\rangle |\psi_j\rangle \quad \text{with} \quad \gamma_{i,j} = v_i w_j \quad (2.80)$$

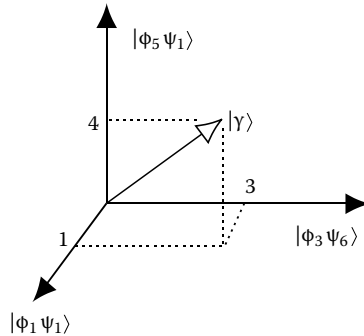


FIGURE 2.17 The decomposition of the vector $|\gamma\rangle$ in direct product space.

However, given the vector $|\gamma\rangle$ in direct product space, one cannot uniquely find vectors in V and W to give $|\gamma\rangle$. The reason is that the number $\gamma_{i,j}$ cannot be uniquely factored into v_i and w_j .

The adjoint operator “+” maps the vector (ket)

$$|v, w\rangle \in V \otimes W$$

into the projection operator (bra) as

$$|v, w\rangle^+ = \langle v, w| = \langle v|\langle w|$$

where $\langle v, w| \in [V \otimes W]^+ = V^+ \otimes W^+$. Given that the adjoint represents an isomorphism, the size of the original direct product space must be the same as that of the dual direct product space. As will become apparent, there is not much point in switching the order of the direct product vectors under the action of the adjoint. The basis set for the dual space is

$$\{\langle \phi_i, \psi_j| = \langle \phi_i|\langle \psi_j|\}$$

How are inner products formed? We must keep track of which dual space acts on which vector space. In particular, inner products can only be formed between V^+ and V , and between W^+ and W . Therefore if

$$|v_1\rangle, |v_2\rangle \in V \quad |w_1\rangle, |w_2\rangle \in W$$

the inner product satisfies

$$\langle v_1 w_1 | v_2 w_2 \rangle = \langle v_1 | v_2 \rangle \langle w_1 | w_2 \rangle \quad (2.81)$$

Of course, $\langle v_1 | v_2 \rangle$ and $\langle w_1 | w_2 \rangle$ are just complex numbers so that Equation 2.81 can also be written as

$$\langle v_1 w_1 | v_2 w_2 \rangle = \langle w_1 | w_2 \rangle \langle v_1 | v_2 \rangle$$

where the factors on the right hand side have been reversed.

The basis vectors are orthonormal in the sense

$$\langle \phi_a \psi_b | \phi_c \psi_d \rangle = \langle \phi_a | \phi_c \rangle \langle \psi_b | \psi_d \rangle = \delta_{ab} \delta_{cd} \quad (2.82)$$

where δ symbolizes the Kronecker delta function. Notice that the components $C_{a,b}$ in a superposition vector (Equation 2.79) can be found by projecting onto a basis vector such as $|\phi_a \psi_b\rangle$. The a, b component will be

$$\langle \phi_a \psi_b | \gamma \rangle = \langle \phi_a \psi_b | \sum_{i,j} C_{i,j} |\phi_i \psi_j\rangle = \sum_{i,j} C_{i,j} \langle \phi_a \psi_b | \phi_i \psi_j \rangle = C_{a,b} \quad (2.83)$$

The closure relation can now be determined by substituting the coefficients $C_{a,b}$ from Equation 2.83 back into Equation 2.79.

$$|\gamma\rangle = \sum_{i,j} C_{i,j} |\phi_i \psi_j\rangle = \sum_{i,j} \langle \phi_i \psi_j | \gamma \rangle |\phi_i \psi_j\rangle = \sum_{i,j} |\phi_i \psi_j\rangle \langle \phi_i \psi_j | \gamma \rangle \quad (2.84)$$

Comparing both sides shows that the resolution of unity must be given by

$$\hat{1} = \sum_{i,j} |\phi_i \psi_j\rangle \langle \phi_i \psi_j| \quad (2.85)$$

2.9.2 INTRODUCTION TO DYADIC NOTATION FOR THE TENSOR PRODUCT OF TWO EUCLIDEAN VECTORS

The previous section shows how to handle vector spaces with discrete basis sets which certainly includes the Euclidean vector spaces. However, dyadic and tensor notation sometimes appears in the literature and text books. The reader should note that the tensor product and direct product as defined here represent the same mathematical entity except we use the tensor product to refer to Euclidean vectors.

If V and W are spanned by the unit vectors in $\{\tilde{x}, \tilde{y}, \tilde{z}\}$ then the tensor product space will be spanned by $\{\tilde{x}\tilde{x}, \tilde{x}\tilde{y}, \tilde{x}\tilde{z}, \dots, \tilde{z}\tilde{z}\}$. A general vector γ in the space $V \otimes W$ can be written as $\gamma = \beta_{11}\tilde{x}\tilde{x} + \beta_{12}\tilde{x}\tilde{y} + \dots + \beta_{33}\tilde{z}\tilde{z}$ or, using $\tilde{e}_i$ ($i = 1, 2, 3$) for the unit vectors, $\gamma = \sum_{i,j} \beta_{i,j} \tilde{e}_i \tilde{e}_j$. The reader will note a similarity with the dyads discussed in Chapter 3. The vector γ is given the notation $\overleftrightarrow{\gamma}$ with the double arrow to show the vectors placed side by side in this case. The inner product requires two dot products to project out the components. Component a, b will be $\beta_{ab} = \tilde{e}_a \cdot \overleftrightarrow{\gamma} \cdot \tilde{e}_b$.

2.9.3 DIRECT PRODUCT SPACE FROM THE FOURIER SERIES

Up to this point in Chapter 2, we have dealt with functions having a single variable x such as $f(x) = \sum \beta_n \phi_n(x)$. The functions $\{\phi_n\}$ form a complete set and define a Hilbert space. Using the definition $\langle x|f \rangle = f(x)$, the expansion can also be written as $|f \rangle = \sum \beta_n |\phi_n \rangle = \sum \beta_n |n \rangle$.

What about functions such as $f(x, y)$? For example, we often solve partial differential equations for 2-D motion on a square rubber membrane (drum head) and find solutions of the form $f(x, y) = \sum_{n,m} \beta_{n,m} \sin(n\pi x/L) \sin(m\pi y/L)$. Now it is necessary to know the basis functions for the x -space and those for the y -space. In other words, f really consists of two separate Hilbert spaces. Let us assume that the functions $\{X_n(x)\}$ and $\{Y_m(y)\}$ are complete orthonormal sets for their respective spaces. Consider x fixed for just a moment; this means $X_n(x)$ must be a constant. For this given value of x , the function f can only be a function of y , say $f(x, y) = g(y)$. We can expand $g(y)$ in terms of the Y basis

$$g(y) = \sum_m \alpha_m Y_m(y) \quad (2.86)$$

Now, if x can take on other values, then clearly the components α_n must be functions of x since changing x on the left side of the equation

$$f(x, y) = \sum_m \alpha_m Y_m(y)$$

must produce changes on the right side. Given that the components depend on x

$$\alpha_m = \alpha_m(x)$$

they can be expanded in the basis set X_n

$$\alpha_m(x) = \sum_n \beta_{nm} X_n(x) \quad (2.87)$$

where β_{nm} must be constants independent of x, y . Combining the two summations in Equations 2.86 and 2.87 provides

$$f(x, y) = \sum_{nm} \beta_{nm} X_n(x) Y_m(y) \quad (2.88)$$

There are alternate ways to write the summation in Equation 2.88 by extending the Dirac notation a little bit. A function of two variables x, y can be written as

$$f(x, y) = \langle x, y | f \rangle$$

where the ket $|x, y\rangle = |xy\rangle$ is the coordinate ket. Similarly, one can write $\langle xy| = \langle x|\langle y|$. Technically, the 2-D coordinate ket represents two Dirac delta functions such as

$$|x'\rangle|y'\rangle = |\delta(x - x')\delta(y - y')\rangle$$

The expansion in Equation 2.88 can be written as

$$\langle xy | f \rangle = \sum_{nm} \beta_{nm} \langle x | X_n \rangle \langle y | Y_m \rangle \quad (2.89)$$

People use the shorthand notation $|X_n\rangle = |n\rangle$ and $|Y_m\rangle = |m\rangle$ keeping in mind that n refers to X and m refers to Y . Next $f(x, y)$ in Equation 2.89 can be written as

$$\langle xy | f \rangle = \langle x | \langle y | \sum_{nm} \beta_{nm} |X_n\rangle |Y_m\rangle \quad (2.90)$$

where now we must keep track that x goes with X_n and y goes with Y_m . Sometimes we track this order by the position of x, y in $\langle xy|$. Comparing both sides of Equation 2.90 shows

$$|f\rangle = \sum_{nm} \beta_{nm} |X_n\rangle |Y_m\rangle \quad (2.91)$$

as one alternative to Equation 2.88.

The reader should recognize the combination of kets in Equation 2.91 as vectors in the direct product space based on the discussion in Section 2.9.1. As with Euclidean vectors, the collection of all linear combinations of the direct product of basis vectors $|X_n Y_m\rangle$ forms the direct product space

$$V = V_x \otimes V_y = Sp\{|X_n\rangle |Y_m\rangle = |X_n Y_m\rangle\}$$

The combinations $|X_n\rangle |Y_m\rangle = |\phi_{nm}\rangle$ form the basis vectors of the product space and can be conveniently written as $|X_n\rangle |Y_m\rangle = |X_n Y_m\rangle = |nm\rangle$. A general function in the product space can be expanded as

$$|f\rangle = \sum_{nm} \beta_{nm} |\phi_{nm}\rangle = \sum_{nm} \beta_{nm} |X_n\rangle |Y_m\rangle = \sum_{nm} \beta_{nm} |nm\rangle \quad (2.92)$$

The combinations such as $\langle xy | nm \rangle$ can be written as

$$\langle xy | nm \rangle = \langle xy | X_n Y_m \rangle = \langle x | \langle y | |X_n\rangle |Y_m\rangle = \langle x | X_n \rangle \langle y | Y_m \rangle = X_n(x) Y_m(y)$$

The set $\{|\phi_{nm}\rangle = |nm\rangle = |X_n Y_m\rangle\}$ is orthonormal

$$\langle n' m' | nm \rangle = \delta_{n' n} \delta_{m' m}$$

as can easily be seen

$$\langle \phi_{n' m'} | \phi_{nm} \rangle = \langle X_{n'} Y_{m'} | X_n Y_m \rangle = \langle X_{n'} | X_n \rangle \langle Y_{m'} | Y_m \rangle = \delta_{n' n} \delta_{m' m} \quad (2.93)$$

2.9.4 COMPONENTS AND CLOSURE RELATION FOR THE DIRECT PRODUCT OF FUNCTIONS WITH DISCRETE BASIS SETS

If $|f\rangle$ is known, how do we write the components β_{nm} (Equation 2.91) in terms of X_n, Y_m ? The answer starts with the definition

$$|f\rangle = \sum_{nm} \beta_{nm} |X_n\rangle |Y_m\rangle$$

and uses the orthonormal properties of $\{X_n\}$ and $\{Y_m\}$. First, operate with $\langle Y_{m'}|$ on both sides

$$\langle Y_{m'}|f\rangle = \sum_{nm} \beta_{nm} |X_n\rangle \langle Y_{m'}|Y_m\rangle \quad (2.94)$$

Notice that the bras $\langle Y_{m'}|$ only operate on the Hilbert space spanned by $\{|Y_m\rangle\}$. Using the orthonormal relation for a discrete set of basis vectors

$$\langle Y_{m'}|Y_m\rangle = \delta_{m'm} \quad (2.95)$$

Therefore the summation in Equation 2.94 becomes

$$\langle Y_{m'}|f\rangle = \sum_n \beta_{nm'} |X_n\rangle \quad (2.96)$$

Because of $|X_n\rangle$ in this summation, the inner product $\langle Y_{m'}|f\rangle$ must be a function of x . In fact, define $|g\rangle = \langle Y_{m'}|f\rangle$ where g is a function of x . The function $g(x)$ can be written as

$$|g\rangle = \sum_n \beta_{nm'} |X_n\rangle$$

Now operate with $\langle X_{n'}|$ on both sides to get

$$\langle X_{n'}|g\rangle = \beta_{n'm'}$$

or

$$\beta_{nm} = \langle X_n|g\rangle = \{\langle X_n|\langle Y_m|\} |f\rangle = \langle X_n Y_m |f\rangle = \langle nm |f\rangle$$

which is the desired result. This result can also be written as an integral, where the domains for X_n and Y_m are assumed to be (a, b) and (c, d) , respectively.

$$\begin{aligned} \beta_{nm} &= \langle nm |f\rangle = \langle n | \langle m |f\rangle \\ &= \int_a^b dx X_n^*(x) \int_c^d dy Y_m^*(y) f(x, y) \\ &= \int_a^b dx \int_c^d dy X_n^*(x) Y_m^*(y) f(x, y) \end{aligned}$$

Notice the complex conjugation on X^* and Y^* .

Next, we demonstrate the closure relation for the basis vectors $|X_n Y_m\rangle$. Starting with the basic definition

$$|f\rangle = \sum_{nm} \beta_{nm} |X_n\rangle |Y_m\rangle$$

and substituting $\beta_{nm} = \langle nm|f\rangle$, yields

$$|f\rangle = \sum_{nm} \langle nm|f\rangle |nm\rangle = \sum_{nm} |nm\rangle \langle nm|f\rangle$$

Comparing both sides (i.e., treating $|f\rangle$ as the arbitrary vector that it is)

$$\underline{1} = \sum_{nm} |nm\rangle \langle nm| = \sum_{nm} |X_n\rangle |Y_m\rangle \langle Y_m| \langle X_n| \quad (2.97)$$

which is the closure relation for the basis vectors (a.k.a., the completeness relation). As usual, the closure relation is equivalent to a representation of the Dirac delta function.

$$\begin{aligned} \langle x'y'|xy\rangle &= \langle x'y'|1|xy\rangle = \sum_{nm} \langle x'y'|nm\rangle \langle nm|xy\rangle \\ &= \sum_{nm} [X_n^*(x')X_n(x)] [Y_m^*(y')Y_m(y)] = \sum_n X_n^*(x')X_n(x) \sum_m Y_m^*(y')Y_m(y) \\ &= \delta(x - x')\delta(y - y') \end{aligned}$$

2.9.5 NOTES ON THE DIRECT PRODUCTS OF CONTINUOUS BASIS SETS

By now, the reader realizes the case for the continuous basis functions can be found from that of the discrete ones simply by replacing summations with integrals and Kronecker delta functions with the Dirac delta functions. If the spaces V and W , respectively, are spanned by the continuous basis sets $\{\phi_k\}$ and $\{\psi_\kappa\}$ where k and κ have a continuous range of values, then the basis set for the direct product space will be $\{|\phi_k \psi_\kappa\rangle\}$. An arbitrary vector $|\gamma\rangle$ in the direct product space is given by

$$|\gamma\rangle = \iint dk d\kappa \beta_{k,\kappa} |\phi_k \psi_\kappa\rangle \quad (2.98a)$$

which should remind the reader of a 2-D Fourier transform. The components and closure relation are then

$$\beta_{k,\kappa} = \iint dk d\kappa \langle \phi_k \psi_\kappa | \gamma \rangle \quad (2.98b)$$

$$\hat{1} = \iint dk d\kappa |\phi_k \psi_\kappa\rangle \langle \phi_k \psi_\kappa| \quad (2.98c)$$

Similarly, one can see that the closure relation for the spatial coordinate kets $|xy\rangle$ is

$$\hat{1} = \int_a^b dx \int_c^d dy |xy\rangle \langle xy| \quad (2.99a)$$

where

$$\langle x'y'|xy\rangle = \delta(x - x')\delta(y - y') \quad (2.99b)$$

2.10 INTRODUCTION TO MINKOWSKI SPACE

The tensor notation commonly found with studies of special relativity provides a compact, simplifying notation in many fields of study. Of special importance for the present chapter, the infrastructure of special relativity incorporates Minkowski space that has a pseudo-inner product. The pseudo-inner product in this case does not satisfy all of the requirements for an inner product. In particular, it does not require a vector be zero when the inner product has a zero value. The special theory of relativity requires this inner product in order to ensure the speed of light remains independent of the translational motion of the observer.

2.10.1 COORDINATES AND PSEUDO-INNER PRODUCT

Minkowski space has four dimensions with coordinates (x_0, x_1, x_2, x_4) where for special relativity, the first coordinate is related to the time t . Rather than defining the inner product as $\langle v|w \rangle = \sum_n v_n w_n$, the inner product has the form

$$\langle v|w \rangle = v_0 w_0 - (v_1 w_1 + v_2 w_2 + v_3 w_3) \quad (2.100)$$

Based on this definition, the inner product for Minkowski space does not satisfy all the properties of the inner product. In particular, the pseudo-inner product in Equation 2.100 does not require the vectors v or w to be zero when the inner product has the value of zero.

The theory of relativity uses two types of notation. In the first, Minkowski 4-vectors use an imaginary number i to make the “inner product” appear similar to Euclidean inner products. In the second, a “metric” matrix is defined along with specialized notation. Additionally, a constant multiplies the time coordinate t in order to give it the same units as the spatial coordinates.

2.10.2 PSEUDO-ORTHOGONAL VECTOR NOTATION

One variant of the 4-vector notation uses an imaginary i with the time coordinate $x_\mu = (ict, x, y, z) = (ict, \vec{r})$. The constant c , the speed of light, converts the time t into a distance. The pseudo-inner product of the vector with itself then has the form

$$x_\mu x_\mu \equiv \sum_{\mu=1}^4 x_\mu x_\mu = (ict, \vec{r}) \cdot (ict, \vec{r}) = -c^2 t^2 + x^2 + y^2 + z^2 \quad (2.101)$$

The imaginary number $i = \sqrt{-1}$ makes the calculation of length look like Pythagorean’s theorem but produces the same result as for the pseudo-inner product in Equation 2.100. Notice the “Einstein repeated summation convention” where repeated indices indicate a summation. The indices appear as subscripts. Notice this pseudo-inner product does not require x_μ to be zero when $x_\mu x_\mu = 0$. For this notation, the μ can appear as either a subscript or superscript without any change in meaning.

2.10.3 TENSOR NOTATION

As an alternate notation, the imaginary number can be removed by using a “metric” matrix. As is conventional, we use natural units with the speed of light $c = 1$ and $\hbar = 1$ for convenience. The various constants can be reinserted if desired.

One represents the basic 4-vector with the index in the upper position. For example, we can represent the space–time 4-vector in component form as

$$x^\mu = (t, x, y, z) = (t, \vec{r}) \quad (2.102)$$

where time t comprises the $\mu = 0$ component. Notice the conventional order of the components. The position of the index is significant. To take a pseudo-inner product, we could try writing $x^\mu x^\mu = t^2 + x^2 + \dots$ where we have used a repeated index convention. However, the result needs an extra minus sign. Instead, if we write

$$x_\mu = (t, -\vec{r}) \quad (2.103)$$

then the summation becomes $x_\mu x^\mu = (t, -\vec{r}) \cdot (t, \vec{r}) = t^2 - r^2$ where the “extra” minus sign appears. Again the position of the index is important. Lowering an index places a minus sign on the spatial part of the 4-vector.

A metric (matrix) provides a better method of tracking the minus signs. Consider the following metric

$$g^{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} = g_{\mu\nu} \quad (2.104)$$

Ordinary matrix multiplication then produces

$$x_\mu = g_{\mu\nu} x^\nu \quad (2.105a)$$

Notice the form of this result and the fact that we sum over the ν index by the summation convention. We can also write

$$x^\mu = g^{\mu\nu} x_\nu \quad (2.105b)$$

Therefore to take a pseudo-inner product, we write

$$x_\mu x^\mu = (g_{\mu\nu} x^\nu) x^\mu = (t, -\vec{r}) \cdot (t, \vec{r}) = t^2 - r^2 \quad (2.106)$$

The metric given here is the “West Coast” metric since it became most common on the west coast of the United States. The east coast metric contains a minus sign on the time component and the rest have a “+” sign.

2.10.4 DERIVATIVES

Derivatives naturally have lower indices.

$$\partial_\mu = (\partial_0, \partial_1, \partial_2, \partial_3) = \left(\frac{\partial}{\partial x^0}, \frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3} \right) = \left(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) = \left(\frac{\partial}{\partial t}, \nabla \right) = \square \quad (2.107)$$

Notice the location of the indices. The upper-index case gives

$$\partial^\mu = g^{\mu\nu} \partial_\nu = (\partial_0, -\partial_1, -\partial_2, -\partial_3) = \left(\frac{\partial}{\partial t}, -\nabla \right) \quad (2.108)$$

Let us consider a few examples. The complex plane wave has the form

$$e^{i(\vec{k} \cdot \vec{r} - \omega t)} = e^{-i(\omega t - \vec{k} \cdot \vec{r})} = e^{-ik_\mu x^\mu}$$

where $k^\mu = (\omega, \vec{k})$. Also notice that the wave equation

$$\left(\nabla^2 - \frac{\partial^2}{\partial t^2}\right)\psi = 0$$

can be written as

$$\partial_\mu \partial^\mu \psi = 0$$

Just keep in mind the repeated index convention.

As a note, any valid theory must transform correctly. The inner product is relativistically correct since it is invariant with respect to Lorentz transformations.

2.11 BRIEF DISCUSSION OF PROBABILITY AND VECTOR COMPONENTS

The quantum theory provides the mathematical apparatus to deal with the inherent uncertainty in nature. The vectors of the theory, which have an exact mathematical representation and represent the physical properties of the quantum objects, must also be associated with probability theory. Therefore, an introductory section on the relation between the vectors and the probability theory is in order. For simple formulas for probability, the quantum theory uses vectors all normalized to unity and therefore differs from the typical vector space. In fact, the set of wave functions for the quantum theory does not form a Hilbert space at all in this case. However, the quantum theory can be formulated without the normalizations so long as the definitions for the probability separately account for the normalization.

2.11.1 SIMPLE 2-D SPACE FOR STARTERS

A 2-D space has only two basis vectors denoted by $|1\rangle$ and $|2\rangle$. In the physical world these might represent the two possible energy levels for an electron or perhaps the spin-up or spin-down conditions for an electron. As a side note regarding visualization, someone might picture the spin as pointing up or down (separated by 180°) whereas the basis vector differ by only 90° in the Hilbert space. We will see the actual physical difference in spin is not exactly 180° but somewhat less. However, the important point is that each basis vector represents one of the independent states regardless of their physical geometry.

Suppose a vector $|\psi\rangle = \beta_1|1\rangle + \beta_2|2\rangle$ represents a particle. Physically, the particle will only be “found” in either state 1 or state 2, represented by $|1\rangle$ or $|2\rangle$ respectively as shown in Figure 2.18. Chapter 5 will discuss in more detail how the particle actually exists in both states (i.e., represented by the superposition $|\psi\rangle$) at the same time but upon examining the electron, it will drop out of the

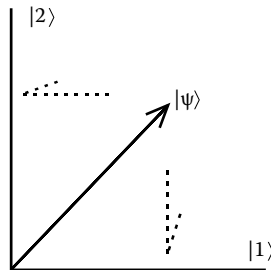


FIGURE 2.18 Superposition vector with two components.

superposition and it will be found in exactly one of the basis states (miracle and mystery of the quantum theory)—sometimes termed “the collapse of the wave function.” So the issue becomes one of asking how to mathematically relate the superposition vector to the probability of finding the particle in state 1 or state 2 upon examining it in detail.

To orient our thinking, one would quite readily agree that the probability of the particle being found in state 1 for the superposition $|\psi\rangle = \frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle$ would be 0.5 since the components of the vector have equal size. Similarly one would say the probability of state 1 for the vector $|\psi\rangle = \frac{-1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle$ would be 0.5 for the same reason even though the first component is negative.

What computational method should be used to calculate the probability of the particle being found in one basis state or the other? One begins to wonder if the probability of finding the particle in state $n = 1$, for example, should be given by $P(1) = |\beta_1|/\{|\beta_1| + |\beta_2|\}$ and something similar for the probability of finding the particle in the second state $P(2) = |\beta_2|/\{|\beta_1| + |\beta_2|\}$. On the surface, these probabilities appear to be fine in that they range between 0 and 1, and $P(1) + P(2) = 1$. There are several reasons why one should not consider such expressions for probability. First and foremost, nature does not experimentally follow this pattern. Second, the probability $P(1) = |\beta_1|/\{|\beta_1| + |\beta_2|\}$ would consist of a series of nonintuitive sharp changes when $|\psi\rangle$ makes an angle of 90° , 270° , (and so on) with respect to the $|1\rangle$ axis. That is, the first derivative of $P(1)$ with respect to angle would not be smooth.

One might speculate that the probability such as $P(1)$ should be smooth in the angle between the wave function $|\psi\rangle$ and the $|1\rangle$ axis. One can see this leads to an equation for $P(1)$, for example, which agrees with our assumption that $P(1) = |\beta_1|^2$ for a wave function $|\psi\rangle = \beta_1|1\rangle + \beta_2|2\rangle$ normalized to unity by requiring $|\beta_1|^2 + |\beta_2|^2 = 1$. Let us assume that we are dealing with a real vector space and that we do not need to worry about complex coefficients. That is, assume the vector $|\psi\rangle = \beta_1|1\rangle + \beta_2|2\rangle$ has unit length with real coefficients which requires $\beta_1^2 + \beta_2^2 = 1$. Use the notation $P(i|\beta_1, \beta_2)$ to mean the probability of state i given the coefficients have the values β_1 and β_2 ; however, the coefficients are not independent and we reduce the notation to $P(1|\beta_1)$ and $P(2|\beta_2)$.

The simplest “smooth” equation for $P(1)$ is a polynomial in β_1 , which can actually be terminated at linear powers of β_1 as will be seen below. Suppose we include the quadratic power as

$$P(1|\beta_1) = a_2\beta_1^2 + a_1\beta_1 + a_0 \quad (2.109)$$

Here the coefficients must be constants independent of the value of β_1 . The coefficients in Equation 2.109 can be determined by some general considerations. First, the probability of the particle being found in state $|1\rangle$ for $\beta_1 = 0$ must be zero which determines the coefficient a_0 as $0 = P(1|\beta_1 = 0) = a_0$. So now we have

$$P(1|\beta_1) = a_2\beta_1^2 + a_1\beta_1 \quad (2.110)$$

Next consider a_1 in Equation 2.110. Consider the case for β_1 very small but either $\beta_1 < 0$ or $\beta_1 > 0$. The fact that β_1 should be very small indicates the term with β_1^2 must be negligible (we cannot adjust a_2 since it must be independent of β_1). Now the two cases of $\beta_1 < 0$ and $\beta_1 > 0$ would require $a_1 < 0$ and $a_1 > 0$, respectively, in order to keep $P(1)$ positive. Then we must require $a_1 = 0$ to prevent a contradiction with a_1 not depending on the β_1 . Now the probability reduces to

$$P(1|\beta_1) = a_2\beta_1^2 \quad (2.111)$$

Finally, the condition $P(1|\beta_1 = 1) = 1$ requires $a_2 = 1$ and therefore

$$P(1|\beta_1) = \beta_1^2 \quad (2.112)$$

as expected from previous discussion in the chapter for the normalized wave vector $|\psi\rangle = \beta_1|1\rangle + \beta_2|2\rangle$.

In quantum theory, the wave functions are all normalized to unity. Therefore, for a 2-D space, all wave functions must terminate on the unit circle. The probability of finding the particle in any basis state (upon measurement) only depends on direction in the space through the components β_i . Sometimes people forget to normalize the wave functions ahead of time, in which case, the probability of state 1 becomes

$$P(1) = \frac{|\beta_1|^2}{\langle\psi|\psi\rangle} \quad (2.113)$$

which is the ratio of the side squared to the radius squared of the vector. The probability associated with those wave vectors without unit length would then be found as $P(1) = \beta_1^2 / \{\beta_1^2 + \beta_2^2\}$ which shows the length (squared) of the vector is used for normalization purposes and we recover $P(1) + P(2) = 1$. We see that the absolute value formula for probabilities would not provide this same intuitive simplicity in that the factor $|\beta_1| + |\beta_2|$ does not directly relate to the vector length.

Example 2.26

Consider a two level atom with states $|1\rangle$ and $|2\rangle$. Assume the electron has the wave function given by

$$|\psi\rangle = \frac{i}{\sqrt{2}}|1\rangle + \frac{-i}{\sqrt{2}}|2\rangle$$

where $i = \sqrt{-1}$. Find the probability that the electron will be found in state 2.

SOLUTION

$$|\beta_2|^2 = \frac{1}{2}$$

2.11.2 INTRODUCTION TO APPLICATIONS OF THE PROBABILITY

At this point, the classical concepts for probability theory can be applied to calculate the statistical moments (e.g., see the appendices for a review). If $P(i)$ represents the probability that the particle will be found in state i and if E_i represents the value of a quantity such as energy in state i , then the average energy will be given by

$$\langle E \rangle = \sum_i E_i P(i) \quad (2.114)$$

Interestingly, for a particle in the superposed state $|\psi\rangle = \beta_1|1\rangle + \beta_2|2\rangle$ (normalized to unity), the average energy $\langle E \rangle = |\beta_1|^2 E_1 + |\beta_2|^2 E_2$ has a value between E_1 and E_2 as it should when it has the characteristics of both basis states.

Objects with inherent randomness, will show some variation about the average. The variation is represented by the variance σ^2 and more specifically the standard deviation σ . The variance has the form

$$\sigma^2 = \langle (E - \bar{E})^2 \rangle = \langle E^2 \rangle - \bar{E}^2 \quad (2.115)$$

where $\bar{E} = \langle E \rangle$.

2.11.3 DISCRETE AND CONTINUOUS HILBERT SPACES

One can see from the previous sections in the present section that the “probability amplitude,” which is also the vector component, is given by

$$\beta_n = \langle n | \psi \rangle \quad (2.116a)$$

where once again we consider wave functions normalized to unity for simplicity. The probability of state n is then given by

$$P(n) = |\beta_n|^2 = |\langle n | \psi \rangle|^2 = \langle n | \psi \rangle \langle \psi | n \rangle \quad (2.116b)$$

We will later see that the quantity $|\psi\rangle\langle\psi|$ is the density operator. One should note the form of Equation 2.116b. The probability is formed by projecting the wave function onto the basis vectors. For the case of non-normalized wave functions, the second form of the equation provides the clue as to how to normalize the probability. One must normalize each wave function in Equation 2.116b as

$$P(n) = \langle n | \left\{ \frac{|\psi\rangle}{\|\psi\|} \frac{\langle\psi|}{\|\psi\|} \right\} | n \rangle = \frac{\langle n | \psi \rangle \langle \psi | n \rangle}{\langle \psi | \psi \rangle} = \frac{|\beta_n|^2}{\sum_n |\beta_n|^2} \quad (2.116c)$$

For the normalized wave function, the average of a quantity A_n is defined by

$$\langle A \rangle = \sum_n A_n P(n) = \sum_n A_n |\beta_n|^2 \quad (2.117)$$

The Hilbert spaces with continuous basis sets produce “similar” structures for the probability. We start with the form setup in Equation 2.116b with the projection onto the basis states. Consider the normalized wave functions and consider an example using the coordinate basis set. The probability amplitude is defined as

$$\psi(x) = \langle x | \psi \rangle \quad (2.118a)$$

where the wave function has the basis expansion $|\psi\rangle = \int dx' |x'\rangle \langle x' | \psi \rangle = \int dx' \psi(x') |x'\rangle$ and so $\psi(x)$ corresponds to something similar to β_x in the previous notation above. Rather than probability, one finds the probability density when using a form similar to 2.116b

$$P(x) = \langle x | \psi \rangle \langle \psi | x \rangle = \psi^*(x) \psi(x) \quad (2.118b)$$

If the wave functions are not normalized to unity then the probability needs to be modified according to

$$P(x) = \frac{\psi^*(x) \psi(x)}{\langle \psi | \psi \rangle} \quad (2.118c)$$

$P(x)$ is called a density since it is the probability per unit x . Integrals replace summations and the average has the form (normalized wave function case)

$$\langle A(x) \rangle = \int dx A(x) \psi^* \psi \quad (2.118d)$$

We will see in Chapter 5 that the correct form (especially for operators) is

$$\langle A(x) \rangle = \int dx \psi^* A(x) \psi \quad (2.118e)$$

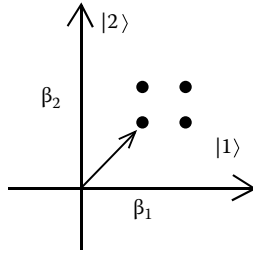


FIGURE 2.19 A random vector with four possible values.

2.11.4 CONTRAST WITH RANDOM VECTORS

One must understand the distinction between the “probability of a particle dropping into a basis vector when it previously existed in the superposition” and the “probability that a random vector takes on a particular (vector) value.” A random vector variable can be defined (in a 2-D space for example) as

$$|\psi\rangle = \beta_1|1\rangle + \beta_2|2\rangle \quad (2.119)$$

where the β_n become random variables (possibly complex) but the basis vectors do not have any randomness. Assume for the present discussion that the β_n are real and statistically independent. For example, consider Figure 2.19 showing four possible values for the random vector $|\psi\rangle$ and two possible values for each component β_n . Knowing the probability of each component $P(\beta_n)$ leads one to calculate the probability of one of the four vector values as

$$P = P(\beta_1)P(\beta_2) \quad (2.120)$$

One can sum (or integrate when appropriate) over the components to find the probability of a cluster of possible vector values.

With the random vectors, one assumes a probability distribution for the components to find the probability of a given vector value. However, for the case of the “collapsing wave function,” the probability of the particular wave function is one and we look for the probability that the particle will end up in one of the basis states. It should be clear that these two types of probability are quite different.

2.12 REVIEW EXERCISES

- 2.1** Show that the set of Euclidean vectors $\{\vec{v} = a\tilde{x} + b\tilde{y}; a, b \in \mathcal{R}\}$ forms a vector space when the binary operation is ordinary vector addition. $\mathcal{R}$ denotes real numbers and $\tilde{x}, \tilde{y}$ represent basis vectors.
- 2.2** Show that the set of Euclidean vectors $\{\vec{v} = a\tilde{x} + b\tilde{y}; a, b \in \mathcal{C}\}$ forms a vector space when the binary operation is ordinary vector addition. $\mathcal{C}$ denotes complex numbers and $\tilde{x}, \tilde{y}$ represent basis vectors.
- 2.3** Show that the set of 2-D Euclidean vectors terminating on the unit circle

$$\{\vec{v}: |\vec{v}| = 1\}$$

do not form a vector space.

- 2.4** Show that the dot product satisfies the properties of the inner product.

- 2.5** Explain what it means to say that $\vec{v} = a\tilde{x} + b\tilde{y}$ represents a mixture of the properties represented by $\tilde{x}$, $\tilde{y}$. Assume the “properties” refer to direction.
- 2.6** If $\vec{W} = -4\tilde{x} + 3\tilde{y}$ what are $\langle 1|w\rangle$, $\langle 2|w\rangle$, $\langle 3|w\rangle$?
- 2.7** If $\vec{W} = -j\tilde{x} + (3 + j2)\tilde{y}$ with $j = \sqrt{-1}$, find $\langle 1|w\rangle$, $\langle 2|w\rangle$, $\langle 3|w\rangle$.
- 2.8** If $\vec{W} = (-2 - j)\tilde{x} + (-1 + 2j)\tilde{y}$ write $\langle W|$ in terms of $\langle 1|$, $\langle 2|$.
- 2.9** If $\vec{W} = j\tilde{x} + (2 - j3)\tilde{y}$ write $\langle W|$ in terms of $\langle 1|$, $\langle 2|$.
- 2.10** If $\vec{W}_1 = -j\tilde{x} + (1 - j)\tilde{y}$ and $\vec{W}_2 = j\tilde{x} + (1 + j)\tilde{y}$ then find $\langle W_1|W_2\rangle$.
- 2.11** Show that if $\vec{v} = a\tilde{x} + b\tilde{y}$ with a, b real then $\|\vec{v}\| = 0$ requires $a = 0 = b$. There are a couple of methods to prove this but perhaps the easiest method consists of considering the factors $(a + ib)(a - ib) = 0$ where $i = \sqrt{-1}$.
- 2.12** For the basis set $\{\tilde{x}, \tilde{y}\}$ write out the closure relation.
- 2.13** Find the length of $f(x) = x$ for $x \in [0, 2]$. Show $g(x) = f(x)/\|f\|$ has unit length.
- 2.14** Prove the triangle inequality

$$|\|\vec{a}\| - \|\vec{c}\|| \leq \|\vec{a} + \vec{c}\| \leq \|\vec{a}\| + \|\vec{c}\|$$

for a 2-D vector space defined by

$$V = \{\vec{v} = \tilde{x}x + \tilde{y}y \text{ such that } x \text{ and } y \text{ real}\}$$

You can directly use the norm defined by

$$\|\vec{v}\| = \sqrt{v_x^2 + v_y^2}$$

- 2.15** Prove the triangle inequality for two vectors $\vec{f}$ and $\vec{g}$

$$|\|\vec{f}\| - \|\vec{g}\|| \leq \|\vec{f} + \vec{g}\| \leq \|\vec{f}\| + \|\vec{g}\|$$

without reference to a specific form of the inner product.

- 2.16** Find $\|\vec{v}\|^2$ when $|v\rangle = 2j|1\rangle + 3|2\rangle$ where $j = \sqrt{-1}$.
- 2.17** Find $\|\vec{v}\|^2$ when $|v\rangle = 2j|1\rangle + (3 + 2j)|2\rangle$.
- 2.18** Starting with the vector space V , show that the dual space V^* must also be a vector space. That is, show that the vectors in V^* satisfy the properties required of a vector space. Hint: use the adjoint operator.
- 2.19** Show that the adjoint operator induces an inner product on the dual space V^* . That is, show that we can define an inner product on V^* .
- 2.20** Show the set of integers (positive, negative, and zero) is countable.
- 2.21** Show the set of rational fractions is countable. A rational fraction has an integer for the numerator and denominator. Hint: Consider the set of integers on the x -axis (denominator) and the set of integers on the y -axis (numerator). Each ordered pair of integers corresponds to a different rational fraction (counted twice because of minus signs). Form a spiral around the origin and show the counting.
- 2.22** Show the even integers and the odd integers each form separate countable sets. They each therefore have the same size (cardinality). They have the same size as the set of all integers.
- 2.23** Determine if the union of two countable sets is countable. Prove or disprove in detail.
- 2.24** Write the closure relation (using a Dirac delta function) for the basis set

$$\left\{ \frac{e^{in\pi x/L}}{\sqrt{2L}} : n = 0, \pm 1, \dots \right\}$$

2.25 If $f(x) = x$ and $g(x) = x^2$ find $\langle jf | g \rangle$ on the interval $(-1, 1)$ where $j = \sqrt{-1}$.

2.26 Use change of variables to find

$$\int_{-1}^1 f(x) \delta(ax) \, dx$$

where $a > 0$

2.27 Use integration by parts to find

$$\int_{-1}^1 dx \, f(x) \delta'(x)$$

where $\delta'(x) = \frac{d}{dx} \delta(x)$

2.28 Use change of variables to find $\int_{-1}^1 f(x) \delta(ax - b) \, dx$ for $(a=2 \text{ and } b=1/2)$, and $(a=2 \text{ and } b=2)$.

2.29 Find $\int_{-1}^1 f(x, y) \delta(ax - b) \delta(cy - d) \, dx dy$. Consider all cases.

2.30 Suppose a function f has magnitude $\|f\| = 2$ in a 2-D Hilbert space with basis $\{\phi_1, \phi_2\}$. Assume the angle with respect to $|\phi_1\rangle$ is given by the parameter t . Draw a plot showing the collection of points that $|f\rangle$ traces in the $\phi_1 - \phi_2$ plane as a function of t .

2.31 Suppose a function f has magnitude $\|f\| = t$ (where t is a time parameter) in a 2-D Hilbert space with basis $\{\phi_1, \phi_2\}$. Assume the angle with respect to $|\phi_1\rangle$ is also t . Draw a plot showing the collection of points that $|f\rangle$ traces in the $\phi_1 - \phi_2$ plane as a function of t .

2.32 Consider a function f in a finite-dimensional Hilbert space with basis vectors $|\phi_i\rangle$ with $i = 1, \dots, N$ and expansion coefficients c_i . Show that if one of the coefficients is made larger, then the value of function $f(x)$ must become larger. It might be easiest to consider two function f and g corresponding to the two different sets of coefficients.

2.33 Consider the function f defined on the interval $[0, L]$ as

$$f(x) = \begin{cases} -1 & x = \text{irrational} \\ +1 & x = \text{rational} \end{cases}$$

Find $\|f\|$.

2.34 Find the constant c that normalizes the following functions to unity on the interval $[0, L]$.

a. $f(x) = c \sin(x)$ with $L = 2\pi$.

b. $f(x) = c \sin(kx)$ with $k = n\pi/L$ and $n = 1, 2, 3, \dots$

c. $f(x) = \begin{cases} -1 & x = \text{irrational} \\ +1 & x = \text{rational} \end{cases}$.

2.35 Determine if the two vectors in the following sets are independent.

a. $\{2\tilde{x}, 3\tilde{x} + 4\tilde{y}\}$.

b. $\{\tilde{x} + 2\tilde{y}, 2\tilde{x} + \tilde{y}\}$.

c. $\{\tilde{z}, 2\tilde{x} \times 3\tilde{y}\}$.

2.36 Prove the two vectors in the set $\{2\tilde{x}, 3\tilde{x} + 4\tilde{y}\}$ are independent and then use the Gram-Schmidt orthonormalization procedure to find a basis set.

2.37 Prove the two vectors in the set $\{\tilde{x} + 2\tilde{y}, 2\tilde{x} + \tilde{y}\}$ are independent and then use the Gram-Schmidt orthonormalization procedure to find a basis set.

2.38 Show that if two functions f and g are independent then the two functions $\{f, \beta_1 f + \beta_2 g\}$ are independent (where β_1, β_2 represent constants) so long as β_2 is not zero.

- 2.39** Suppose the functions f and g are independent. Find two orthonormal vectors $\{\phi_1, \phi_2\}$ (i.e., a basis set to span the same space as spanned by f and g) such that the vector ϕ_1 is parallel to $f + g$. Show that ϕ_2 is proportional to

$$|h\rangle = |g\rangle \left\{ \frac{\|f\|^2 + \langle g|f\rangle}{\|f + g\|^2} \right\} + |f\rangle \left\{ \frac{\|g\|^2 + \langle f|g\rangle}{\|f + g\|^2} \right\}$$

- 2.40** Suppose a set consists of two vectors $\{X_1(x) = \frac{1}{2L}, X_2(x) = c_1x + c_2\}$ where $x \in (-L, L)$. Find the values of c_1 and c_2 that make this a basis set. Do not use the Gram–Schmidt process. Consider orthogonality first.
- 2.41** Use the Gram–Schmidt orthonormalization procedure to turn the set of functions $\{1, x, x^2\}$ into a basis set on the interval $x \in (-1, 1)$. The results should remind you of the Legendre polynomials.
- 2.42** Use the Gram–Schmidt orthonormalization procedure to turn the set of functions $\{1, x, x^2, x^3\}$ into a basis set on the interval $x \in (0, 1)$.
- 2.43** Starting with $|f\rangle = |h\rangle + c_1|\phi_1\rangle + c_2|\phi_2\rangle$ in Section 2.6.2 for the Gram–Schmidt procedure, show $|f\rangle = |h\rangle + |\phi_1\rangle\langle\phi_1|f\rangle + |\phi_2\rangle\langle\phi_2|f\rangle$.
- 2.44** Show the set of even functions form a vector space.
- 2.45** Show the set of odd functions form a vector space.
- 2.46** Consider the sine basis functions for the space of functions defined on the interval $x \in (0, L)$

$$B_s = \left\{ \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \quad n = 1, 2, 3, \dots \right\} = \{\psi_n(x): n = 1, 2, 3, \dots\}$$

- Show the space can be expanded to include functions that repeat every $2L$ along the x -axis.
 - Consider a function defined on all reals including $(-L, 0)$. What values must the function have on $(-L, 0)$ compared with its values on the interval $(0, L)$?
 - Are there restrictions on functions defined in the interval $(L, 2L)$? Explain. Hint: consider your answers to parts a and b.
- 2.47** Consider the sine basis functions for the space of functions defined on the interval $x \in (0, L)$

$$B_c = \left\{ \frac{1}{\sqrt{L}}, \sqrt{\frac{2}{L}} \cos\left(\frac{n\pi x}{L}\right), \dots \quad \text{for } n = 1, 2, 3, \dots \right\} = \{\phi_0, \phi_1, \dots\}$$

- Show the space can be expanded to include functions that repeat every $2L$ along the x -axis.
 - Consider a function defined on all reals including $(-L, 0)$. What values must the function have on $(-L, 0)$ compared with its values on the interval $(0, L)$?
 - Are there restrictions on functions defined in the interval $(L, 2L)$? Explain. Hint: consider your answers to parts a and b.
- 2.48** Find the Fourier transform of $\delta(x - 1)$ and of $\frac{1}{2}\delta(x - 1) + \frac{1}{2}\delta(x + 1)$.
- 2.49** Show that the Fourier series basis of sines and cosines must be equivalent to the alternate basis set defined in terms of complex exponentials

$$f(x) = \sum_{n=-\infty}^{\infty} D_n \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right)$$

Hint: Start with the Fourier series expansion

$$f(x) = \frac{\alpha_0}{\sqrt{2L}} + \sum_{n=1}^{\infty} \alpha_n \frac{1}{\sqrt{L}} \cos\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} \beta_n \frac{1}{\sqrt{L}} \sin\left(\frac{n\pi x}{L}\right)$$

and rewrite the sines and cosine in terms of complex exponentials. In the summation $\sum_{n=1}^{\infty} \frac{1}{\sqrt{L}} \left(\frac{\alpha_n + i\beta_n}{2}\right) \exp(-i\frac{n\pi x}{L})$ replace n with $-n$. Combine all terms under the summation and define new constants D_n . Relate these new coefficients to the old ones as in Equation 2.58 in the chapter.

2.50 Show that the basis vectors

$$\left\{ \frac{1}{\sqrt{2L}} \exp\left(i\frac{n\pi x}{L}\right) \quad \text{for } x \in (-L, L), n = 0, \pm 1, \dots \right\}$$

must be orthonormal.

2.51 Find the sine series expansion of the function $\cos(x)$ for $x \in (0, \pi)$.

2.52 Find the cosine series expansion of the function $\sin(x)$ for $x \in (0, \pi)$.

2.53 Find the Fourier transform of $\cos(x)$ and $\sin(x)$ for x any real number in $(-\infty, \infty)$. What is the transform if the interval is restricted to $(-L, L)$.

2.54 Find the Fourier transform of the following functions

a. e^{-x^2}

b. $\frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

2.55 Suppose the unit vector $|1'\rangle$ makes an angles of α, β, γ with respect to the only three basis vectors $\{|1\rangle, |2\rangle, |3\rangle\}$. Find a relation between the three angles.

2.56 Consider the unit vectors $|1'\rangle, |2'\rangle$ that make respective angles of α, β, γ and α', β', γ' to the only three basis vectors $\{|1\rangle, |2\rangle, |3\rangle\}$. Find a relation between the angles assuming that $|1'\rangle, |2'\rangle$ are orthogonal to each other. Hint: consider an inner product for the primed vectors.

2.57 If $|v\rangle = |1\rangle_v + 4|2\rangle_v$ and $|w\rangle = 4|1\rangle_w + |2\rangle_w$ then find $|v\rangle \otimes |w\rangle$. Here the subscripts refer to the vector space V or W .

2.58 If $|v\rangle = |1\rangle_v + j|2\rangle_v$ and $|w\rangle = -4j|1\rangle_w + |2\rangle_w$ then find $|v\rangle \otimes |w\rangle$. Here $j = \sqrt{-1}$ and the subscripts refer to the vector space V or W .

2.59 Consider the direct product space $V \otimes W$ with $|\gamma\rangle = 2|1, 1\rangle + 2|2, 1\rangle$ and $\dim(V)=2, \dim(W)=1$, find the collection of vectors in V and W that produce the vector $|\gamma\rangle$.

2.60 Consider two vector spaces V and W . As discussed in connection with the direct product spaces, inner products can only be formed between V^+ and V , and between W^+ and W . If $|v_1\rangle, |v_2\rangle \in V$ and $|w_1\rangle, |w_2\rangle \in W$ then show that the definition of inner product for the direct product space $\langle v_1 w_1 | v_2 w_2 \rangle = \langle v_1 | v_2 \rangle \langle w_1 | w_2 \rangle$ satisfies the properties for inner products given in Section 2.1. Discuss any assumptions that you might make for the proof.

2.61 Consider the 2-D spaces V and W with respective basis sets $\{\tilde{x}, \tilde{y}\}$ and $\left\{ \sqrt{\frac{2}{L}} \sin\left(\frac{\pi x}{L}\right), \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi x}{L}\right) \right\}$ where the variable x has a value in the interval $(0, L)$.

a. Write the set of basis vectors for the direct product space $V \otimes W$.

b. Write the closure relation.

c. If $|\gamma\rangle = \{3|\phi_1\rangle - 2j|\phi_2\rangle\}\{j|\psi_1\rangle + 3|\psi_2\rangle\}$ where $j = \sqrt{-1}$ then find the components.

2.62 Consider the 2-D space V and the vector space W with respective basis sets $\{\tilde{x}, \tilde{y}\}$ and $\left\{ \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \quad n = 1, 2, 3, \dots \right\}$ where the variable x has a value in the interval $(0, L)$.

a. Write the set of basis vectors for the direct product space $V \otimes W$.

b. Write the closure relation.

c. If $|\gamma\rangle = \{3|\phi_1\rangle - 2j|\phi_2\rangle\}$ where $j = \sqrt{-1}$ then find the components. Hint, Fourier decompose the number “1” multiplying the $\{ \}$.

- 2.63** Suppose vector space V and W are spanned, respectively, by the Fourier transform basis sets $\{\phi_{k_x}(x) = e^{ik_x x}/\sqrt{2\pi}\}$ and $\{\phi_{k_y}(y) = e^{ik_y y}/\sqrt{2\pi}\}$.
- Write the set of basis vectors for the direct product space $V \otimes W$.
 - If $\gamma = \gamma(x, y)$ is a vector in the direct product space, write the general summation over the basis vectors.
 - Find an expression for the expansion coefficients.
 - Write the closure relation.
- 2.64** Consider the function

$$f(x, y) = \begin{cases} 1 & x \in (-1, 1) \quad y \in (-1, 1) \\ 0 & \text{otherwise} \end{cases}$$

Find the components of f in the tensor product space when the individual basis sets are

$$\left\{ e^{ik_x x}/\sqrt{2\pi} \right\} \quad \text{and} \quad \left\{ e^{ik_y y}/\sqrt{2\pi} \right\}.$$

- 2.65** Show that the tensor product space forms a Hilbert space with the given definition for the inner product.
- 2.66** Show that the expansion of a vector in its basis set is unique.
- 2.67** Suppose a linear operator appears in a partial differential equation of the form

$$\hat{L}\psi - \frac{\partial}{\partial t}\psi = \delta(x)$$

where the operator does not have any time dependence. Further suppose the operator has an eigenvector equation of the form

$$\hat{L}\phi_n(x) = c_n\phi_n(x)$$

where the set $\{\phi_n\}$ forms a basis set. Setting $\psi = \sum_n \phi_n(x)T_n(t)$, expanding $\delta(x)$ in terms of the basis set, find a solution for T .

- 2.68** Show a set of orthonormal functions $\{\phi_i\}$ must be linearly independent.
- 2.69** Suppose $\{|x\rangle\}$ represents a coordinate basis set.
- Find alternate expressions for the parameters c_x in $|f\rangle = \int dx \, c_x |x\rangle$.
 - Find $\langle k|f\rangle$ using the results of part a where $\{|k\rangle\}$ represents the Fourier transform basis.
- 2.70** Prove the remainder of the properties for the space $(W, \&)$ discussed at the end of Section 2.1.
- 2.71** Show that if a function f is an isomorphism $f: V \rightarrow W$ then so is $f^{-1}: W \rightarrow V$.
- 2.72** Determine if the set of order pairs (m, n) using typical addition and SM properties satisfy the requirements of a vector space when m, n are integers and the field of number consists of all real numbers.
- 2.73** Show the set of real numbers $(\mathcal{R}, +)$ forms a vector space.
- 2.74** Determine if $(W, *)$ is a vector space ($*$ is ordinary multiplication of real numbers) when $W = \{2^x \text{ with } x = \text{real}\}$ by directly using the properties in Section 2.1.
- 2.75** For the previous problem, does an isomorphism exist between $(\mathcal{R}, +)$ and $(W, *)$? If so, show that it is an isomorphism. If not, what property is not satisfied?

REFERENCES AND FURTHER READINGS

Classics

1. Dirac P.A.M., *The Principles of Quantum Mechanics*, 4th ed., Oxford University Press, Oxford (1978).
2. Von Neumann J., *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, NJ (1996).

Introductory

3. Krause E.F., *Introduction to Linear Algebra*, Holt, Rinehart and Winston, New York (1970).
4. Bronson R., *Matrix Methods, An Introduction*, Academic Press, New York (1970).

Standard

5. Byron F.W. and Fuller R.W., *Mathematics of Classical and Quantum Physics*, Dover Publications, New York (1970).
6. Cushing J.T., *Applied Analytical Mathematics for Physical Scientists*, John Wiley & Sons, Inc., New York (1975).
7. Von Neumann J., *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, NJ (1996).

Involved

8. Loomis L.H. and Sternberg S., *Advanced Calculus*, Addison-Wesley Publishing Co., Reading, MA (1968).
9. Stakgold I., *Green's Functions and Boundary Value Problems*, 2nd ed., John Wiley & Sons, New York (1998).

Fourier series as generalized expansions and partial differential equations

10. Brown J.W. and Churchill R.V., *Fourier Series and Boundary Value Problems*, 5th ed., McGraw-Hill, Inc., New York (1993).
11. Farlow S.J., *Partial Differential Equations for Scientists and Engineers*, Dover Publications Inc., New York (1993).
12. Weinberger H.F., *A First Course in Partial Differential Equations with Complex Variables and Transform Methods*, Dover Publications, Inc., New York (1995).
13. Davis H.F., *Fourier Series and Orthogonal Functions*, Dover Publications, Inc., New York (1963).

Mathematics—Miscellaneous and interesting

14. Naber G.L., *The Geometry of Minkowski Spacetime: An Introduction to the Mathematics of the Special Theory of Relativity*, Dover Publications, Mineola, NY (1992, 2003).
15. Kunzig R., A head for numbers, *Discover*, July issue, 108–115 (1997).
16. Dawson J.W., Godel and the limits of logic, *Scientific American*, June issue, 76–81 (1999).

3 Operators and Hilbert Space

Although Hilbert spaces are interesting mathematical objects with important physical applications, the study of linear algebra remains incomplete without a study of linear operators (i.e., linear transformations). In fact, the set of linear transformations itself forms a vector space and therefore has a basis set. This chapter uses a basis-set expansion of an operator to demonstrate the relation between the set of linear operators defined on the vector space and the vector space itself. Chapter 2 already discussed the relation by introducing projection operators and demonstrating the closure relation—a basis vector expansion of the unit operator.

Every linear operator can be represented as a matrix once having identified a basis set for the vector space. Although the operator and matrix appear to be different, the two mappings produce the same result once the results are suitably interpreted. Stated in other words, there exists an isomorphism between the space of linear operators and matrices that ensures that the two vector spaces have identical properties. Therefore, the theorems in the present chapter can be stated and proved using either the matrix or operator formalism.

A Hermitian (self-adjoint) operator produces a basis set within a Hilbert space. The basis set comes from the eigenvector equation for the particular operator. The relation between the Hermitian operator and the basis set has particular importance for quantum mechanics. Observables such as energy or momentum correspond to Hermitian operators while the basis corresponds to fundamental “states” of the system. These concepts can be approached from an alternate point of view using classical mathematical theory for boundary value problems (partial differential equations). The fundamental equation for the dynamics of the quantum system essentially comes from energy conservation and has the form of a partial differential equation—the Schrödinger wave equation. The basis set comes from the Sturm–Liouville system associated with the partial differential equation. Regardless of the method of finding the basis set, the vectors in the set can be used to expand functions/vectors that reside in the Hilbert space. The technique of separating variables produces the Sturm–Liouville system and the resulting eigenvectors provide the generalized summation to satisfy the boundary value problem.

The present chapter discusses the notion of linear operators and several representations including matrices and expansions in projection operators (a combination of bras and kets). An isomorphism links the linear operators with these representations; therefore, the spaces of operators and the representations have identical properties and dimensions. As mentioned previously, a linear operator that is self-adjoint (Hermitian) produces a basis set. The chapter also discusses methods of finding eigenvectors, change of basis, matrix properties, raising and lowering operators, and creation and annihilation operators (and their matrix representations). This chapter extends the concepts presented in Chapter 2 where we considered Hilbert spaces with both a finite and infinite number of dimensions. Many physical theories require the concept of unitary transformation as a change of basis, as will also be discussed.

3.1 INTRODUCTION TO OPERATORS AND GROUPS

Operators have the important roles of describing the transformations or the evolution of a dynamical system. One simple operator, the translation operator, displaces a system from one region to another. However, because a system is a physical object and the operator is a mathematical object, the operator must translate the vectors representing the system and its subparts. The rotation operator represents another simple operator that maps one vector into another having the same length but

often making a different angle with respect to the axes. Many of the operators will be linear in the sense that operating on the sum of two vectors produces the sum of each individual image vectors. This first section briefly introduces the idea of a linear operator and most importantly, illustrates how knowledge of the mapping of basis vectors determines the mapping of all vectors and therefore defines the linear operator.

3.1.1 LINEAR OPERATOR

Linear operators map vectors in one vector space (the domain) into vectors in another vector space (the range). The domain and range spaces can be the same or different spaces. If V and W are two vector spaces, then a linear operator acting between the spaces can be defined as $\hat{T}: V \rightarrow W$ (Figure 3.1). Note the use of the caret above the letter to denote an operator. To say that the operator $\hat{T}$ is linear means that if $|v_1\rangle$ and $|v_2\rangle$ are elements of the vector space V , and c_1, c_2 are in the set of complex numbers (denoted by C), then

$$\hat{T}[c_1|v_1\rangle + c_2|v_2\rangle] = c_1\hat{T}|v_1\rangle + c_2\hat{T}|v_2\rangle \quad (3.1)$$

where the image vectors $|w_1\rangle = \hat{T}|v_1\rangle$ and $|w_2\rangle = \hat{T}|v_2\rangle$ are members of the vector space W . Linear operators therefore have the property of superposition.

3.1.2 TRANSFORMATIONS OF THE BASIS VECTORS DETERMINE THE LINEAR OPERATOR

The linear operator $\hat{T}$ maps elements $|v\rangle$ of the domain space into other elements $\hat{T}|v\rangle$ of the range space. However, each element $|v\rangle$ in the domain must be a linear combination of the basis vectors $|\phi_i\rangle$. It seems reasonable that if we know how $\hat{T}$ affects each basis vector $|\phi_i\rangle$ then, by the property of linear superposition, we know how $\hat{T}$ maps all vectors $|v\rangle$. Therefore, we know how the linear operator $\hat{T}$ maps the entire domain space based on a “few” basis vectors.

To see how the transformation of the basis vectors determines the transformation of all vectors, consider the following example. Consider a linear operator $\hat{T}: V \rightarrow V$ that maps a vector space V into itself (Figure 3.2). Assume that the vector space $\text{Dim}(V) = 2$ with the basis set $\{|\phi_1\rangle, |\phi_2\rangle\}$ or equivalently $\{|1\rangle, |2\rangle\}$. Suppose the linear operator $\hat{T}$ produces the following mappings of the basis vectors.

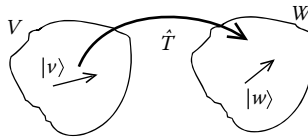


FIGURE 3.1 The operator T maps vectors from V into W .

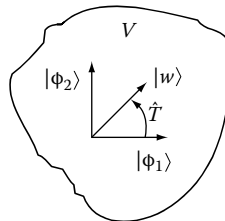


FIGURE 3.2 The operator $\hat{T}: V \rightarrow V$ maps the vector space V into itself.

$$\hat{T}|1\rangle = \frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle \quad (3.2a)$$

$$\hat{T}|2\rangle = -\frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle \quad (3.2b)$$

The example in Equations 3.2 illustrates that the image vectors $|w_1\rangle = \hat{T}|1\rangle$ and $|w_2\rangle = \hat{T}|2\rangle$ are vectors in the vector space V . As a result, the vectors $|w_1\rangle, |w_2\rangle$ can be expressed in terms of the basis vectors of V . The example shows that the operator maps each basis vector into a specific linear combination of basis vectors. Based on Equations 3.2 and Figure 3.2, the linear operator $\hat{T}$ rotates the basis vectors by 45° . We therefore expect the linear operator $\hat{T}$ to rotate every vector by 45° . As a check, consider the vector $|v\rangle = |1\rangle/\sqrt{2} + |2\rangle/\sqrt{2}$ which has unit length and initially makes a 45° angle with respect to the basis vectors. The operator then has the following effect using Equations 3.2:

$$\hat{T}|v\rangle = \hat{T}\left\{\frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle\right\} = \frac{1}{\sqrt{2}}\hat{T}|1\rangle + \frac{1}{\sqrt{2}}\hat{T}|2\rangle = |2\rangle \quad (3.3)$$

The result of the operation produces a unit vector making an angle of 90° with respect to the $|1\rangle$ axis. The operator $\hat{T}$ therefore rotates the vector by 45° (without changing its length) as expected. “Rotation” operators also have the names of orthogonal or unitary depending on the type of domain space.

We can represent a linear transformation $\hat{T}$ by a matrix. A representation of an operator $\hat{T}$ refers to a mathematical object that performs the same operations as $\hat{T}$ but has one of many different forms. We have already encountered two different representations of a function $|f\rangle$, namely the x -coordinate representation $\langle x|f\rangle = f(x)$ and the Fourier transform representation $\langle k|f\rangle = \langle \phi_k|f\rangle = f(k)$. Both of these represent the essential properties of $|f\rangle$ but in different forms. As we will see, a matrix of the operator $\hat{T}$ represents the operator by describing the effect the operator has on the basis vectors of the space. For example, the coefficients in Equation 3.2 provide the 2×2 matrix for the operator $\hat{T}$.

3.1.3 INTRODUCTION TO ISOMORPHISMS

An isomorphism is a special function (i.e., operator) that maps one set V into another W and maintains the binary operations. The set of linear operators forms one set and the image space of the isomorphism then defines the various representations. We will see that the set of matrices forms one representation while the basis vector expansion forms another.

The isomorphism is defined to be a “1–1 onto” linear function f . A function is 1–1 when for each element y in the range of f there is only one element x in the domain of f such that $f(x) = y$. On the other hand, the definition of the function already provides the condition that each element x in the domain of f maps into exactly one element in the range. In this manner, a 1–1 function always pairs exactly one element in the domain with exactly one element in the range. These same conditions provide a method to compare the size of sets of numbers. The “onto” part ensures that the space w is the same as the range of the function f . The “onto” is defined by requiring each element of W to be in the range of f ; that is, each element y in W has a preimage x in V such that $f(x) = y$. The reader will recognize that the conditions of “1–1 onto” ensure the existence of the inverse function.

3.1.4 COMMENTS ON GROUPS AND OPERATORS

A group G is a set on which multiplication (i.e., composition for operators) is defined and satisfies the following properties assuming x, y, z are elements of G .

Property	Description
Closure	If x and y in G then $x \cdot y$ is in G
Identity	There is an identity e in G such that $x \cdot e = e \cdot x = x$
Inverse	For every x in G , there is an inverse x^{-1} such that $x \cdot x^{-1} = x^{-1} \cdot x = e$
Associative	$x \cdot (y \cdot z) = (x \cdot y) \cdot z$

We are interested in a group of operations and in particular, symmetry operations. A symmetry operation maps a particular system into itself.

Example 3.1

Consider the set of operations in a two-dimensional (2-D) plane $\{R_\theta \text{ for } \theta = 0^\circ, 120^\circ, 240^\circ\}$ where R_θ refers to a rotation through an angle θ . The following table shows the multiplicative results.

Mult	R_0	R_{120}	R_{240}
R_0	R_0	R_{120}	R_{240}
R_{120}	R_{120}	R_{240}	R_0
R_{240}	R_{240}	R_0	R_{120}

R_0 is the identity while R_{120} is the inverse of R_{240} , and so on.

The present chapter generally uses the term “representation” of an operator to refer to an alternate form of the operator. For example, the abstract operator might be represented as a matrix or a basis vector expansion. However in group theory, the term “representation” has a definite meaning. For each element g of the group G , consider the mapping D that produces the image $D(g)$, which will be the representation of g . We require

$$D(g_1 \cdot g_2) = D(g_1)D(g_2)$$

(3.4)

$D(g)$ is another manner of representing g . Later, we will have primary interest in unitary operators. So if g is a rotation of a physical object, then $D(g)$ will be the unitary operator in the Hilbert space that rotates vectors. For every physical operation, there will be a corresponding mathematical one in the Hilbert space. Equation 3.4 shows the essential requirement for the representation is that the group properties should be preserved. The sequential operation by two group elements (left-hand side) should give rise to two sequential mathematical operations in Hilbert space.

For group theory, the representation most often refers to matrices. Consider a group of rotations. Each group element g will correspond to a matrix M . However, we know nothing about the matrix M except that it must represent rotations. What is the size of the matrix? This will depend on the number of physical dimensions that we are considering (for example). Rotations restricted to two dimensions will be 2×2 . Those in three dimensions will be 3×3 . So there can be a set of 2×2 matrices that represent the group G and there can also be a set of 3×3 matrices. For groups, one must specify the desired image space for the mapping D to be well defined. For the case of linear operators, we will discuss the isomorphism between the set of abstract linear operators and the set of matrices (for example) or the set of operators in a basis vector expansion (for another example). For the linear algebra, the representation is not necessarily limited to matrices so long as the multiplication properties of the operators is sustained.

A number of definitions are important for group theory:

1. The order of the finite group is the number of elements in the group.
2. A group for which all elements commute is defined to be a commutative (or abelian) group.
3. A subgroup S in the group G consists of a set of elements in G that satisfies the group properties.
4. The right coset C_g of the subgroup $S \subseteq G$ is $C_g = Sg \triangleq \{sg: s \in S\}$ for any $g \in G$.
5. A group becomes an “algebra” by defining addition and scalar multiplication. For a group of order h , the set for the algebra must contain all objects of the form

$$\sum_{i=1}^h c_i g_i$$

If we define $g_{ij} = g_i g_j \in G$ then the product of elements of the set become

$$\sum_{i=1}^h c_i g_i \sum_{j=1}^h c_j g_j = \sum_{i,j=1}^h c_i c_j g_i g_j = \sum_{i,j=1}^h c_i c_j g_{ij}$$

3.1.5 PERMUTATION GROUP AND A MATRIX REPRESENTATION: AN EXAMPLE

The permutation group provides a common example for group theory and how matrices represent the operations. On first reading of the present chapter, one can safely bypass this discussion without loss of continuity.

For convenience, consider five objects arranged in buckets as shown in Figure 3.3. It is most natural to denote the permutation by transformation notation as, for example, in

$$\hat{T}[1, 2, 3, 4, 5] = [3, 2, 1, 4, 5]$$

which switches the item in location 3 (i.e., bucket 3 in Figure 3.3) with that in location 1 (i.e., bucket 1). Each pair or type of switching would require a symbol for the operation. It is easier to use another notation for the transformations. For example, the notation $[3, 2, 1, 4, 5]$ means to take the object presently in position #3 and place it in position #1 and take the object in position #1 and place it in position #3 as shown in Figure 3.3.

We can see that the set of all permutations forms a group. The identity can be identified as $[1, 2, 3, 4, 5]$. An inverse can be identified for every element. For example, the element $[4, 2, 1, 3, 5]$

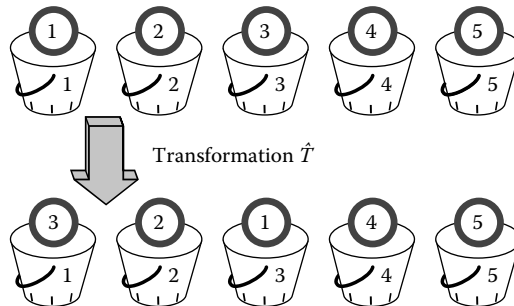


FIGURE 3.3 The permutation of objects 1 and 2. The order of the buckets corresponds to the position in the brackets $[\]$.

has the inverse $[3, 2, 4, 1, 5]$ so that $[3, 2, 4, 1, 5] [4, 2, 1, 3, 5] = [1, 2, 3, 4, 5]$. One must always remember to focus on the operations and not the objects. The operations form the group. The objects show the results for the operations.

We can now demonstrate a matrix representation. For simplicity, consider the permutation group on three objects. The objects might be “ γ, ξ, η ” originally arranged as a column matrix

$$\begin{pmatrix} \gamma \\ \xi \\ \eta \end{pmatrix} \quad (3.5)$$

The identity element of the group has the form:

$$e = [1, 2, 3] \Rightarrow D(e) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

since it does not change the order of the objects in the column matrix. Next consider the operation that switches the first two elements:

$$[2 \ 1 \ 3] \Rightarrow D[2 \ 1 \ 3] = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

One can easily see the switch of objects in positions 1 and 2 as

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma \\ \xi \\ \eta \end{pmatrix} = \begin{pmatrix} \xi \\ \gamma \\ \eta \end{pmatrix}$$

Similarly one can show the full set of matrices:

$$\begin{aligned} D[1 \ 2 \ 3] &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} & D[2 \ 1 \ 3] &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} & D[3 \ 2 \ 1] &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \\ D[1 \ 3 \ 2] &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} & D[2 \ 3 \ 1] &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} & D[3 \ 1 \ 2] &= \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{aligned} \quad (3.6)$$

3.2 MATRIX REPRESENTATIONS

Every linear operator $\hat{T}$ can be represented as a matrix $\underline{T}$. The result of a linear transformation operating on a vector can be found by first determining how the operator affects each basis vector and then adding together the results to form the image vector. The matrix of $\hat{T}$ describes the results of the transformation of the basis vectors. Operators map vectors into other vectors whereas matrices map vector components into other components. Matrices represent linear operators after the basis set has been identified for the vector space. Although the operator and matrix have different mathematical forms, once suitably interpreted, the two mappings do in fact produce the same result. The space

of linear operators and the space of matrices are isomorphic which allows the terms “operator” and “matrix” to be used interchangeably.

3.2.1 DEFINITION OF MATRIX FOR AN OPERATOR WITH IDENTICAL DOMAIN AND RANGE SPACES

First, we define the matrix for a linear operator $\hat{T}$ mapping a vector space V into itself according to $\hat{T}: V \rightarrow V$. Let V be an N -dimensional Hilbert space with basis

$$B = \{\phi_i = |\phi_i\rangle = |i\rangle : i = 1, 2, \dots, N\}$$

The matrix of the operator $\hat{T}$ with respect to the basis set B is

$$\underline{T}_{ij} = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1N} \\ T_{21} & T_{22} & \cdots & T_{2N} \\ \vdots & & & \vdots \\ T_{N1} & T_{N1} & \cdots & T_{NN} \end{bmatrix} = \underline{T}$$

where T_{ij} is defined in the relation

$$\hat{T}\phi_j = \sum_i T_{ij}\phi_i \quad (3.7)$$

Note the order of i, j on the matrix element T_{ij} . Equation 3.7 can also be written as

$$\hat{T}|\phi_j\rangle = \sum_i T_{ij}|\phi_i\rangle \quad \text{or} \quad \hat{T}|j\rangle = \sum_i T_{ij}|i\rangle$$

The collection of matrix elements will be denoted by $\underline{T}$ and the number of rows is the same as the number of columns (i.e., square matrix) for this case.

Notice how one defines the matrix in terms of the basis set. The numbers T_{ij} are related to the image vector $\hat{T}|\phi_j\rangle$. The image vector $\hat{T}|\phi_j\rangle$ must be another vector in the Hilbert space V and therefore, can be expanded in the basis set. For example, Figure 3.4 shows a 2-D space and $|v_1\rangle = \hat{T}|\phi_1\rangle$. However, $|v_1\rangle$ is also an element of the vector space V and so it can be expanded in the basis set to obtain $|v_1\rangle = a|\phi_1\rangle + b|\phi_2\rangle$ where a and b represent numbers. The same operator $\hat{T}$ would map the second basis vector $|\phi_2\rangle$ into another vector $|v_2\rangle$ in V and so we would need to use another set of constants c, d to describe the expansion of $|v_2\rangle = \hat{T}|\phi_2\rangle$ in the basis set. We would have

$$\hat{T}|\phi_1\rangle = |v_1\rangle = a|\phi_1\rangle + b|\phi_2\rangle \quad (3.8a)$$

$$\hat{T}|\phi_2\rangle = |v_2\rangle = c|\phi_1\rangle + d|\phi_2\rangle \quad (3.8b)$$

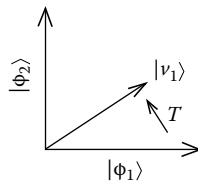


FIGURE 3.4 The operator T maps $|\phi_1\rangle$ into the vector $|v\rangle$ which itself must be a linear combination of the basis vectors.

Instead, one can invent an indexing scheme whereby the indices on a coefficient link (1) the “domain” basis vector (for example, the $|\phi_1\rangle$ on the left-hand side of Equation 3.8a) and (2) a “range” basis vector (for example, either $|\phi_1\rangle$ or $|\phi_2\rangle$ on the right-hand side of Equation 3.8a) to (3) the particular coefficient. Furthermore, rather than use $a, b, c, \dots$, we use numbers represented by a T to indicate which operator produced the mapping. So for example $T_{21} = b$ where “1” in the subscript refers to the domain vector $|\phi_1\rangle$ and the “2” refers to the component of the image vector corresponding to $|\phi_2\rangle$. Equation 3.8a, and 3.8b can be rewritten as

$$\hat{T}|\phi_1\rangle = |v_1\rangle = T_{11}|\phi_1\rangle + T_{21}|\phi_2\rangle \quad \text{and} \quad \hat{T}|\phi_2\rangle = |v_2\rangle = T_{12}|\phi_1\rangle + T_{22}|\phi_2\rangle \quad (3.8c)$$

Compare Equation 3.8a and b with Equation 3.8c until the indexing scheme becomes clear. Notice that T_{ij} represent numbers (the matrix elements); the reason for the order of the indices will become clearer once we have examined the Dirac notation for matrices. In general, $\hat{T}|\phi_j\rangle$ must be a linear combination of the basis vectors

$$\hat{T}|\phi_j\rangle = \sum_i T_{ij}|\phi_i\rangle$$

and T_{ij} are the components of the resulting vector $\hat{T}|\phi_j\rangle$.

Example 3.2

For the 2-D space with an operator $\hat{T}: V \rightarrow V$ according to

$$\hat{T}|\phi_1\rangle = |\phi_1\rangle - i|\phi_2\rangle \quad \text{and} \quad \hat{T}|\phi_2\rangle = |\phi_1\rangle + 3|\phi_2\rangle, \quad \text{find the matrix } \underline{T}.$$

SOLUTION

Equation 3.7 shows that the matrix has the form:

$$\underline{T} = \begin{bmatrix} 1 & -1 \\ -i & 3 \end{bmatrix}$$

Example 3.3

If $\hat{T}|\phi_1\rangle = T_{11}|\phi_1\rangle + T_{21}|\phi_2\rangle$, find an expression for T_{11} in terms of an inner product of the form $\langle\phi_a|\hat{T}|\phi_b\rangle$.

SOLUTION

The operator $\hat{T}$ produces $\hat{T}|\phi_1\rangle = T_{11}|\phi_1\rangle + T_{21}|\phi_2\rangle$. So T_{11} describes how much of the image vector $|v_1\rangle = \hat{T}|\phi_1\rangle$ runs along the basis vector $|\phi_1\rangle$. We can find this number by applying a projection operator $\langle\phi_1|$ to obtain $\langle\phi_1|\hat{T}|\phi_1\rangle = T_{11}\langle\phi_1|\phi_1\rangle + T_{21}\langle\phi_1|\phi_2\rangle = T_{11}$ by orthonormality of the basis set.

3.2.2 MATRIX OF AN OPERATOR WITH DISTINCT DOMAIN AND RANGE SPACES

Next consider a linear transformation acting between two distinct vector spaces such as $\hat{T}: V \rightarrow W$ where the vector space V has the basis set $B_v = \{|\phi_j\rangle: j = 1, 2, \dots, M\}$ and the vector space W has the basis set $B_w = \{|\psi_i\rangle: i = 1, 2, \dots, N\}$. The basis B_v does not necessarily have the same number of basis vectors as B_w . The resulting matrix will be square when $N = M$ and nonsquare otherwise. The matrix equation for $\hat{T}$ has the form

$$\hat{T}|\phi_j\rangle = |w\rangle = \sum_{i=1}^N T_{ij}|\psi_i\rangle \quad \text{for } j = 1, \dots, M \quad (3.9)$$

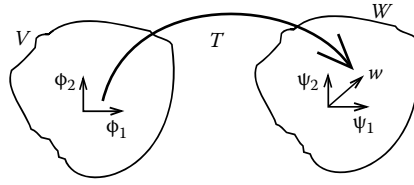


FIGURE 3.5 The linear operator T maps between vector spaces. The figure shows that the operator maps the basis vector $|\phi_1\rangle$ into the vector $|w\rangle$ which must be a linear combination of basis vectors in W .

Figure 3.5 shows that the operator maps the basis vector $|\phi_1\rangle$, for example, into a vector $|w\rangle$. Equation 3.9 then indicates that this image vector $|w\rangle$ must be a linear combination of the basis vectors for W . Once again we see that the transformation $\hat{T}$ can be defined by how it affects each of the basis vectors in V .

We do not require the $\text{Dim}[\text{domain}(T)]$ to be the same as $\text{Dim}[\text{range}(T)]$, and the $\text{Range}(T)$ does not need to be the same as the W although the range must be a subset of W . For example, as will become clear in the next sections, the operator $\hat{T} = |1\rangle\langle 1| + |1\rangle\langle 2|$ maps every vector $|v\rangle$ into a multiple of just one vector namely $|1\rangle$. For example,

$$\hat{T}[2|1\rangle + 3|2\rangle] = [|1\rangle\langle 1| + |1\rangle\langle 2|][2|1\rangle + 3|2\rangle] = 5|1\rangle$$

The dimension of the domain of $\hat{T}$ is 2 because $|1\rangle, |2\rangle$ presumably span the domain. However, the range is spanned by only a single unit vector namely $|1\rangle$ and so it has the dimension of 2.

Matrices $\underline{T}$ are not the same as operators $\hat{T}$! Matrices are arrays of “numbers” that act on the vector “components.” Operators act on “vectors.”

3.2.3 DIRAC NOTATION FOR MATRICES

Dirac notation treats Euclidean and function spaces the same although there exists some distinction between discrete and continuous basis sets. Discrete basis sets require summations for generalized expansions and Kronecker delta functions for the orthonormality relation. Continuous basis sets require integrals for the generalized summations and Dirac delta functions for the orthonormality relations. It should be kept in mind that functions can have either discrete or continuous basis sets regardless of whether the function itself is continuous or not. Now let us continue with the definition of matrices using Dirac notation.

Sometimes the order of the indices on T_{ij} for the definition of matrix

$$\hat{T}|\phi_j\rangle = \sum_i T_{ij}|\phi_i\rangle$$

might appear to be backward since the first one i refers to the basis vector on the right-hand side $|\phi_i\rangle$ and the second index j refers to the basis vector $|\phi_j\rangle$ on the left-hand side. Dirac notation straightens that out and provides a nice picture for the components T_{ij} . For simplicity, consider an operator that maps a vector space into itself according to $\hat{T}: V \rightarrow V$. As before, assume the basis vectors (Euclidean or functions) $B_V = \{|\phi_i\rangle = |i\rangle: i = 1, 2, \dots, N\}$ span the vector space V . The defining relation for the matrix of the operator $\hat{T}$ can be written as

$$\hat{T}|\phi_b\rangle = \sum_i T_{ib}|\phi_i\rangle \quad (3.10a)$$

Operating with a projection operator $\langle \phi_a |$, we have

$$\langle \phi_a | \hat{T} | \phi_b \rangle = \langle \phi_a | \sum_i T_{ib} | \phi_i \rangle = \sum_i T_{ib} \langle \phi_a | \phi_i \rangle = \sum_i T_{ib} \delta_{ai} = T_{ab} \quad (3.10b)$$

or, simply

$$T_{ab} = \langle a | \hat{T} | b \rangle \quad (3.10c)$$

So inner products involving basis vectors and the linear transformation $\hat{T}$ are really elements of a matrix. Note the order of the indices a, b . In fact, this last expression explains why the order of the indices i, j in $T | \phi_j \rangle = \sum_i T_{ij} | \phi_i \rangle$ appears to be backward (but is not).

The expression $T_{ab} = \langle a | \hat{T} | b \rangle$ can be easily interpreted: the vector $|v_1\rangle = \hat{T} | b \rangle$ comes from the linear operator $\hat{T}$ acting on the unit vector $|b\rangle$; then the number $T_{ab} = \langle a | \hat{T} | b \rangle$ must be the result of projecting $|v_1\rangle = \hat{T} | b \rangle$ onto the unit vector $|a\rangle$. That is, $T_{ab} = \langle a | \hat{T} | b \rangle$ gives the a th component of the vector $\hat{T} | b \rangle$. Figure 3.6 shows an example for the operator mapping the first basis vector into the vector v_1 and then projecting back onto the first basis vector to give the number T_{11} . This component view will be important for quantum mechanics for the following reason. The operators in quantum mechanics represent dynamical variables and produce changes in the state vectors (corresponding to the physical states of the particle or system). So $|b\rangle$ represents the original state and $\hat{T} | b \rangle$ represents the changed state. The number $T_{ab} = \langle a | \hat{T} | b \rangle$ then represents the probability that the particle transitions from state b to state a for the particular process at hand.

Obviously, expressions similar to Equations 3.10 can be written for a linear operator $\hat{T}: V \rightarrow W$ where the two sets of basis vectors are

$$B_v = \{ | \phi_a \rangle : a = 1, 2, \dots, M \} \quad B_w = \{ | \psi_i \rangle : i = 1, 2, \dots, N \}$$

and the operator $\hat{T}$ is defined by

$$\hat{T} | \phi_b \rangle = \sum_{i=1}^N T_{ib} | \psi_i \rangle \quad (3.11)$$

Notice that the vector $\hat{T} | \phi_b \rangle$ must be a vector in W and must therefore be a linear combination of the basis set for W , namely $\sum_{i=1}^N T_{ib} | \psi_i \rangle$. To continue, recall that each Hilbert space has a dual space $V \leftrightarrow V^+$ and $W \leftrightarrow W^+$; the basis set for W^+ consists of projection operators $\{ \langle \psi_j | \}$. Now because $\hat{T} | \phi_b \rangle$ must be a vector in W , we can operate on Equation 3.11 with say $\langle \psi_a |$ to find

$$\langle \psi_a | \hat{T} | \phi_b \rangle = \langle \psi_a | \sum_i T_{ib} | \psi_i \rangle = \sum_i T_{ib} \langle \psi_a | \psi_i \rangle = \sum_i T_{ib} \delta_{ai} = T_{ab}$$

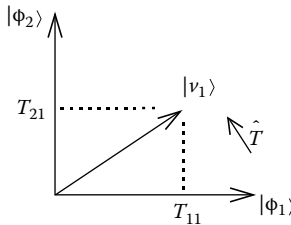


FIGURE 3.6 The operator maps basis vectors into vectors that have components in the original basis set.

Again notice that matrix elements come from inner products of operators between “basis” vectors. We will see later that quantities such as

$$\langle v|\hat{T}|w\rangle \quad \text{or} \quad \langle \phi_a|\hat{T}|\phi_b\rangle$$

can also be interpreted as expectation values (i.e., averages).

Example 3.4

Let $\hat{T}: V \rightarrow V$ and suppose that $\hat{T}$ is the unit operator; that is, $\hat{T} = 1$. Find the matrix of the transformation.

SOLUTION

To find a matrix, we need a basis set although we do not care about the exact mathematical form of the vectors in the set. We assume the following basis set for the vector space V

$$B_V = \{|\phi_j\rangle = |j\rangle : j = 1, 2, \dots, N\}$$

For each basis vector $|a\rangle \in B_V$ we can write $\hat{T}|a\rangle = \sum_{j=1}^N T_{ja}|j\rangle$ from the basic definition of the matrix. We know $\hat{T} = 1$ so that $\hat{T}|a\rangle = |a\rangle$ for each basis vector $|a\rangle$ and therefore $|a\rangle = \sum_{j=1}^N T_{ja}|j\rangle$. Now operate on both sides with the dual basis vector $\langle b|$ to find $\langle b|a\rangle = \sum_{j=1}^N T_{ja}\langle b|j\rangle = \sum_{j=1}^N T_{ja}\delta_{bj} = T_{ba}$ but we also know that the inner products between two basis vectors $|a\rangle, |b\rangle$ must be $\langle a|b\rangle = \delta_{ab}$. Therefore, by combining the last two expressions, we conclude that $T_{ba} = \delta_{ab}$. The matrix elements T_{ba} have nonzero elements only on the diagonal

$$T = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & & & \ddots \end{bmatrix}$$

3.2.4 OPERATING ON AN ARBITRARY VECTOR

The mapping of each vector $|v\rangle$ by the operator $\hat{T}$ can be determined based on how $\hat{T}$ maps each “basis” vector. The scheme works because of the linearity of $\hat{T}$. Suppose $\hat{T}: V \rightarrow V$ maps a Hilbert space into itself where V has the basis set $B_V = \{|\phi_i\rangle = |i\rangle\}$. If $|v\rangle$ is an element of the Hilbert space then we can write

$$|v\rangle = \sum \chi_n |\phi_n\rangle$$

where the symbols χ_n represent the components of the vector. Now the effect of operating with $\hat{T}$ can be found

$$\hat{T}|v\rangle = \sum_n \chi_n \hat{T}|\phi_n\rangle = \sum_n \chi_n \sum_m T_{mn} |\phi_m\rangle = \sum_{nm} (T_{mn} \chi_n) |\phi_m\rangle$$

We know the complex numbers T_{mn} and χ_n along with the basis vectors, and so we know how the operator $\hat{T}$ maps each vector $|v\rangle$ in the space. The coefficients $\sum_m (T_{mn} \chi_n)$ give the m th component of the resulting vector $|w\rangle = \hat{T}|v\rangle$.

3.2.5 MATRIX EQUATION

This section shows how an operator equation such as

$$T|v\rangle = |w\rangle \quad (3.12)$$

can be transformed into a matrix equation. For example, consider a linear transformation between distinct Hilbert spaces $\hat{T}: V \rightarrow W$ where the spaces have basis vectors given by

$$B_v = \{|\phi_i\rangle\} \quad \text{and} \quad B_w = \{|\psi_j\rangle\}$$

Assume that the vectors $|v\rangle \in \text{Sp } B_v$ and $|w\rangle \in \text{Sp } B_w$ have expansions

$$|v\rangle = \sum_n x_n |\phi_n\rangle \quad \text{and} \quad |w\rangle = \sum_m y_m |\psi_m\rangle \quad (3.13)$$

where x_n, y_m are the expansion coefficients.

We can proceed most simply by substituting Equations 3.13 into Equation 3.12 to find

$$\sum_n \hat{T}|\phi_n\rangle x_n = \sum_m y_m |\psi_m\rangle$$

Operate with $\langle\psi_m|$ on both sides to obtain

$$\sum_n \langle\psi_m|\hat{T}|\phi_n\rangle x_n = y_m \quad (3.14a)$$

The term in the summation can be identified as the matrix element because $|\phi_n\rangle, |\psi_m\rangle$ are basis vectors

$$T_{mn} = \langle\psi_m|\hat{T}|\phi_n\rangle$$

So in other words

$$\sum_n T_{mn} x_n = y_m \quad (3.14b)$$

By defining rectangular and column matrices as

$$\underline{T} = \begin{bmatrix} T_{11} & T_{12} & \cdots \\ T_{21} & \cdots & \\ \vdots & & \end{bmatrix} \quad \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} \quad \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix}$$

Equation 3.14b can be rewritten as a matrix product as

$$\begin{bmatrix} T_{11} & T_{12} & \cdots \\ T_{21} & \cdots & \\ \vdots & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} \quad (3.15)$$

In summary, the column vectors $\underline{x}, \underline{y}$ consist of the expansion coefficients from Equations 3.13, namely $|v\rangle = \sum_n x_n |\phi_n\rangle$ and $|w\rangle = \sum_m y_m |\psi_m\rangle$. The elements of $\underline{T}$ come from $T_{mn} = \langle\psi_m|\hat{T}|\phi_n\rangle$.

Example 3.5

Use the closure relation in the vector space V to find the results given in Equations 3.14b and 3.15.

SOLUTION

Start with the equation $\hat{T}|v\rangle = |w\rangle$ and insert a unit operator between $\hat{T}$ and $|v\rangle$ so as to find $\hat{T}1|v\rangle = |w\rangle$. Using the completeness relation for the vector space V , $1 = \sum_b |\phi_b\rangle\langle\phi_b|$ gives upon substituting it into the previous equation

$$\hat{T} \sum_b |\phi_b\rangle\langle\phi_b|v\rangle = |w\rangle \quad \text{and therefore} \quad \sum_b \hat{T}|\phi_b\rangle\langle\phi_b|v\rangle = |w\rangle$$

Now, because $|v\rangle = \sum_n x_n |\phi_n\rangle$, the inner product provides $\langle\phi_b|v\rangle = x_b$ and so the last expression can be rewritten as $\sum_b \hat{T}|\phi_b\rangle x_b = |w\rangle$. Next operate on both sides with one of the basis vectors $\langle\psi_a|$ in the dual vector space W^+

$$\sum_b \langle\psi_a|\hat{T}|\phi_b\rangle x_b = \langle\psi_a|w\rangle$$

Now evaluate the terms. Equation 3.13 shows that $\langle\psi_a|w\rangle = y_a$ and also, by definition of the matrix element, $\langle\psi_a|\hat{T}|\phi_b\rangle = T_{ab}$ (since ψ_a, ϕ_b are basis vectors). Substituting these terms, Equation 3.14a becomes

$$\sum_b T_{ab} x_b = y_a \quad \text{or} \quad \underline{T} \underline{x} = \underline{y}$$

$$\begin{bmatrix} T_{11} & T_{12} & \cdots \\ T_{21} & \cdots & \\ \vdots & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix}$$

The expansion coefficients of the vectors appear in the column matrices.

Example 3.6

Find the matrix representation of an operator $\hat{T}: V \rightarrow V$ that maps a 2-D vector space (Euclidean or function) into itself according to

$$\begin{aligned} \hat{T}|1\rangle &= \frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle \\ \hat{T}|2\rangle &= -\frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle \end{aligned} \tag{3.16}$$

where the vector space has the basis set

$$B_V = \{\phi_1, \phi_2 = |1\rangle, |2\rangle\}$$

using Dirac notation.

SOLUTION

Figure 3.7 shows the image of the basis vectors as indicated by the labels $\hat{T}|1\rangle, \hat{T}|2\rangle$. The image vectors $\hat{T}|1\rangle, \hat{T}|2\rangle$ must be linear combinations of the original basis vectors as given by

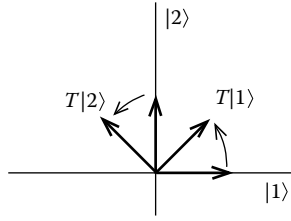


FIGURE 3.7 The operator T rotates the basis vectors.

Equations 3.16. Inner products $\langle i|\hat{T}|j\rangle$ provide the matrix elements of the operator $\hat{T}$. Using Equations 3.16 and operating with $\langle 1|$ and $\langle 2|$ on each of them, we find

$$\begin{aligned} T_{11} &= \langle 1|\hat{T}|1\rangle = \frac{1}{\sqrt{2}} & T_{12} &= \langle 1|\hat{T}|2\rangle = -\frac{1}{\sqrt{2}} \\ T_{21} &= \langle 2|\hat{T}|1\rangle = \frac{1}{\sqrt{2}} & T_{22} &= \langle 2|\hat{T}|2\rangle = \frac{1}{\sqrt{2}} \end{aligned}$$

so that

$$\underline{T} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

The reader will recognize the operator T as a rotation through a 45° angle.

Example 3.7

Continue the previous example and find the matrix representation of the operator equation $\hat{T}|v\rangle = |v'\rangle$ where the vectors are expressed in the basis set as

$$|v\rangle = v_x|1\rangle + v_y|2\rangle \quad |v'\rangle = v'_x|1\rangle + v'_y|2\rangle$$

The column matrix representation of each vector can be found by operating on both sides of both equations with $\langle 1|$ and $\langle 2|$ so that

$$\underline{v} = \begin{bmatrix} \langle 1|v\rangle = v_x \\ \langle 2|v\rangle = v_y \end{bmatrix} \quad \underline{v'} = \begin{bmatrix} \langle 1|v'\rangle = v'_x \\ \langle 2|v'\rangle = v'_y \end{bmatrix}$$

Therefore, the matrix representation of the operator equation $\hat{T}|v\rangle = |v'\rangle$ is

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} v'_x \\ v'_y \end{bmatrix}$$

3.2.6 MATRICES FOR FUNCTION SPACES

First, consider the general meaning of an object such as $\langle w|\hat{T}|v\rangle$ when $w = w(x)$ and $v = v(x)$ are functions. The object $\langle w|\hat{T}|v\rangle$ is not to be thought of as an operator. The simplest case assumes $\hat{T}$ is diagonal in the spatial variable x such as for $\hat{T} \sim d/dx$. Diagonal in the “spatial” coordinate means that

$$\langle x'|\hat{T}|x''\rangle = \hat{T}(x'')\langle x'|x''\rangle \quad (3.17)$$

For this diagonal case, the expectation values $\langle w|T|v\rangle$ can be calculated by using the spatial-coordinate closure relation a couple of times.

$$\begin{aligned} \langle w|\hat{T}|v\rangle &= \langle w|\hat{1}\hat{T}\hat{1}|v\rangle = \int dx' \int dx'' \langle w|x'\rangle \langle x'|\hat{T}|x''\rangle \langle x''|v\rangle \\ &= \int dx' dx'' W^*(x') \hat{T}(x'') \langle x'|x''\rangle v(x'') \\ &= \int dx' dx'' W^*(x') \hat{T}(x'') \delta(x'' - x') v(x'') \\ &= \int dx' W^*(x') \hat{T}(x') v(x') \end{aligned}$$

More general quantities will have the form $\langle x'|T|x''\rangle \sim T(x', x'')$.

Example 3.8

Find the matrix representation of the operator

$$T = -\frac{d^2}{dx^2}$$

for the basis vectors given by

$$B = \{\phi_m(x)\} = \left\{ \sqrt{\frac{2}{L}} \sin\left(\frac{m\pi x}{L}\right) : x \in (0, L), \quad m = 1, 2, \dots \right\}$$

The matrix is found by calculating matrix elements of the form:

$$T_{mn} = \langle \phi_m|\hat{T}|\phi_n\rangle = \langle m|\hat{T}|n\rangle$$

The matrix element

$$T_{mn} = \langle \phi_m|\hat{T}|\phi_n\rangle = \langle \phi_m|\hat{T}\phi_n\rangle$$

has the form of an inner product which is an integral for functions:

$$\begin{aligned} T_{mn} &= \langle \phi_m|\hat{T}|\phi_n\rangle = \langle \phi_m|\hat{T}\phi_n\rangle \equiv \int_0^L dx \phi_m^*(x) \left[-\frac{\partial^2}{\partial x^2} \right] \phi_n(x) \\ &= \int_0^L dx \phi_m^*(x) \left[-\frac{\partial^2}{\partial x^2} \right] \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) = \int_0^L dx \phi_m^*(x) \left[\frac{n\pi}{L} \right]^2 \phi_n(x) \end{aligned}$$

The last line can now be written as

$$T_{mn} = \left(\frac{n\pi}{L}\right)^2 \langle \phi_m | \phi_n \rangle = \left(\frac{n\pi}{L}\right)^2 \delta_{mn}$$

The matrix can be written as

$$T_{ij} = \begin{bmatrix} \left(\frac{\pi}{L}\right)^2 & 0 & \dots \\ 0 & \left(\frac{2\pi}{L}\right)^2 & \\ \vdots & & \end{bmatrix}$$

3.2.7 INTRODUCTION TO OPERATOR EXPECTATION VALUES

It will be important in the quantum theory to find the expectation value of operators. Given that Hermitian operators represent physically observable quantities (such as energy), the average of the operator actually refers to the average of the particular physical quantity. We now provide a mathematical discussion of the average (and other statistical moments) of an operator. Chapter 5 will provide a more complete physical picture.

The average of an operator $\hat{O}$ for the state $|\psi\rangle$ has the form:

$$\langle \hat{O} \rangle = \langle \psi | \hat{O} | \psi \rangle \quad (3.18)$$

Usually the operators are required to be Hermitian which have eigenvectors that can be used for basis sets. Physical observables correspond to Hermitian operators because they have real eigenvalues and a complete set of eigenvectors (as we will see later in the chapter). We use the Hermitian operators with the eigenvectors $|n\rangle$ as the basis $B = \{|1\rangle, \dots, |n\rangle, \dots, |N\rangle\}$ with

$$\hat{O}|n\rangle = o_n|n\rangle \quad (3.19)$$

to give some idea on how Equation 3.18 represents an average.

We will need the concept from the last section of Chapter 2 of how a vector

$$|\psi\rangle = \sum_n \beta_n |n\rangle \quad (3.20)$$

gives rise to the probability

$$P(n) = |\beta_n|^2 \quad (3.21)$$

of finding a particular basis vector in $|\psi\rangle$. Now we can better understand the definition of average by expanding Equation 3.18

$$\langle \hat{O} \rangle = \langle \psi | \hat{O} | \psi \rangle = \sum_{mn} \beta_m^* \beta_n \langle m | \hat{O} | n \rangle \quad (3.22a)$$

and using Equation 3.19 to find

$$\langle \hat{O} \rangle = \langle \psi | \hat{O} | \psi \rangle = \sum_n o_n |\beta_n|^2 = \sum_n o_n P(n) \quad (3.22b)$$

We recognize the last term as the classical definition for an average.

Now one might interpret the average as follows. The value o_n represents the value of the operator $\hat{O}$ for the state $|n\rangle$ by virtue of Equation 3.19 (i.e., $o_n = O_{nn} = \langle n|\hat{O}|n\rangle$). But when we try to find a particular basis vector $|n\rangle$, we know that the probability of finding it will be $P(n) = |\beta_n|^2$. This means that the probability the operator will have the value o_n must also be $P(n) = |\beta_n|^2$. So therefore, the expected value of the operator must be given by Equation 3.22b.

Other types of averages can be defined similarly. The average of an operator $\hat{O}^2 = \hat{O}\hat{O}$ for the state $|\psi\rangle$ will be

$$\langle \hat{O}^2 \rangle = \langle \psi | \hat{O}^2 | \psi \rangle \quad (3.23)$$

One can also define a variance σ^2 and standard deviation $\sigma = \sqrt{\sigma^2}$. Again, one prefers Hermitian operators which produce real eigenvalues, and have eigenvectors that span the space (i.e., complete basis) and produce real averages and variances.

$$\sigma^2 = \langle (\hat{O} - \bar{O})^2 \rangle = \langle \hat{O}^2 \rangle - \bar{O}^2 \quad (3.24)$$

where $\bar{O} = \langle \hat{O} \rangle$. One should notice, for this quantum mechanical style average, one must always specify the state $|\psi\rangle$ for the average to have meaning. The standard deviation measures how close $|\psi\rangle$ is to exactly one of the basis vectors as illustrated in the next example.

Example 3.9

Calculate the variance when $|\psi\rangle = |n\rangle$, one of the basis vectors for which $\hat{O}|n\rangle = o_n|n\rangle$.

SOLUTION

Start with the quantities in Equation 3.24

$$\bar{O} = \langle n | \hat{O} | n \rangle = \langle n | o_n | n \rangle = o_n \langle n | n \rangle = o_n$$

Similarly,

$$\langle \hat{O}^2 \rangle = \langle n | \hat{O}^2 | n \rangle = \langle n | o_n^2 | n \rangle = o_n^2$$

As a result, we find

$$\sigma^2 = \langle (\hat{O} - \bar{O})^2 \rangle = \langle \hat{O}^2 \rangle - \bar{O}^2 = 0$$

3.2.8 MATRIX NOTATION FOR AVERAGES

Quantum theory represents observables (such as energy or momentum) by Hermitian operators. Often we have an interest in knowing the average value of an observable. We therefore defined the average of a linear operator $\hat{T}: V \rightarrow V$ for the state $|v\rangle$ defined by

$$\langle \hat{T} \rangle = \langle v | \hat{T} | v \rangle$$

Transitions of electrons between states (such as for optical transitions) requires an expectation-style value be defined for unlike states $|v\rangle, |w\rangle$

$$\langle v | \hat{T} | w \rangle$$

In general, the vectors $|v\rangle, |w\rangle$ can be members of a single vector space or in two distinct spaces depending on the nature of $\hat{T}$.

These expectation values can be written in matrix notation. Identical expressions hold for either Euclidean or function space. We now show the matrix form of the inner product $\langle w|\hat{T}|v\rangle$. Consider two vectors in their respective spaces $|v\rangle \in V = Sp\{|n\rangle\}$, $|w\rangle \in W = Sp\{|m\rangle\}$. We assume that the operator $\hat{T}: V \rightarrow W$ maps between the vector spaces. We can write

$$\begin{aligned}\langle w|\hat{T}|v\rangle &= \langle w|1\hat{T}1|v\rangle = \langle w|\left(\sum_m |m\rangle\langle m|\right)\hat{T}\left(\sum_n |n\rangle\langle n|\right)|v\rangle \\ &= \sum_{m,n} \langle w|m\rangle\langle m|\hat{T}|n\rangle\langle n|v\rangle = \sum_{m,n} w_m^* T_{mn} v_n \\ &\equiv \underline{w}^+ \underline{T} \underline{v}\end{aligned}$$

Notice that we define the Hermitian conjugate of the column vector as follows:

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}^+ = [w_1^* \quad w_2^* \quad \cdots]$$

What alternate form does the inner product $\langle w|\hat{T}|v\rangle$ take for Euclidean and functions spaces? First of all, the object $\langle w|\hat{T}|v\rangle$ can be called an inner product because $\hat{T}|v\rangle = |\hat{T}v\rangle$ is an element of the W space and therefore $\langle w|\hat{T}|v\rangle = \langle w|\hat{T}v\rangle$ is an inner product between two vectors in the W space. Next, the inner product can be written for either Euclidean space or for function spaces. For Euclidean space

$$\langle w|\hat{T}|v\rangle = \sum_{ij} w_i^* T_{ij} v_j$$

and for function space

$$\langle w|\hat{T}|v\rangle = \int dx \, w^*(x) \hat{T}(x) v(x)$$

3.3 COMMON MATRIX OPERATIONS

The previous discussions have shown that every linear operator $\hat{T}$ corresponds to a matrix T_{ab} . The space L of all linear operators (acting between two vector spaces) is isomorphic to a space of matrices. In fact, the set L itself forms a vector space. We review the composition of operators, determinants, inverses, and trace.

3.3.1 COMPOSITION OF OPERATORS

Suppose $\hat{S}: U \rightarrow V$ and $\hat{T}: V \rightarrow W$ are two linear operators and U, V, W are three distinct vector spaces with the following basis sets (Figure 3.8)

$$B_u = \{|\chi_i\rangle\} \quad B_v = \{|\phi_j\rangle\} \quad B_w = \{|\psi_k\rangle\}$$

The composition (i.e., product) $\hat{R} = \hat{T}\hat{S}$ first maps the space U to the space V and then maps V to W . The matrix of $\hat{R} = \hat{T}\hat{S}$ must involve the basis vectors B_u and B_w according to the basic definition of the matrix as found in Section 3.2. The operator $\hat{R} = \hat{T}\hat{S}$ corresponds to the product of matrices.

$$R_{ab} = \langle \psi_a | \hat{R} | \chi_b \rangle = \langle \psi_a | \hat{T} \hat{S} | \chi_b \rangle$$

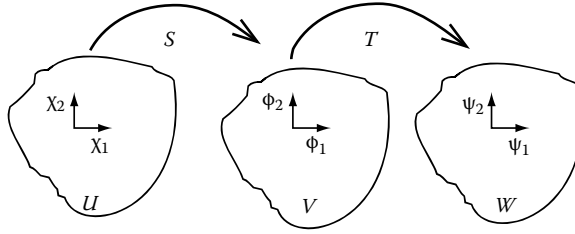


FIGURE 3.8 Three vector spaces for the composition of functions.

Inserting the closure relation between $\hat{T}$ and $\hat{S}$ gives

$$R_{ab} = \langle \psi_a | \hat{T} \hat{S} | \chi_b \rangle = \langle \psi_a | \hat{T} \left(\sum_c |\phi_c\rangle \langle \phi_c| \right) \hat{S} | \chi_b \rangle = \sum_c \langle \psi_a | \hat{T} | \phi_c \rangle \langle \phi_c | \hat{S} | \chi_b \rangle = \sum_c T_{ac} S_{cb} \quad (3.25)$$

Notice that the closure relation for the set V is inserted between $\hat{T}$ and $\hat{S}$ which corresponds to the range of $\hat{S}$ and the domain of $\hat{T}$. The last equation shows that the composition of operators corresponds to the multiplication of matrices $\underline{R} = \underline{T}\underline{S}$.

3.3.2 ISOMORPHISM BETWEEN OPERATORS AND MATRICES

The existence of an isomorphism (a 1–1, onto, linear) $M: \{\hat{T}\} \rightarrow \{\underline{T}\}$ between a set of operators and a set of matrices ensures identical properties for each. The properties of one set can be deduced from the properties of the other. The requirement of “linear” applies to the vector space aspects of operators and matrices. The set of operators forms a group with respect to the addition of operators. However, a group can also be formed from the operators with respect to composition (i.e., multiplication) which can be used to deduce the definition of matrix multiplication.

We already know an isomorphic mapping that relates the operator $\hat{T}$ to the matrix $\underline{T}$. The relation is given by $\hat{T} = \sum_{ab} T_{ab} |\phi_a\rangle \langle \psi_b|$ where $V_1 = \text{Sp}\{|\phi_a\rangle\}$, $V_2 = \text{Sp}\{|\psi_a\rangle\}$, and $\hat{T}: V_1 \rightarrow V_2$. Each different linear operator $\hat{T}$ gives a different collection of matrix elements $\underline{T}$ and vice versa (1–1 and onto).

Requiring $M(\hat{S})M(\hat{T}) = M(\hat{S}\hat{T})$ gives the required matrix multiplication as follows. Let $\hat{U} = \hat{S}\hat{T}$ where $\hat{S}: V_2 \rightarrow V_3$ and $V_3 = \text{Sp}\{|\omega_a\rangle\}$ so that $\hat{U}: V_1 \rightarrow V_3$. Then the multiplication property of M produces

$$\begin{aligned} \underline{S}\underline{T} &= M(\hat{S})M(\hat{T}) = M(\hat{S}\hat{T}) = M\left\{ \sum_{ab} S_{ab} |\omega_a\rangle \langle \psi_b| \sum_{cd} T_{cd} |\psi_c\rangle \langle \phi_d| \right\} \\ &= M\left\{ \sum_{abd} S_{ab} T_{bd} |\omega_a\rangle \langle \phi_d| \right\} \end{aligned}$$

where the orthonormality on V_2 has been used. Then the resulting matrix of the product operator is given by

$$\underline{S}\underline{T} = \left\{ \sum_b S_{ab} T_{bd} \right\}$$

where “ $\{\}$ ” refers to the collection of matrix elements. Notice that M essentially “picks off” the coefficients such as S_{ab} in $M(\hat{S})$. This agrees with the usual definition for matrix multiplication.

3.3.3 DETERMINANT

The “determinant of an operator” is defined to be the determinant of the corresponding matrix

$$\det(\hat{T}) = \det(\underline{T})$$

Generally, we assume for simplicity that the operator $\hat{T}$ operates within a single vector space (since the matrix needs to be square). The determinant can be written in terms of a completely “antisymmetric” tensor $\varepsilon_{ijk\dots}$, often termed the Levi-Cevita symbol.

$$\det(\underline{T}) = \sum_{i,j,k\dots} \varepsilon_{ijk\dots} T_{1i} T_{2j} T_{3k} \dots \quad \text{where } \varepsilon_{ijk\dots} = \begin{cases} +1 & \text{even permutations of } 1, 2, 3, \dots \\ -1 & \text{odd permutation of } 1, 2, 3, \dots \\ 0 & \text{if any of } i = j = k \text{ holds} \end{cases} \quad (3.26)$$

For example $\varepsilon_{132} = -1$, $\varepsilon_{312} = +1$, and $\varepsilon_{133} = 0$.

Another common method to evaluate the determinant is to “expand” along a row or column. Consider expanding a 3×3 determinant $\underline{T}$ along the top row.

$$\begin{vmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{vmatrix} = T_{11} \begin{vmatrix} T_{22} & T_{23} \\ T_{32} & T_{33} \end{vmatrix} - T_{12} \begin{vmatrix} T_{21} & T_{23} \\ T_{31} & T_{33} \end{vmatrix} + T_{13} \begin{vmatrix} T_{21} & T_{22} \\ T_{31} & T_{32} \end{vmatrix}$$

The same technique can be used for any column or row for any square matrix. Keep in mind that every other term must have a minus sign. Expanding along the second column, for example, requires the leading term to start with a minus sign as does every other term after that. The rules for minus signs easily follow from the basic definition of the determinant in Equation 3.26.

Here is several useful properties (see the chapter review exercises) for the matrices of operators that map a vector space V into itself.

1. The inverse of a square matrix exists provided its determinant is not zero.
2. $\text{Det}(\underline{A} \underline{B} \underline{C}) = \text{Det}(\underline{A}) \text{Det}(\underline{B}) \text{Det}(\underline{C})$.
3. $\text{Det}(c\underline{A}) = c^N \text{Det}(\underline{A})$ where $N = \text{Dim}(V)$ and c is a complex number.
4. $\text{Det}(\underline{A}^T) = \text{Det}(\underline{A})$ where T signifies transpose.
5. The $\text{Det}(\underline{A})$ is independent of the particular basis chosen for the vector space.

The proofs will be found in the subsequent sections and as some of the chapter review problems. The proof of property 5 will become more obvious after discussing orthogonal and unitary transformations. For now, we mention that unitary operators change basis. Unitary operators have the property that $\hat{u}^+ = \hat{u}^{-1}$. Applying a unitary operator to the operator produces

$$\hat{A}' = \hat{u} \hat{A} \hat{u}^+$$

Then using property 1, we find

$$\text{Det}(\hat{A}') = \text{Det}(\hat{u}) \text{Det}(\hat{A}) \text{Det}(\hat{u}^+) = \text{Det}(\hat{A}) \text{Det}(\hat{u} \hat{u}^+) = \text{Det}(\hat{A}) \text{Det}(1) = \text{Det}(\hat{A})$$

We will later see more properties of the determinant as related to the type of linear operator and eigenvalues.

Example 3.10

Evaluate the following

$$\text{Det}(\underline{A}) = \text{Det} \begin{bmatrix} 4 & 0 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

using the antisymmetric tensor.

SOLUTION

The matrix $\underline{A}$ has three rows and columns so there will be three indices on the antisymmetric tensor.

$$\text{Det}(\underline{A}) = \sum_{i,j,k} \varepsilon_{ijk} A_{1i} A_{2j} A_{3k} = \varepsilon_{111} A_{11} A_{21} A_{31} + \varepsilon_{112} A_{11} A_{21} A_{32} + \cdots$$

Terms with repeated indices in the Levi-Civita symbol produce zero. We are left with

$$\begin{aligned} \text{Det}(\underline{A}) &= \sum_{i,j,k} \varepsilon_{ijk} A_{1i} A_{2j} A_{3k} \\ &= \varepsilon_{123} A_{11} A_{22} A_{33} + \varepsilon_{132} A_{11} A_{23} A_{32} + \varepsilon_{213} A_{12} A_{21} A_{33} + \varepsilon_{231} A_{12} A_{23} A_{31} \\ &\quad + \varepsilon_{312} A_{13} A_{21} A_{32} + \varepsilon_{321} A_{13} A_{22} A_{31} \\ &= A_{11} A_{22} A_{33} - A_{11} A_{23} A_{32} + A_{12} A_{23} A_{31} - A_{12} A_{21} A_{33} + A_{13} A_{21} A_{32} - A_{13} A_{22} A_{31} \\ &= 4 \cdot 2 \cdot 1 - 4 \cdot 2 \cdot 0 + 0 \cdot 2 \cdot 0 - 0 \cdot 0 \cdot 1 + 4 \cdot 0 \cdot 0 - 4 \cdot 2 \cdot 0 \\ &= 8 \end{aligned}$$

Example 3.11

Calculate the same determinant by expanding along the bottom row.

$$\text{Det}(A) = \text{Det} \begin{bmatrix} 4 & 0 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix} = 0 \begin{vmatrix} 0 & 4 \\ 2 & 2 \end{vmatrix} - 0 \begin{vmatrix} 4 & 4 \\ 0 & 2 \end{vmatrix} + 1 \begin{vmatrix} 4 & 0 \\ 0 & 2 \end{vmatrix} = 8$$

Example 3.12

Show that $\text{Det}(c\underline{A}) = c^N \text{Det}(\underline{A})$ for the simple case of

$$\underline{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

SOLUTION

$$\text{Det } c \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \text{Det} \begin{bmatrix} 1c & 2c \\ 3c & 4c \end{bmatrix} = -2c^2 = c^2 \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

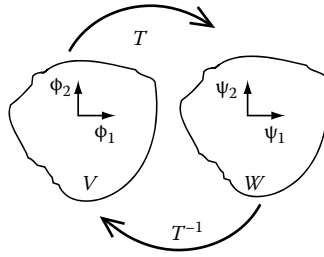


FIGURE 3.9 Inverse of an operator.

3.3.4 INTRODUCTION TO THE INVERSE OF AN OPERATOR

Given an operator $\hat{T}: V \rightarrow W$, we want to find an operator $\hat{T}^{-1}$ such that $\hat{T}\hat{T}^{-1} = 1 = \hat{T}^{-1}\hat{T}$. In such a case, an equation of the form $\hat{T}|v\rangle = |w\rangle$ can be inverted to give $|v\rangle = \hat{T}^{-1}|w\rangle$. If $\hat{T}: V \rightarrow W$ operates between spaces or even within one space, the function T must be “1–1” and “onto” to have an inverse (Figure 3.9). The term “1–1” means that every vector in the vector space V is mapped into a unique vector in the space W . The term “onto” means that every vector $|w\rangle \in W$ in the vector space W has a preimage $|v\rangle \in V$ such that $\hat{T}|v\rangle = |w\rangle$.

The null space (also known as the kernel) provides a means for determining if a linear operator $\hat{T}: V \rightarrow W$ can be inverted. We define the null space to be the set of vectors $N = \{|v\rangle\}$ such that $\hat{T}|v\rangle = 0$. Obviously, if the null space contains more than a single element (i.e., an element other than zero), the operator does not have any inverse since an element of the range has multiple preimages. Furthermore, the end-of-chapter problems demonstrate the relation:

$$\text{Dim}(V) = \text{Dim}(W) + \text{Dim}(N) \quad (3.27)$$

for $\hat{T}: V \rightarrow W$ where $W = \text{Range}(\hat{T})$. This particular definition of W automatically requires the operator to be “onto.” In this case, the value of $\text{Dim}(N)$ dictates whether or not the operator $\hat{T}$ is 1–1 and therefore whether or not it has any inverse. We assure the 1–1 property of the operator when we require $\text{Dim}(N) = 0$. Alternatively, we can also require the determinant to be nonzero $\text{Det}(\hat{T}) \neq 0$ for the operator to be invertible.

Example 3.13

Using $\underline{A} = \begin{bmatrix} 4 & 0 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix}$ calculate the following quantities

- Find $\underline{A}^{-1}$ if it exists.
- What are the basis vectors? (Trick question)

SOLUTIONS

- Inverse operator

First note that the determinant of the operator is $\text{Det}(\hat{A}) = 8$ and not zero so it makes sense to find the inverse. We see that the determinant is not zero and so we can find an inverse matrix. Although inverse matrices can be found by using determinants, we use elementary row operations on the composite matrix given by

$$\left[\begin{array}{ccc|ccc} 4 & 0 & 4 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right]$$

The right-hand side consists of the unit matrix and the left-hand side as the original matrix to be inverted. The objective is to transform the left-hand side into the unit matrix by using elementary row operations and the right-hand side will be the inverse matrix. Notice that the row operations apply to the entire six-element row. We use the notation

$$R1/4 \quad R2 - R3 \rightarrow R3$$

to mean “divide first row by 4” and “subtract the third row from the second row and substitute the results into the third row.”

$$\begin{aligned} & \left[\begin{array}{ccc|ccc} 4 & 0 & 4 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right] \xrightarrow[R2/2]{R1/4} \left[\begin{array}{ccc|ccc} 1 & 0 & 1 & 0.25 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right] \\ & \xrightarrow{R1-R3 \rightarrow R1} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.25 & 0 & -1 \\ 0 & 1 & 1 & 0 & 0.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right] \xrightarrow{R2-R3 \rightarrow R2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.25 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0.5 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right] \end{aligned}$$

So we can write the inverse matrix as

$$\underline{A}^{-1} = \begin{bmatrix} 0.25 & 0 & -1 \\ 0 & 0.5 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

- b. The exact form of the basis vectors remains unspecified. The set $\{|\hat{i}\rangle\}$ can be $\{\hat{x}, \hat{y}, \hat{z}\}$ or even $\left\{\sqrt{\frac{2}{L}}\sin\left(\frac{\pi x}{L}\right), \sqrt{\frac{2}{L}}\sin\left(\frac{2\pi x}{L}\right), \sqrt{\frac{2}{L}}\sin\left(\frac{3\pi x}{L}\right)\right\}$. The matrix tells you nothing about the exact nature of the vector space. This is part of the reason why matrices have such general application to so many different fields.

Example 3.14

If an operator $\hat{H}$ can be written as

$$\hat{H} = \sum_a E_a |a\rangle\langle a|$$

with $E_a \neq 0$ for every allowed index a , show that the inverse of the operator is given by

$$\hat{H}^{-1} = \sum_b \frac{1}{E_b} |b\rangle\langle b|$$

We need to show that $HH^{-1} = H^{-1}H = 1$ (both must be true). We will only show that

$$HH^{-1} = 1$$

Substituting the expansions for the operators gives us

$$\begin{aligned} \hat{H}\hat{H}^{-1} &= \sum_a E_a |a\rangle\langle a| \sum_b \frac{1}{E_b} |b\rangle\langle b| = \sum_{ab} \frac{E_a}{E_b} |a\rangle\langle a|b\rangle\langle b| \\ &= \sum_{ab} \frac{E_a}{E_b} |a\rangle\langle b| \delta_{ab} = \sum_a |a\rangle\langle a| = 1 \end{aligned}$$

where of course the last result is obtained by closure on the Hilbert space.

3.3.5 TRACE

The trace of an operator $\hat{T}: V \rightarrow V$ is the trace of the corresponding matrix (which is assumed to be square). For this definition, the inverse operator of $\hat{T}$ (i.e., $\hat{T}^{-1}$) does not need to exist. The trace of a matrix is found by adding up all of the diagonal elements of the matrix

$$\text{Tr}(\hat{T}) \equiv \text{Tr} \begin{bmatrix} T_{11} & T_{12} & \cdots \\ T_{21} & T_{22} & \\ \vdots & & \ddots \end{bmatrix} = \sum_n T_{nn}$$

If the basis for V is $B_v = \{|n\rangle\}$, then the trace of an operator can also be written as

$$\text{Tr}(\hat{T}) \equiv \sum_n \langle n | \hat{T} | n \rangle = \sum_n T_{nn}$$

The trace of an operator $\hat{T}$ is the sum of the diagonal elements of the matrix $\underline{T}$. The trace for an operator acting in a space V with a continuous basis set $B = \{|k\rangle\}$ has the form

$$\text{Tr}(\hat{T}) = \int dk \langle k | \hat{T} | k \rangle$$

which again represents a generalized summation over diagonal matrix elements.

The trace is extremely important in quantum mechanics for calculating averages using the density operator. As a comment, for $T: V \rightarrow W$ the spaces V and W can be fundamentally different types. V might be a 3-D Euclidean space while W can be a function space.

Here are some important properties for the Trace. Assume that the operators $\hat{A}, \hat{B}, \hat{C}$ have a domain and range within a single vector space V with basis vectors $B_v = \{|a\rangle\}$.

$$1. \text{Tr}(\hat{A}\hat{B}) = \text{Tr}(\hat{B}\hat{A})$$

This is easy to see by starting with the basic definition of trace

$$\text{Tr}(\hat{A}\hat{B}) = \sum_n \langle n | \hat{A}\hat{B} | n \rangle = \sum_n \langle n | \hat{A} | m \rangle \langle m | \hat{B} | n \rangle = \sum_{nm} \langle n | \hat{A} | m \rangle \langle m | \hat{B} | n \rangle$$

Next, use the fact that $\langle n | \hat{A} | m \rangle, \langle m | \hat{B} | n \rangle$ are just numbers, to commute them to get

$$(\hat{A}\hat{B}) = \sum_{nm} \langle n | \hat{A} | m \rangle \langle m | \hat{B} | n \rangle = \sum_{nm} \langle m | \hat{B} | n \rangle \langle n | \hat{A} | m \rangle = \sum_m \langle m | \hat{B} \hat{A} | m \rangle = \text{TR}(\hat{B}\hat{A})$$

where the closure relation is used to obtain the fourth term.

$$2. \text{TR}(ABC) = \text{TR}(BCA) = \text{TR}(CAB).$$

3. The trace of the operator $\hat{T}$ is “independent” of the chosen basis set as will be shown later.

The proof is similar to the one for the determinant.

Example 3.15

Find the trace of the operator $\hat{T} = \sum_{ab} T_{ab} |\phi_a\rangle \langle \phi_b|$, which the next section shows to be the basis vector expansion of the operator $\hat{T}$. We will see that the numbers T_{ab} are the matrix elements. For the present case, we assume $\hat{T}: V \rightarrow V$ where V has the basis $B_v = \{|\phi_a\rangle\}$.

SOLUTION

The trace of

$$\hat{T} = \sum_{ab} T_{ab} |\phi_a\rangle \langle \phi_b|$$

can be found by using the basic definition of trace given in the previous formula.

$$\begin{aligned} \text{Tr}(\hat{T}) &= \sum_c \langle \phi_c | \hat{T} | \phi_c \rangle = \sum_c \langle \phi_c | \left(\sum_{ab} T_{ab} |\phi_a\rangle \langle \phi_b| \right) | \phi_c \rangle \\ &= \sum_c \sum_{ab} T_{ab} \langle \phi_c | \phi_a \rangle \langle \phi_b | \phi_c \rangle = \sum_c \sum_{ab} T_{ab} \delta_{ac} \delta_{bc} = \sum_c T_{cc} \end{aligned}$$

which is a sum over all diagonal elements as expected. Apparently, the trace can be calculated for an operator $T: V \rightarrow W$ so long as $\dim(V) = \dim(W)$.

Example 3.16

Find the trace of the following composite operator assuming $\hat{A}: W \rightarrow V$ and $\hat{B}: V \rightarrow W$ where $V = \text{Sp}\{|\phi_m\rangle\}$ and $W = \text{Sp}\{|\psi_n\rangle\}$

$$\hat{O} = \hat{A}\hat{B}$$

SOLUTION

In this case, the operator maps V into itself and so one takes the trace using the basis vectors of V .

$$\text{Tr}(\hat{O}) = \sum_m \langle \phi_m | \hat{A}\hat{B} | \phi_m \rangle = \sum_{m,n} \langle \phi_m | \hat{A} | \psi_n \rangle \langle \psi_n | \hat{B} | \phi_m \rangle = \sum_{m,n} \hat{A}_{mn} \hat{B}_{nm}$$

where the closure relation on W was inserted to obtain the second summation.

Example 3.17

Find the trace of an operator $\hat{O}$ that maps a direct product space W into itself.

SOLUTION

Suppose $W = \text{Sp}\{|m n\rangle\}$ then

$$\text{Tr}(\hat{O}) = \sum_{a,b} \langle ab | \hat{O} | ab \rangle$$

The double summation occurs since each basis vector $|a b\rangle$ is characterized by two parameters a, b .

3.3.6 TRANSPOSE AND HERMITIAN CONJUGATE OF A MATRIX

The transpose operation means to interchange elements across the diagonal. For example

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

This is sometimes written as

$$(R^T)_{ab} = R_{ba} \quad (3.28a)$$

Note the interchange of the indices a and b . Sometimes this is also written as

$$R_{ab}^T = R_{ba} \quad (3.28b)$$

The Hermitian conjugate (i.e., the adjoint) of the matrix requires the complex conjugate so that

$$(R^+)_{ab} = R_{ba}^* \quad (3.28c)$$

One should note that R_{ab} refers to a single number. Sometimes people say that the R_{ab} refers to the entire matrix but they mean the entire collection $\{R_{ab}\}$ refers to the entire matrix (along with the matrix properties). Writing R_{ab} as a “number without reference to the matrix” would provide $R_{ab}^+ = R_{ab}^*$ since the adjoint of a number is the complex conjugate. The notation in Equation 3.28a through c indicates the “ a, b element” of the matrix.

3.4 OPERATOR SPACE

Linear operators have representations other than the matrix one. Perhaps the most conceptually useful representation treats the linear operator as a vector in a vector space for which it has a basis vector expansion. Such a representation clearly shows mathematical structure without burdensome detail sometimes unnecessary for calculations. The notion of a Hilbert space of operators requires an inner product that in turn, gives rise to the idea of the “length” of an operator as well as “angles” between operators.

3.4.1 CONCEPTS AND SECTION SUMMARY

Consider a vector space V with basis vectors $|\phi_a\rangle$ for $a=1, 2, \dots, N=\text{Dim}(V)$. The set of linear operators $L = \{\hat{T}: V \rightarrow V\}$ forms a vector space with basis set $B_L = \{|\phi_a\rangle\langle\phi_b|\}$ where $a, b=1, 2, \dots, N$. For this discrete case, the dimension of the space L must be N^2 . As will be shown in Section 3.4.2, every linear operator $\hat{T}$ in the set L can be written as a linear combination over a basis set of the form

$$\hat{T} = \sum_{ab} T_{ab} |\phi_a\rangle\langle\phi_b| \quad (3.29)$$

where T_{ab} appear as the components of the vector (i.e., expansion coefficients of the summation). One imagines L to be a vector space with basis vectors as shown in Figure 3.10 for example. The components T_{ab} can easily be seen to be the same as the matrix elements by operating on $\hat{T} = \sum_{ij} T_{ij} |\phi_i\rangle\langle\phi_j|$ with $\langle\phi_a|$ and $|\phi_b\rangle$ and using the orthonormality of the basis for V to find $T_{ab} = \langle\phi_a|\hat{T}|\phi_b\rangle$. The proof that the set L constitutes a vector space follows from a simple application of the basic definition for linear operators in Section 3.2. Needless to say, each basis vector in B_L lives in the space L and in a sense, represents the simplest operators in the space L . The reader has seen a similar basis vector expansion for the unit operator $\hat{1} = \sum_a |\phi_a\rangle\langle\phi_a|$.

The basis expansion of the operator (Equation 3.29) has many advantages over the matrix representation. First, all of the “parts” of the operator are present including the range represented by the kets and the domain represented by the bras, as well as the mixture of the fundamental operators (i.e., the basis vectors) through the components T_{ab} . Second, this representation gives a sense of the transformation (i.e., mapping) properties of the operator because of the particular combination of kets and bras in the basis set. For example, the fundamental operator $|\phi_2\rangle\langle\phi_1|$ maps

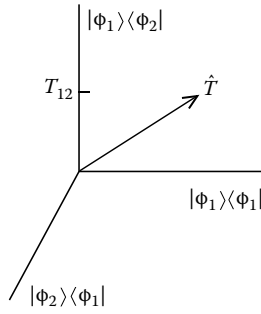


FIGURE 3.10 Example conceptual diagram showing the operator as a vector and the basis vectors. The matrix element T_{12} appears as a component of the vector.

$|\phi_1\rangle$ into $|\phi_2\rangle$ as easily seen by calculating the sequence $\{|\phi_2\rangle\langle\phi_1||\phi_1\rangle = |\phi_2\rangle$ using the orthogonality of the basis for V namely $\langle\phi_a|\phi_b\rangle = \delta_{ab}$. The combinations of the form $|\phi_i\rangle\langle\phi_j|$ can be read from right to left and shows that the vector $|\phi_j\rangle$ will be mapped into the vector $|\phi_i\rangle$. Third, the basis expansion shows all of the possible mappings by the operator. One can see how the operator has the possible mappings built right into it. On the other hand, the matrix representation provides an easy method for calculating.

The next few sections of discussion will show how the basis vector expansion of the operator follows from the basic definition of the matrix in Section 3.2. The discussion demonstrates an inner product for the operator space. One will find that the inner product is not unique although it never is unique anyway. For example, the dot product could be changed just by requiring an extra constant multiplying the results. First, however, we complete the present discussion with examples that will become more familiar later in the book.

Example 3.18

For the linear operator $\hat{T}: V \rightarrow V$ find an operator that maps the basis vectors as follows:

$$|1\rangle \rightarrow |2\rangle \quad \text{and} \quad |2\rangle \rightarrow -|1\rangle$$

SOLUTION

Form the following two combinations: $|2\rangle\langle 1|$ and $-|1\rangle\langle 2|$. Notice how these combinations map the domain vector into the range vector by the association of the corresponding kets and bras. One can see the mappings do in fact work:

$$\begin{aligned} \{|2\rangle\langle 1||1\rangle &= |2\rangle\langle 1|1\rangle = |2\rangle \\ \{-|1\rangle\langle 2||2\rangle &= -|1\rangle\langle 2|2\rangle = -|1\rangle \end{aligned}$$

We therefore speculate that the desired operator must be

$$\hat{T} = |2\rangle\langle 1| - |1\rangle\langle 2|$$

The reader should try the operator on both basis vectors. Try it on the first basis vector

$$\hat{T}|1\rangle = (|2\rangle\langle 1| - |1\rangle\langle 2|)|1\rangle = |2\rangle$$

The transformation T describes a rotation by 90° . The mapping of the basis vectors defines unique operator.

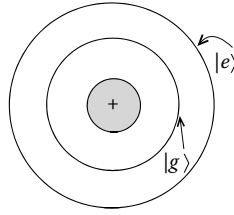


FIGURE 3.11 Cartoon drawing of a two-level atom.

Example 3.19

A two-level atom has two possible electron states labeled $|e\rangle$ and $|g\rangle$ which correspond to the first excited and ground state respectively (Figure 3.11). Find an operator that describes the absorption of light by the atom.

SOLUTION

The Hamiltonian has the form $\hat{\mathcal{H}} = c_1 |e\rangle\langle g|$ where, as will be seen later, c_1 depends on other operators since the absorption of the photon must also be described. This particular form of the operator shows the changes that the electron will undergo when the atom absorbs light. Reading from right to left, shows that the electron will be promoted from the ground state $|g\rangle$ to the excited state $|e\rangle$. The c_1 in the operator must account for the fact that a photon will be absorbed. The interaction Hamiltonian $\hat{\mathcal{H}}$ will have the form $\hat{\mathcal{H}} = c_2 \hat{a} |e\rangle\langle g|$ where the operator “ $\hat{a}$ ” is the annihilation operator for the photon field. The annihilation operator removes one photon from the incident light beam while $|g\rangle$ essentially removes one electron from the ground state and $|e\rangle$ makes the electron reappear in the excited state. As a final comment, notice how the state vectors (i.e., the actual vectors $|g\rangle$ and $|e\rangle$, not the operator) represent the state of the electron in the atom.

3.4.2 BASIS EXPANSION OF A LINEAR OPERATOR

We now demonstrate how the basic definition of the matrix leads to the representation of the linear operators as a summation over the basis vectors. We apply the procedure to an operator acting (1) within a single space with a discrete basis, (2) between two distinct spaces with discrete basis, and on (3) spaces with continuous basis sets.

First, consider the case of an operator $T: V \rightarrow V$ with the vector space V having basis set $B_v = \{|a\rangle = |\phi_a\rangle\}$ and $\text{Dim}(V) = n$. The result of $\hat{T}$ operating on one of the basis vectors $|b\rangle$ can be written as

$$\hat{T}|b\rangle = \sum_a T_{ab} |a\rangle$$

where T_{ab} represents the matrix elements. We want to isolate the operator $\hat{T}$ by producing the resolution of unity on the left-hand side. To this end, multiply this last equation by $\langle b|$ from the right, to find

$$\hat{T}|b\rangle\langle b| = \sum_a T_{ab} |a\rangle\langle b|$$

Now sum both sides over the index b

$$\hat{T} \sum_b |b\rangle\langle b| = \sum_{a,b} T_{ab} |a\rangle\langle b|$$

where $\hat{T}$ moves past the summation since $\hat{T}$ is linear. The closure relation $\sum_b |b\rangle\langle b| = \hat{1}$ for the vector space V provides

$$\hat{T} = \sum_{a,b} T_{ab} |a\rangle\langle b| \quad \text{or} \quad \hat{T} = \sum_{a,b} T_{ab} |\phi_a\rangle\langle\phi_b| \quad (3.30)$$

The dimension of the vector space of operators in this case must be n^2 . These basis vector representations of an operator have a form very reminiscent of the closure relation. In fact, we can recover the closure relation if the operator $\hat{T}$ is taken as the unit operator $\hat{T} = 1$ so that the matrix elements are $T_{ab} = \delta_{ab}$.

Similar to the previous discussion, the procedure can be applied to an operator $\hat{T}: V \rightarrow W$ acting between two distinct spaces. Assume that the two basis sets have the form $B_v = \{|\phi_i\rangle\}$ and $B_w = \{|\psi_j\rangle\}$. As before, start with the basic definition of the matrix $\hat{T}|\phi_b\rangle = \sum_a T_{ab} |\psi_a\rangle$, multiply by $\langle\phi_b|$ on the right-hand side to find the expression $\hat{T}|\phi_b\rangle\langle\phi_b| = \sum_a T_{ab} |\psi_a\rangle\langle\phi_b|$. The left-hand side of this expression involves vectors their duals from the same space V whereas the right-hand side has a mix from the two spaces. We can then isolate the operator $\hat{T}$ by summing over the index b on both sides to obtain $\hat{T} \sum_b |\phi_b\rangle\langle\phi_b| = \sum_{a,b} T_{ab} |\psi_a\rangle\langle\phi_b|$ and then using the closure relation on V , namely $\sum_b |\phi_b\rangle\langle\phi_b| = \hat{1}$. We obtain the desired final expression:

$$\hat{T} = \sum_{a,b} T_{ab} |\psi_a\rangle\langle\phi_b| \quad (3.31)$$

The formalism discussed to this point holds for either Euclidean or Function spaces so long as the vector spaces V and W have discrete basis sets. Interestingly, the basis set has the form $B_L = B_V \otimes B_W^+$ where B_W^+ is the basis for the dual space of W .

Finally, the operators $\hat{T}: V \rightarrow W$ acting between two different function spaces with continuous basis set $B_v = \{|\phi_k\rangle\}$ and $B_w = \{|\psi_{k'}\rangle\}$ have similar expansions except integrals instead of discrete summations. For example, these basis sets might be the Fourier transform sets with k and k' representing wave vectors. The operator $\hat{T}$ maps a basis vector such as $|\phi_k\rangle$ into a linear combination of basis vectors in space W to produce

$$\hat{T}|\phi_k\rangle = \int dk' T(k', k) |\psi_{k'}\rangle \quad (3.32a)$$

where $T(k', k) = T_{k',k}$. As before, we want to use the resolution of unity $\int dk |\phi_k\rangle\langle\phi_k| = 1$ for vector space V to isolate the operator. Multiply both sides on the right by $\langle\phi_k|$, integrate over the continuous parameter k , to find

$$\int dk \hat{T} |\phi_k\rangle\langle\phi_k| = \int dk \int dk' T(k', k) |\psi_{k'}\rangle\langle\phi_k|$$

The operator can be removed from the integral so that the resolution of unity can be used to obtain the desired final result.

$$\hat{T} = \int \int dk dk' T(k', k) |\psi_{k'}\rangle\langle\phi_k| \quad (3.32b)$$

Example 3.20

For the operator $\hat{T}: V \rightarrow V$ with the function space V having a discrete basis set $B_v = \{|\phi_a\rangle\}$ and the matrix of the operator having the form $T_{ab} = \delta_{ab}$, write Equation 3.31 in terms of coordinate x .

SOLUTION

Operator on both sides of Equation 3.31 with $\langle x'|$ and $|x''\rangle$ provides

$$\langle x'|\hat{T}|x''\rangle = \sum_{a,b} T_{ab}\phi_a(x')\phi_b^*(x'') = \sum_{a,b} \delta_{ab}\phi_a(x')\phi_b^*(x'') = \sum_a \phi_a(x')\phi_a^*(x'') = \delta(x' - x'')$$

Example 3.21

Find the matrix elements for the operator

$$\hat{\mathcal{H}} = 0|1\rangle\langle 1| + 0.5|1\rangle\langle 2| + 1|2\rangle\langle 1| + 3|2\rangle\langle 2|$$

by taking the inner products of both sides

$$\begin{aligned}\hat{\mathcal{H}}_{11} &= \langle 1|\hat{\mathcal{H}}|1\rangle = \langle 1|\{0|1\rangle\langle 1| + 0.5|1\rangle\langle 2| + 1|2\rangle\langle 1| + 3|2\rangle\langle 2|\}|1\rangle \\ &= 0\langle 1|1\rangle\langle 1|1\rangle + 0.5\langle 1|1\rangle\langle 2|1\rangle + \dots = 0\end{aligned}$$

$$\begin{aligned}\hat{\mathcal{H}}_{12} &= \langle 1|\hat{\mathcal{H}}|2\rangle = \langle 1|\{0|1\rangle\langle 1| + 0.5|1\rangle\langle 2| + 1|2\rangle\langle 1| + 3|2\rangle\langle 2|\}|2\rangle \\ &= 0\langle 1|1\rangle\langle 1|2\rangle + 0.5\langle 1|1\rangle\langle 2|2\rangle + \dots = 0.5\end{aligned}$$

$$\begin{aligned}\hat{\mathcal{H}}_{21} &= \langle 2|\hat{\mathcal{H}}|1\rangle = \langle 2|\{0|1\rangle\langle 1| + 0.5|1\rangle\langle 2| + 1|2\rangle\langle 1| + 3|2\rangle\langle 2|\}|1\rangle \\ &= 1\end{aligned}$$

$$\begin{aligned}\hat{\mathcal{H}}_{22} &= \langle 2|\hat{\mathcal{H}}|2\rangle = \langle 2|\{0|1\rangle\langle 1| + 0.5|1\rangle\langle 2| + 1|2\rangle\langle 1| + 3|2\rangle\langle 2|\}|2\rangle \\ &= 3\end{aligned}$$

Example 3.22

Using $\underline{A} = \begin{bmatrix} 4 & 0 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix}$ calculate (a) the basis vector expansion and (b) the inverse operator in the basis vector expansion.

SOLUTIONS

a. The basis vector expansion for the operator $\hat{A}: V \rightarrow V$ is

$$\hat{A} = \sum_{i,j=1}^3 A_{ij}|\hat{i}\rangle\langle j| = 4|1\rangle\langle 1| + 0|1\rangle\langle 2| + 4|1\rangle\langle 3| + 0 + 2|2\rangle\langle 2| + 2|2\rangle\langle 3| + 0 + 0 + 1|3\rangle\langle 3|$$

b. *Inverse operator*

The inverse matrix is

$$\underline{A}^{-1} = \begin{bmatrix} 0.25 & 0 & -1 \\ 0 & 0.5 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

which provides the following operator

$$\hat{A}^{-1} = \sum_{mn} (\hat{A}^{-1})_{mn} |m\rangle\langle n| = 0.25|1\rangle\langle 1| - 1|1\rangle\langle 3| + 0.5|2\rangle\langle 2| - 1|2\rangle\langle 3| + |3\rangle\langle 3|$$

An ambitious reader should show that $\hat{A}^{-1}\hat{A} = 1$ without resorting to matrix notation.

Example 3.23

If an operator $\hat{H}$ can be written as

$$\hat{H} = \sum_a E_a |a\rangle\langle a|$$

with $E_a \neq 0$ for every allowed index a , show that the inverse of the operator is given by

$$\hat{H}^{-1} = \sum_b \frac{1}{E_b} |b\rangle\langle b|$$

We need to show that $HH^{-1} = H^{-1}H = 1$ (both must be true). We will only show that

$$HH^{-1} = 1$$

Substituting the expansions for the operators gives us

$$\begin{aligned} \hat{H}\hat{H}^{-1} &= \sum_a E_a |a\rangle\langle a| \sum_b \frac{1}{E_b} |b\rangle\langle b| = \sum_{ab} \frac{E_a}{E_b} |a\rangle\langle a|b\rangle\langle b| \\ &= \sum_{ab} \frac{E_a}{E_b} |a\rangle\langle b| \delta_{ab} = \sum_{ab} \frac{E_a}{E_a} |a\rangle\langle a| = \sum_a |a\rangle\langle a| = 1 \end{aligned}$$

where of course the last line is obtained by closure on the Hilbert space.

Example 3.24

As will be discussed in a subsequent section, a Hermitian operator $\hat{\mathcal{H}}: V \rightarrow V$ can be “diagonalized” by choosing its eigenvectors $\{|e\rangle\}$ (normalized to unit length) as the basis set. Assume that the operator has the form $\hat{\mathcal{H}} = \sum_e E_e |e\rangle\langle e|$. Show (1) that if $|g\rangle, |h\rangle$ are basis vectors then $H_{gh} = \langle g|\hat{\mathcal{H}}|h\rangle = E_g \delta_{gh}$ (definition of a diagonal matrix) and (2) $\hat{\mathcal{H}}|g\rangle = E_g |g\rangle$.

SOLUTION

1. Apply $\langle g|$ and $|h\rangle$ to the operator and use orthonormality to obtain

$$H_{gh} = \langle g|\hat{\mathcal{H}}|h\rangle = \sum_e E_e \langle g|e\rangle\langle e|h\rangle = \sum_e E_e \delta_{ge} \delta_{eh} = E_g \delta_{gh}$$

2. Apply the operator to the vector $|g\rangle$ and use orthonormality to find

$$\hat{\mathcal{H}}|g\rangle = \sum_e E_e |e\rangle\langle e|g\rangle = \sum_e E_e |e\rangle \delta_{eg} = E_g |g\rangle$$

3.4.3 INTRODUCTION TO THE INNER PRODUCT FOR A HILBERT SPACE OF OPERATORS

The notion of a Hilbert space of operators gives rise to the idea of the “length” of an operator as well as “angles” between operators. What does this “length” mean? What length would one assign to the unit operator or perhaps to an operator that doubles the length of every vector in its domain? One answer would be to assign unit length to the unit operator and perhaps a length of two to the

doubling operator. Many different “lengths” can be imagined depending on how one defines the inner product between operators.

Consider another point. Suppose that we know an operator $\hat{T}$ but not the expansion coefficients $T_{ij} = \beta_{ij}$ in the generalized expansion

$$\hat{T} = \sum_{ij} \beta_{ij} |\phi_i\rangle \langle \phi_j| = \sum_{ij} \beta_{ij} \hat{Z}_{ij} \quad (3.33)$$

where $B_L = \{\hat{Z}_{ij} = |\phi_i\rangle \langle \phi_j|\}$ represents the basis vectors for the operator space L . How can we find a specific component $T_{ab} = \beta_{ab}$? One method would be to apply $\langle \phi_a|$ on the left-hand sides and $|\phi_b\rangle$ on the right-hand sides. However, Chapter 2 shows that components of vectors can be found by applying a single inner product. In the case of the Hilbert space of linear operators L with the summation in Equation 3.33, we need to define an inner product between operators to apply the vector formalism developed in Chapter 2. We would like to project the operator $\hat{T}$ onto the basis vector $\hat{Z}_{ab}$ to find $\beta_{ab} = \langle \hat{Z}_{ab} | \hat{T} \rangle$. The inner product leads to the orthonormality of the basis set $B_L = \{\hat{Z}_{ab} = |\phi_a\rangle \langle \phi_b|\}$.

To discuss orthonormality of B_L , an inner product must be defined. To get a clue as to how to define the inner product, consider the basis set for linear operators mapping V into W given by $\hat{Z}_{ab}: V \rightarrow W$. The inner product will need to combine basis vectors to produce a number. A combination of the form $\hat{Z}_{ab}\hat{Z}_{cd}$ is not defined since the first operator $\hat{Z}_{cd}$ would produce a vector in W but the second operator $\hat{Z}_{ab}$ can only operate on one in V . So one can reverse the mapping to produce one from W to V by using the adjoint to reverse the order of the bra and ket to obtain $\hat{Z}_{ab}^+: W \rightarrow V$. Then products of the form

$$\hat{Z}_{ab}^+ \hat{Z}_{cd} = (|\phi_b\rangle \langle \psi_a|)(|\psi_c\rangle \langle \phi_d|) \quad (3.34)$$

map the vector space V into the same space V (i.e., $\hat{Z}_{ab}^+ \hat{Z}_{cd}: V \rightarrow V$). However, we need a complex number as the value for the inner product rather than a vector as would be produced by Equation 3.34. One suspects that it will be the inner products on the individual spaces that give rise to the inner product for $\text{Sp}(B_L)$. Equation 3.34 already has an inner product for W which produces a complex number, but it still needs one on V . To solve two problems at once, namely the need for complex numbers rather than vectors and the need for an inner product on V , one needs to move $|\phi_b\rangle$ from the front to the back. Taking the trace over V allows one to accomplish this. If $|n\rangle$ (i.e., $|\phi_n\rangle$) represents a basis vector then

$$\begin{aligned} \text{Tr}(\hat{Z}_{ab}^+ \hat{Z}_{cd}) &= \sum_n \langle n | (|\phi_b\rangle \langle \psi_a|)(|\psi_c\rangle \langle \phi_d|) | n \rangle = \sum_n \langle \psi_a | \psi_c \rangle (\langle \phi_d | n \rangle \langle n | \phi_b \rangle) \\ &= \langle \psi_a | \psi_c \rangle \langle \phi_d | \phi_b \rangle = \delta_{ac} \delta_{bd} \end{aligned}$$

where the second summation follows by moving the complex numbers, and the third result follows from the closure relation on V . Of course, one could use orthonormality on the first summation to obtain the same result.

The reader should realize the difference between single objects such as $|\psi_m\rangle$, $\langle \phi_n|$ and those of the form $|\psi_m\rangle \langle \phi_n|$. The $|\psi_m\rangle$ and $\langle \phi_n|$ are usually thought of as vectors. Yes, $\langle \phi_n|$ is an operator (i.e., a projector), but it is considered elementary and has the mapping $\langle \phi_n|: V \rightarrow \mathbb{C}$ where $\mathbb{C}$ is the set of complex numbers. Operators such as $|\psi_m\rangle \langle \phi_n|$ are more complicated. Yes, they are typically thought of as “operators” with the mapping $|\psi_m\rangle \langle \phi_n|: V \rightarrow W$ but (as a second thought) they are also vectors in the vector space L .

Section 3.4.4 shows that the proposed inner product between operators, which relies on the definition for the inner product within the vector spaces V and W ,

$$\langle \hat{L}_1 | \hat{L}_2 \rangle = \text{Tr}(\hat{L}_1^+ \hat{L}_2) \quad (3.35)$$

does in fact satisfy all of the requirements for an inner product found in Section 2.1. We also see that the basis vectors (i.e., basis operators) $B_L = \{\hat{Z}_{ab} = |\psi_a\rangle\langle\psi_b|\}$ are orthonormal based on this definition. One can also show the equivalence between the “length” of the operator $\hat{O}$ corresponding to the trace definition in Equation 3.35 and the magnitude of the image vector $\hat{O}|v\rangle$ where $|v\rangle = \sum_n v_n |n\rangle$ and $|v_n|^2 = 1$. This second definition shows that the length of the operator has a direct relation to how it maps the vectors. An operator that doubles the length of the vector $|v\rangle$ can therefore be expected to have a length double that of the unit operator. The proof is left for the review exercises at the end of the chapter.

Example 3.25

Use the inner product of Equation 3.35, to find the length of the unit operator defined for a single vector space V of dimension N . Show the results using both the basis vector expansion and matrices.

SOLUTION

The basis vector expansion of the unit operator has the form $\hat{1} = \sum_{n=1}^N |n\rangle\langle n|$. Then

$$\langle \hat{1} | \hat{1} \rangle = \text{Tr}(\hat{1}^+ \hat{1}) = \text{Tr}(\hat{1}) = \text{Tr}\left(\sum_{m=1}^N |m\rangle\langle m|\right) = \sum_{n=1}^N \langle n | \left(\sum_{m=1}^N |m\rangle\langle m|\right) | n \rangle = \sum_{n=1}^N \delta_{nn} = N$$

The solution for the unit matrix gives the same results.

$$\langle \hat{1} | \hat{1} \rangle = \text{Tr}(\hat{1}^+ \hat{1}) = \text{Tr}(\hat{1}) = \text{Tr} \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \\ \vdots & & \ddots \end{bmatrix} = N$$

The end-of-chapter exercises show that if the inner product is redefined by dividing by N , then the inner product for the unit operators will produce the value of 1. The same revised definition then provides intuitively satisfying ‘lengths’ for other operators as well.

3.4.4 PROOF OF THE INNER PRODUCT

We now turn our attention to showing the proposed inner product

$$\langle \hat{A} | \hat{B} \rangle = \text{Trace}(\hat{A}^+ \hat{B})$$

satisfies the three requirements given in Section 2.1 and reproduced here:

1. $\langle f | g \rangle = \langle g | f \rangle^*$ with $f, g \in F$ and “*” denotes complex conjugate
2. $\langle \alpha f + \beta g | h \rangle = \alpha^* \langle f | h \rangle + \beta^* \langle g | h \rangle$ and $\langle h | \alpha f + \beta g \rangle = \alpha \langle h | f \rangle + \beta \langle h | g \rangle$ where $f, g, h \in F$ and $\alpha, \beta \in \mathbb{C}$, the complex numbers.
3. $\langle f | f \rangle \geq 0$ for every f and $\langle f | f \rangle = 0$ if and only if $f = 0$ (except at possibly a few points for the piecewise continuous functions $C_p[a, b]$).

For simplicity, assume that the space L consists of operators that map a vector space V into itself $L = \{\hat{A}: V \rightarrow V\}$.

Let us prove the first property. Assume that $\hat{A}, \hat{B}$ represent operators in the set L and that the vector space V has basis $\{|a\rangle\}$. Using the basis expansion of $\hat{A}, \hat{B}$

$$\hat{A} = \sum_{aa'} A_{aa'} |a\rangle \langle a'| \quad \hat{B} = \sum_{bb'} B_{bb'} |b\rangle \langle b'|$$

the complex conjugate of the candidate inner product can be written as

$$\begin{aligned} \langle \hat{A} | \hat{B} \rangle^* &= \text{Trace} \{ \hat{A}^\dagger \hat{B} \}^* = \text{Trace} \left\{ \left[\sum_{aa'} A_{aa'} |a\rangle \langle a'| \right]^\dagger \left[\sum_{bb'} B_{bb'} |b\rangle \langle b'| \right] \right\}^* \\ &= \text{Trace} \left\{ \sum_{aa' bb'} A_{aa'}^* B_{bb'} |a'\rangle \langle a| b\rangle \langle b'| \right\}^* = \left[\sum_{aa' bb'} A_{aa'}^* B_{bb'} \langle a| b\rangle \langle b'| a'\rangle \right]^* \\ &= \sum_{aa' bb'} A_{aa'} B_{bb'}^* \langle a| b\rangle^* \langle b'| a'\rangle^* = \sum_{aa' bb'} B_{bb'}^* A_{aa'} \langle b| a\rangle \langle a'| b'\rangle \\ &= \text{Trace} \sum_{aa' bb'} B_{bb'}^* |b'\rangle \langle b| A_{aa'} |a\rangle \langle a'| = \text{Trace} \left[\sum_{bb'} B_{bb'}^* |b\rangle \langle b'| \right]^\dagger \left[\sum_{aa'} A_{aa'} |a\rangle \langle a'| \right] \\ &= \langle \hat{B} | \hat{A} \rangle \end{aligned}$$

Notice that the third line uses the fact that $\langle a| b\rangle^* = \langle b| a\rangle$ since $\langle a| b\rangle$ is an inner product.

We can easily prove the second property because the trace of the sum equals the sum of the traces.

The third property can be demonstrated as follows:

$$\begin{aligned} \langle \hat{A} | \hat{A} \rangle &= \text{Trace} \left\{ \sum_{aa'} A_{aa'}^* |a'\rangle \langle a| \sum_{bb'} A_{bb'} |b\rangle \langle b'| \right\} = \sum_{aa' bb'} A_{aa'}^* A_{bb'} \langle a| b\rangle \langle b'| a'\rangle \\ &= \sum_{aa' bb'} A_{aa'}^* A_{bb'} \delta_{ab} \delta_{a'b'} = \sum_{ab} |A_{ab}|^2 \geq 0 \end{aligned}$$

3.4.5 BASIS FOR MATRICES

Writing $\hat{T}$ as a sum over basis vectors is essentially the same as writing a matrix as a sum of “unit” matrices. For example, a 4×4 matrix can be written as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + d \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

So for real matrices

$$\underline{T} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the “basis set” consists of

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

3.5 OPERATORS AND MATRICES IN DIRECT PRODUCT SPACE

The direct product (tensor product) space has important fundamental applications in the quantum theory. The dimension of the space has direct relation with the number of degrees of freedom of the system. Mathematically, the direct product space combines two or more vector spaces into one space. A direct product space can be formed from any type or number of vector spaces. The present section focuses primarily on the operator and matrices related to the direct product space.

3.5.1 REVIEW OF DIRECT PRODUCT SPACES

Vector spaces V and W can be combined into product spaces with basis vectors of the form

$$\{|\phi_i\rangle \otimes |\psi_j\rangle = |\phi_i\rangle|\psi_j\rangle = |\phi_i, \psi_j\rangle\}$$

where the individual spaces have the basis vectors

$$B_v = \{|\phi_i\rangle\} \quad B_w = \{|\psi_j\rangle\}$$

and the spaces V and W do not need to be the same size. The size of the direct product space $V \otimes W$ is given by

$$\text{Dim}[V \otimes W] = \text{Dim}(V) \text{Dim}(W).$$

when Dim is defined. The adjoint operator “+” maps the vector (ket)

$$|v, w\rangle \in V \otimes W$$

into the projection operator (bra) as

$$|v, w\rangle^+ = \langle v, w| = \langle v|\langle w|$$

where $\langle v, w| \in [V \otimes W]^+$. The basis set for the dual space is

$$\{\langle \phi_i, \psi_j| = \langle \phi_i|\langle \psi_j|\}$$

How are inner products formed? Recall that we must keep track of which dual space acts on which space. In particular, inner products can only be formed between V^+ and V , and between W^+ and W . Therefore, if

$$|v_1\rangle, |v_2\rangle \in V \quad |w_1\rangle, |w_2\rangle \in W$$

the inner product is

$$\langle v_1 w_1 | v_2 w_2 \rangle = \langle v_1 | v_2 \rangle \langle w_1 | w_2 \rangle \quad (3.36)$$

Of course, $\langle v_1 | v_2 \rangle$ and $\langle w_1 | w_2 \rangle$ are just complex numbers, and Equation 3.36 can also be written as

$$\langle v_1 w_1 | v_2 w_2 \rangle = \langle w_1 | w_2 \rangle \langle v_1 | v_2 \rangle$$

where the factors on the right-hand side have been reversed. The problems ask the reader to determine whether or not the direct product space forms a Hilbert space.

Vectors in the direct product $V \otimes W$ space do not necessarily factor into a unique set of vectors consisting of a vector from V and another from W . For example, the basis vector $|1, 1\rangle = |1\rangle|1\rangle$ can be alternatively written as $|1, 1\rangle = (0.5|1\rangle)(2|1\rangle)$. This lack of unique factoring becomes important for the quantum theory.

3.5.2 OPERATORS

Operators $\hat{O}$ can operate between direct product spaces such as $\hat{O}: V \otimes W \rightarrow X \otimes Y$ or within a given direct product space such as $\hat{O}: V \otimes W \rightarrow V \otimes W$. For simplicity, we consider the second case in this section.

We might have an operator $\hat{O}^{(V)}: V \rightarrow V$ and another operator $\hat{O}^{(W)}: W \rightarrow W$. The direct product of the two operators $\hat{O} = \hat{O}^{(V)} \otimes \hat{O}^{(W)}$ maps the direct product space $V \otimes W$ into itself. To find the image of the operator $\hat{O}$ when acting on a vector $|v\rangle|w\rangle$ in the product space, we just need to remember that $\hat{O}^{(V)}$ operates only on vectors in V and $\hat{O}^{(W)}$ operates only on vectors in W . Therefore, we have

$$\hat{O}|v\rangle|w\rangle = \hat{O}^{(V)}\hat{O}^{(W)}|v\rangle|w\rangle = \hat{O}^{(V)}|v\rangle\hat{O}^{(W)}|w\rangle = |x\rangle|y\rangle$$

where $|x\rangle|y\rangle \in V \otimes W$. The inner product behaves in a similar manner.

$$\langle q|\langle r|\hat{O}|v\rangle|w\rangle = \langle q|\langle r|\hat{O}^{(V)}\hat{O}^{(W)}|v\rangle|w\rangle = \langle q|\hat{O}^{(V)}|v\rangle\langle r|\hat{O}^{(W)}|w\rangle$$

where $|q\rangle \in V$, $|r\rangle \in W$, and $\langle q|\langle r|$ is a projector in the dual space $(V \otimes W)^+ = W^+ \otimes V^+ = V^+ \otimes W^+$ where the last relation follows if we do not care about the order.

Another notation is quite common in the literature. It helps to distinguish between ordinary multiplication and the direct product type; this distinction becomes especially important for writing the matrix of a vector in the direct product space. If we have an operator $\hat{O}^{(V)}: V \rightarrow V$ then we can use the unit operator on W to write $\hat{O}^{(V)} \otimes \hat{1}: V \otimes W \rightarrow V \otimes W$ then $\{\hat{O}^{(V)} \otimes \hat{1}\}\{|v\rangle \otimes |w\rangle\} = \{\hat{O}^{(V)}|v\rangle\} \otimes \{\hat{1}|w\rangle\}$. More generally, we can write

$$\{\hat{O}^{(V)} \otimes \hat{O}^{(W)}\}\{|v\rangle \otimes |w\rangle\} = \{\hat{O}^{(V)}|v\rangle\} \otimes \{\hat{O}^{(W)}|w\rangle\}$$

What about the addition of two operators?

$$\{\hat{O}^{(V)} + \hat{O}^{(W)}\}\{|v\rangle \otimes |w\rangle\} \equiv \{\hat{O}^{(V)} \otimes \hat{1} + \hat{1} \otimes \hat{O}^{(W)}\}\{|v\rangle \otimes |w\rangle\}$$

Distributing terms gives

$$\{\hat{O}^{(V)} + \hat{O}^{(W)}\}\{|v\rangle \otimes |w\rangle\} = \{\hat{O}^{(V)} \otimes \hat{1}\}\{|v\rangle \otimes |w\rangle\} + \{\hat{1} \otimes \hat{O}^{(W)}\}\{|v\rangle \otimes |w\rangle\}$$

Simplifying gives

$$\{\hat{O}^{(V)} + \hat{O}^{(W)}\}\{|v\rangle \otimes |w\rangle\} = \{\hat{O}^{(V)}|v\rangle\} \otimes |w\rangle + |v\rangle \otimes \{\hat{O}^{(W)}|w\rangle\}$$

as expected. The notation helps signify that the addition between vectors must be on the direct product space.

3.5.3 MATRICES OF DIRECT PRODUCT OPERATORS

The operators $\hat{O}$ acting on the direct product space $V \otimes W$ map one basis vector into another. Assume the basis vectors for the spaces can be written as

$$B_v = \{|\phi_1\rangle, |\phi_2\rangle\} \quad B_w = \{|\psi_1\rangle, |\psi_2\rangle\} \quad B_{V \otimes W} = \{|\phi_a\rangle|\psi_b\rangle\}$$

The matrix of $\hat{O}$ can be defined in the usual way. The operator maps each basis vector into another vector in the direct product Hilbert space. The resulting vector must be a sum over the basis vectors in the space.

$$\hat{O}|\phi_c \psi_d\rangle = \sum_{\alpha, \beta} O_{\alpha, \beta; c, d} |\phi_\alpha \psi_\beta\rangle$$

or, taking the inner product using the projection operator $\langle \phi_c, \psi_d |$, gives

$$\langle \phi_a \psi_b | \hat{O} | \phi_c \psi_d \rangle = \sum_{\alpha, \beta} O_{\alpha, \beta; c, d} \langle \phi_a \psi_b | \phi_\alpha \psi_\beta \rangle = \sum_{\alpha, \beta} O_{\alpha, \beta; c, d} \delta_{a\alpha} \delta_{b\beta} = O_{ab; cd}$$

Matrix Notation

We investigate two different matrix notations. One follows most naturally from the basis vector expansion while, the second more conventional notation has computational benefits.

The basis vector expansion contains the matrix of the operator $\hat{O}$

$$\hat{O} = \sum_{abcd} O_{ab; cd} |\phi_a, \psi_b\rangle \langle \phi_c, \psi_d| = \sum_{abcd} O_{ab; cd} |\phi_a\rangle |\psi_b\rangle \langle \phi_c| \langle \psi_d| \quad (3.37)$$

Note the order of the indices. To write the matrix of an operator on direct product spaces, we need a convention for the indices. For simplicity, suppose the two vectors spaces V and W have dimension 2 (i.e., $\text{Dim}(V) = \text{Dim}(W) = 2$). An operator $\hat{O}$ written in the basis expansion Equation 3.37 becomes

$$\begin{aligned} \hat{O} = & O_{11,11} |\phi_1\rangle |\psi_1\rangle \langle \phi_1| \langle \psi_1| + O_{11,12} |\phi_1\rangle |\psi_1\rangle \langle \phi_1| \langle \psi_2| \\ & + O_{11,21} |\phi_1\rangle |\psi_1\rangle \langle \phi_2| \langle \psi_1| + O_{11,22} |\phi_1\rangle |\psi_1\rangle \langle \phi_2| \langle \psi_2| \\ & + O_{12,11} |\phi_1\rangle |\psi_2\rangle \langle \phi_1| \langle \psi_1| + O_{12,12} |\phi_1\rangle |\psi_2\rangle \langle \phi_1| \langle \psi_2| \\ & + \dots \end{aligned}$$

The matrix in this case could be written as

$$\begin{array}{c} \uparrow \\ a, b \\ \downarrow \end{array} \begin{array}{c} \leftarrow c, d \rightarrow \\ \begin{bmatrix} O_{11,11} & O_{11,12} & O_{11,21} & O_{11,22} \\ O_{12,11} & O_{12,12} & O_{12,21} & O_{12,22} \\ O_{21,11} & O_{21,12} & O_{21,21} & O_{21,22} \\ O_{22,11} & O_{22,12} & O_{22,21} & O_{22,22} \end{bmatrix} \end{array}$$

Conventional Matrix Notation

Although the previous convention provides a perfectly fine representation of the direct product matrix, another index convention proves more useful when calculating the direct product matrix from two other matrices rather than from other operators. To this end, let us rearrange the basis vectors in Equation 3.37 and write

$$\hat{O} = \sum_{abcd} O_{ab, cd} [|\phi_a\rangle \langle \phi_c|] [|\psi_b\rangle \langle \psi_d|]$$

Interchanging dummy indices b and c produces

$$\hat{O} = \sum_{abcd} O_{ac, bd} [|\phi_a\rangle \langle \phi_b|] [|\psi_c\rangle \langle \psi_d|] \quad (3.38)$$

When necessary, we make the index convention that, for each a and b , the summation is performed first over d and then over c . The object $O_{ac,bd}$ is a single number (an element of a matrix). The collection of $O_{ac,bd}$ of complex numbers forms a matrix that cannot, most of the time, be divided into the product of two matrices.

We now show how the convention works using two cases.

$$\text{Case 1: } \hat{O} = \hat{O}^{(V)} \otimes \hat{O}^{(W)} = \hat{O}^{(V)} \hat{O}^{(W)}$$

This case supposes that the operator $\hat{O}$ operating on the direct product space $V \otimes W$ comes from two operators $\hat{O} = \hat{O}^{(V)} \hat{O}^{(W)}$ where the individual operators $\hat{O}^{(V)}$ and $\hat{O}^{(W)}$ map a single space into itself according to $\hat{O}^{(V)}: V \rightarrow V$ and $\hat{O}^{(W)}: W \rightarrow W$. For simplicity, we again assume $\text{Dim}(V) = \text{Dim}(W) = 2$. The individual operators can be written as basis vector expansions

$$\hat{O}^{(V)} = \sum_{ab} O_{ab}^{(V)} |\phi_a\rangle \langle \phi_b| \quad \text{and} \quad \hat{O}^{(W)} = \sum_{cd} O_{cd}^{(W)} |\psi_c\rangle \langle \psi_d|$$

As usual, we treat the collection of expansion coefficients $\{O_{ab}^{(V)}\}$ and $\{O_{cd}^{(W)}\}$ as matrices. The operator $\hat{O} = \hat{O}^{(V)} \hat{O}^{(W)}$ can now be written as

$$\hat{O} = \hat{O}^{(V)} \hat{O}^{(W)} = \sum_{ab} O_{ab}^{(V)} |\phi_a\rangle \langle \phi_b| \sum_{cd} O_{cd}^{(W)} |\psi_c\rangle \langle \psi_d| = \sum_{abcd} O_{ab}^{(V)} O_{cd}^{(W)} [|\phi_a\rangle \langle \phi_b|][|\psi_c\rangle \langle \psi_d|] \quad (3.39)$$

For each a, b , there exists a set of matrix elements O_{cd} . A comparison of Equations 3.39 and 3.37 shows the matrix elements of $\hat{O} = \hat{O}^{(V)} \hat{O}^{(W)}$ must be related to those for $\hat{O}^{(V)}$ and for $\hat{O}^{(W)}$ by $O_{ac,bd} = O_{ab}^{(V)} O_{cd}^{(W)}$. In matrix notation, this becomes

$$\underline{O} = \underline{O}^{(V)} \otimes \underline{O}^{(W)} = \begin{bmatrix} O_{11}^{(V)} & O_{12}^{(V)} \\ O_{21}^{(V)} & O_{22}^{(V)} \end{bmatrix} \otimes \begin{bmatrix} O_{11}^{(W)} & O_{12}^{(W)} \\ O_{21}^{(W)} & O_{22}^{(W)} \end{bmatrix}$$

This is not the usual “matrix” multiplication! The matrix on the right-hand side is multiplied into each element of the matrix on the left-hand side.

$$\underline{O} = \underline{O}^{(V)} \otimes \underline{O}^{(W)} = \begin{bmatrix} O_{11}^{(V)} \underline{O}^{(W)} & O_{12}^{(V)} \underline{O}^{(W)} \\ O_{21}^{(V)} \underline{O}^{(W)} & O_{22}^{(V)} \underline{O}^{(W)} \end{bmatrix} = \begin{bmatrix} O_{11}^{(V)} O_{11}^{(W)} & O_{11}^{(V)} O_{12}^{(W)} & O_{12}^{(V)} O_{11}^{(W)} & O_{12}^{(V)} O_{12}^{(W)} \\ O_{11}^{(V)} O_{21}^{(W)} & O_{11}^{(V)} O_{22}^{(W)} & O_{12}^{(V)} O_{21}^{(W)} & O_{12}^{(V)} O_{22}^{(W)} \\ O_{21}^{(V)} O_{11}^{(W)} & O_{21}^{(V)} O_{12}^{(W)} & O_{22}^{(V)} O_{11}^{(W)} & O_{22}^{(V)} O_{12}^{(W)} \\ O_{21}^{(V)} O_{21}^{(W)} & O_{21}^{(V)} O_{22}^{(W)} & O_{22}^{(V)} O_{21}^{(W)} & O_{22}^{(V)} O_{22}^{(W)} \end{bmatrix}$$

The product is also called the Kronecker matrix product. Of course each entry $O_{ab}^{(V)} O_{cd}^{(W)}$ is just a single number found by ordinary multiplication between numbers. The above matrix does illustrate the convention for the indices of $\underline{O}$.

As a check on the matrix multiplication for this case, calculate the matrix element of the direct product operator $\hat{O} = \hat{O}^{(V)} \hat{O}^{(W)}$. Recall that matrix elements involve the inner product of basis vectors as

$$\underline{O}_{ab,cd} = \langle \phi_a \psi_b | \hat{O}^{(V)} \hat{O}^{(W)} | \phi_c \psi_d \rangle = \sum_{ef} \langle \phi_a \psi_b | \hat{O}^{(V)} | \phi_e \psi_f \rangle \langle \phi_e \psi_f | \hat{O}^{(W)} | \phi_c \psi_d \rangle$$

which uses the closure relation for direct product space

$$1 = \sum_{ef} |\phi_e \psi_f\rangle \langle \phi_e \psi_f|$$

In the expression for $\underline{Q}_{ab,cd}$, note that $\hat{O}^{(V)}$ operates only on the basis set for V and similarly $\hat{O}^{(W)}$ operates only on the basis set for W . Therefore, the expression for $\underline{Q}_{ab,cd}$ becomes

$$\underline{Q}_{ab,cd} = \sum_{ef} \langle \phi_a | \hat{O}^{(V)} | \phi_e \rangle \delta_{bf} \langle \psi_f | \hat{O}^{(W)} | \psi_d \rangle \delta_{ec} = \langle \phi_a | \hat{O}^{(V)} | \phi_c \rangle \langle \psi_b | \hat{O}^{(W)} | \psi_d \rangle$$

as required.

Case 2: The operator $\hat{O}$ cannot be divided.

The last matrix given in case 1 provides a clue as to how $\underline{Q}$ should be written for the general case, namely

$$\underline{Q} = \begin{bmatrix} O_{11,11} & O_{11,12} & O_{12,11} & O_{12,12} \\ O_{11,21} & O_{11,22} & O_{12,21} & O_{12,22} \\ O_{21,11} & O_{21,12} & O_{22,11} & O_{22,12} \\ O_{21,21} & O_{21,22} & O_{22,21} & O_{22,22} \end{bmatrix}$$

With the index convention, matrices in direct product space can be multiplied together as usual. This case might hold since the operator $\hat{O}$ does not necessarily have a unique decomposition. For example, consider the operator $\hat{O}|vw\rangle = 2|vw\rangle$. The operator $\hat{O}$ can be decomposed in an infinite number of ways to form $\hat{O} = \hat{O}^{(V)} \otimes \hat{O}^{(W)}$ including the following two.

$$\left\{ \begin{array}{l} \hat{O}^{(V)}|v\rangle = 2|v\rangle \\ \hat{O}^{(W)}|w\rangle = |w\rangle \end{array} \right\} \quad \left\{ \begin{array}{l} \hat{O}^{(V)}|v\rangle = |v\rangle \\ \hat{O}^{(W)}|w\rangle = 2|w\rangle \end{array} \right\}$$

3.5.4 MATRIX REPRESENTATION OF BASIS VECTORS FOR DIRECT PRODUCT SPACE

Now let us show how the matrices multiply and define the unit vectors in the cross product space. Again for simplicity consider two 2-D Hilbert spaces V and W and use the product of two operators $\hat{O} = \hat{A}_v \hat{B}_w$ where the v and w indices refer to the originating Hilbert space in $V \otimes W$. Let us convert the operator equation

$$\hat{A}_v \hat{B}_w |v_\otimes\rangle = |v'_\otimes\rangle$$

where the subscript $\otimes$ indicates the vector comes from $V \otimes W$. Operating with $\langle a_v | \langle b_w |$ and inserting the closure relation $\sum_{c,d} |c_v d_w\rangle \langle c_v d_w| = \hat{1}$ produces

$$\sum_{c,d} \langle a_v | \langle b_w | \hat{A}_v \hat{B}_w | c_v d_w \rangle \langle c_v d_w | v_\otimes \rangle = \langle a_v b_w | v'_\otimes \rangle$$

We can write this in matrix notation as

$$\sum_{c,d} A_{a_v} \underbrace{c_v}_2 B_{b_w} \underbrace{d_w}_1 V_{c_v d_w} = V'_{a_v b_w}$$

The extra numbers under c, d indicate that we first sum over d and then over c . Writing this in matrix notation gives us

$$\begin{bmatrix} A_{11}^{(V)} B_{11}^{(W)} & A_{11}^{(V)} B_{12}^{(W)} & A_{12}^{(V)} B_{11}^{(W)} & A_{12}^{(V)} B_{12}^{(W)} \\ A_{11}^{(V)} B_{21}^{(W)} & A_{11}^{(V)} B_{22}^{(W)} & A_{12}^{(V)} B_{21}^{(W)} & A_{12}^{(V)} B_{22}^{(W)} \\ A_{21}^{(V)} B_{11}^{(W)} & A_{21}^{(V)} B_{12}^{(W)} & A_{22}^{(V)} B_{11}^{(W)} & A_{22}^{(V)} B_{12}^{(W)} \\ A_{21}^{(V)} B_{21}^{(W)} & A_{21}^{(V)} B_{22}^{(W)} & A_{22}^{(V)} B_{21}^{(W)} & A_{22}^{(V)} B_{22}^{(W)} \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} v'_{11} \\ v'_{12} \\ v'_{21} \\ v'_{22} \end{bmatrix} \quad (3.40)$$

Notice the order of the factors and the order of the indices in Equation 3.40. The column vectors must come from the direct product of two individual matrices. If $|v_{\otimes}\rangle = |r_v\rangle|s_w\rangle$ then we see

$$\begin{bmatrix} v_{11} \\ v_{12} \\ v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} r_1 s_1 \\ r_1 s_2 \\ r_2 s_1 \\ r_2 s_2 \end{bmatrix} = \begin{bmatrix} r_1 \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \\ r_2 \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \end{bmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \otimes \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \quad (3.41)$$

We therefore realize that the basis vectors can be represented by

$$\begin{aligned} |1\rangle &= |1\rangle_v |1\rangle_w \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ 0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\ |2\rangle &= |1\rangle_v |2\rangle_w \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ 0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \\ |3\rangle &= |2\rangle_v |1\rangle_w \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \\ |4\rangle &= |2\rangle_v |2\rangle_w \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \end{aligned}$$

3.6 COMMUTATORS AND ALGEBRA OF OPERATORS

Operators form more than a vector space, they also form an algebra that includes the multiplication (i.e., composition) of operators. Unlike addition, the multiplication of operators does not necessarily satisfy commutative properties. The degree of noncommutation can be measured by the commutator operator. Later sections will show that noncommutativity produces the Heisenberg uncertainty relation that forms one of the cornerstones for the quantum mechanics.

3.6.1 INITIAL DISCUSSION OF OPERATOR ALGEBRA

The set of linear operators forms a vector space. The vector space properties do not include operator multiplication (i.e., composition). Including a multiplication of operators expands the properties of operators and forms an algebra. In all but a few cases, the operators do not commute under multiplication. The noncommutivity manifests in nonzero “commutators” which play a primary role in the quantum theory. Some operators additionally form a group under multiplication that ensures the existence of an inverse operator.

The linear operators satisfy the following properties:

1. If $\hat{A}$, $\hat{B}$ are linear operators then so is $\hat{A}\hat{B}$.
2. There is an identity operator $\hat{I}$ such that $\hat{I}\hat{A} = \hat{A}\hat{I} = \hat{A}$
One should notice that if the operators act between different spaces then the unit operator has different definition depending on whether it operates on the right- or left-hand side.
3. The associative law holds $\hat{A}(\hat{B}\hat{C}) = (\hat{A}\hat{B})\hat{C}$.
4. In some cases, when the set of operators forms a group, for every operator $\hat{A}$ there exists an inverse operator $\hat{A}^{-1}$ such that $\hat{A}^{-1}\hat{A} = \hat{A}\hat{A}^{-1} = \hat{I}$. We will have significant interest in the unitary operators $\hat{u}$ that have inverse operators $\hat{u}^+$; these are essentially rotation operators in a complex Hilbert space.
5. The operators can be added and there exists an additive inverse along with the other vector space properties.
6. Scalar multiplication is defined as a carryover from the vector space properties

$$a\hat{A} = \hat{A}a$$

where a is a complex number.

7. The distributive law holds

$$\hat{A}(\hat{B} + \hat{C}) = \hat{A}\hat{B} + \hat{A}\hat{C} \quad \text{and} \quad (\hat{A} + \hat{B})\hat{C} = \hat{A}\hat{C} + \hat{B}\hat{C}$$

8. These properties use the definition that two operators $\hat{A}$ and $\hat{B}$ are equal $\hat{A} = \hat{B}$ if $\hat{A}|v\rangle = \hat{B}|v\rangle$ for every vector $|v\rangle$ in the vector space V .

Example 3.26

Show that the linear operator $\hat{A} = |1\rangle\langle 1| + |1\rangle\langle 2|$ does not have an inverse. Assume that the vector space has basis $\{|1\rangle, |2\rangle\}$.

SOLUTION

Notice the unit vectors $|1\rangle$ and $|2\rangle$ map into the same image vector $|1\rangle$ which means that the reverse function (i.e., inverse) would not be well defined in that it would not be able to uniquely map the original image vector $|1\rangle$ into a single preimage vector.

Example 3.27

Show that the operators $\hat{A}$ that map the xy plane into the z -axis and those operators $\hat{B}$ that map the xz plane into the y -axis do not commute.

SOLUTION

Pick two representative operators $\hat{A} = |3\rangle\langle 1| + |3\rangle\langle 2|$ and $\hat{B} = |2\rangle\langle 1| + |2\rangle\langle 3|$ to find $\hat{A}\hat{B} = |3\rangle\langle 1| + |3\rangle\langle 3|$ whereas $\hat{B}\hat{A} = |2\rangle\langle 1| + |2\rangle\langle 2|$.

3.6.2 INTRODUCTION TO COMMUTATORS

The “commutator” operator $\hat{A}\hat{B} - \hat{B}\hat{A}$ provides a measure of the amount by which the operators $\hat{A}, \hat{B}$ do not commute. Our theory of the universe vitally depends on the commutation and noncommutation of operators. The noncommutation of operators underlies all of quantum mechanics! It explains the differences between the classical and quantum worlds.

The previous section discussed the algebraic properties for the multiplication of operators and stated the fact that they do not need to commute. Two operators $\hat{A}$ and $\hat{B}$ commute when $\hat{A}\hat{B} = \hat{B}\hat{A}$ or equivalently $\hat{A}\hat{B} - \hat{B}\hat{A} = 0$. We represent the quantity $\hat{A}\hat{B} - \hat{B}\hat{A}$ by the “commutator” $[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}$. Therefore two operators $\hat{A}$ and $\hat{B}$ commute when $[\hat{A}, \hat{B}] = 0$.

Example 3.28

Show that rotations around the y -axis $\hat{R}_y(\theta_y)$ and those around the z -axis $\hat{R}_z(\theta_z)$ do not commute.

SOLUTION

One method to show this is to find a vector $\vec{v}$ and rotation angles that do not produce the same results for the two composite operations

$$\hat{R}_y(\theta_y)\hat{R}_z(\theta_z) \quad \text{and} \quad \hat{R}_z(\theta_z)\hat{R}_y(\theta_y)$$

For this purpose, consider rotations of 90° around each axis for the initial vector $\tilde{y}$. We find

$$\hat{R}_y(90)\hat{R}_z(90)\tilde{y} = \hat{R}_y(90)(-\tilde{x}) = \tilde{z} \quad \text{and} \quad \hat{R}_z(90)\hat{R}_y(90)\tilde{y} = \hat{R}_z(90)\tilde{y} = -\tilde{x}$$

The difference between the two resulting vectors provides a measure of the noncommutivity

$$[\hat{R}_y(90)\hat{R}_z(90) - \hat{R}_z(90)\hat{R}_y(90)]\tilde{y} = \tilde{z} + \tilde{x}$$

The quantity in “[]” represents another operator say $\hat{O}$. The closer is $\hat{O}$ to zero (i.e., the image vectors have nearly zero length for example) then the more nearly do the operators commute.

Example 3.29

Show $[x, \frac{d}{dx}] \neq 0$

We must treat the commutator

$$\left[x, \frac{d}{dx} \right]$$

as an “operator.” When calculating the commutator, it must operate on a function $f(x)$!

$$\left[x, \frac{d}{dx} \right] f = \left(x \frac{d}{dx} - \frac{d}{dx} x \right) f(x) = x \frac{df}{dx} - \frac{d}{dx} (xf) = x \frac{df}{dx} - \frac{dx}{dx} f - x \frac{df}{dx} = -f \neq 0$$

Notice that the derivative with respect to x operates on everything to the right.

As mentioned, noncommutivity distinguishes the quantum and classical worlds. Later sections show that if two operators $\hat{A}$ and $\hat{B}$ do not commute then $\hat{A}$ and $\hat{B}$ have an associated uncertainty relation:

$$\sigma_A \sigma_B \geq C > 0$$

where σ represents the standard deviation from probability theory. This last relation is a restatement of the Heisenberg uncertainty relation.

3.6.3 SOME COMMUTATOR THEOREMS

The commutators satisfy a number of properties. Let $\hat{A}, \hat{B}, \hat{C}$ be operators and let c represent a complex number.

0. $[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}$	1. $[\hat{A}, \hat{A}] = 0$	2. $[c, \hat{A}] = 0$
3. $[\hat{A}, \hat{B}] = -[\hat{B}, \hat{A}]$	4. $[\hat{A}, \hat{B} + \hat{C}] = [\hat{A}, \hat{B}] + [\hat{A}, \hat{C}]$	5. $[\hat{A} + \hat{B}, \hat{C}] = [\hat{A}, \hat{C}] + [\hat{B}, \hat{C}]$
6. $[\hat{A}, \hat{B}\hat{C}] = [\hat{A}, \hat{B}]\hat{C} + \hat{B}[\hat{A}, \hat{C}]$	7. $[\hat{A}\hat{B}, \hat{C}] = [\hat{A}, \hat{C}]\hat{B} + \hat{A}[\hat{B}, \hat{C}]$	8. $f = f(\hat{A}) \rightarrow [f(\hat{A}), \hat{A}] = 0$

Properties 1 through 7 can be easily proven by expanding the brackets and using the definition of the commutator. For example, property 6 is proved as follows:

$$[\hat{A}, \hat{B}]\hat{C} + \hat{B}[\hat{A}, \hat{C}] = (\hat{A}\hat{B} - \hat{B}\hat{A})\hat{C} + \hat{B}(\hat{A}\hat{C} - \hat{C}\hat{A}) = \hat{A}\hat{B}\hat{C} - \hat{B}\hat{C}\hat{A} = [\hat{A}, \hat{B}\hat{C}]$$

Functions of operators are defined through the Taylor expansion. Property 8 can be proved by Taylor expansion of the function. The Taylor expansion of a function of an operator has the form:

$$f(\hat{A}) = \sum_n c_n \hat{A}^n \quad \text{so that} \quad [f(\hat{A}), \hat{A}] = \left[\sum_n c_n \hat{A}^n, \hat{A} \right] = \sum_n c_n [\hat{A}^n, \hat{A}] = 0$$

where c_n can be a complex number and n is a nonnegative integer. The Taylor expansion of the operator originates in the usual Taylor expansion for a function $f(x)$. Once having written the series of $f(x)$, just replace x with the operator.

The following list of theorems can be proved by appealing to the properties of commutators, derivatives, and functions of operators.

THEOREM 3.1: Operator Expansion Theorem

The operator $\hat{O} = e^{x\hat{A}}\hat{B}e^{-x\hat{A}}$ can be written as

$$\hat{O} = e^{x\hat{A}}\hat{B}e^{-x\hat{A}} = \hat{B} + x[\hat{A}, \hat{B}] + \frac{x^2}{2!}[\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

We can prove this by writing a Taylor expansion of $\hat{O}(x)$ as

$$\hat{O}(x) = \hat{O}(0) + \left. \frac{\partial \hat{O}}{\partial x} \right|_{x=0} x + \frac{1}{2!} \left. \frac{\partial^2 \hat{O}}{\partial x^2} \right|_{x=0} x^2 + \dots$$

where

$$\hat{O}(0) = e^{x\hat{A}}\hat{B}e^{-x\hat{A}}\Big|_{x=0} = \hat{B}$$

and

$$\frac{\partial \hat{O}}{\partial x}\Big|_{x=0} = \frac{\partial}{\partial x} \left(e^{x\hat{A}}\hat{B}e^{-x\hat{A}} \right)_{x=0} = \hat{A}e^{x\hat{A}}\hat{B}e^{-x\hat{A}} - e^{x\hat{A}}\hat{B}e^{-x\hat{A}}\hat{A}\Big|_{x=0} = [\hat{A}, \hat{B}]$$

Higher-order derivatives can be similarly calculated. Putting all of the terms together provides the desired results

$$\hat{O} = e^{x\hat{A}}\hat{B}e^{-x\hat{A}} = \hat{B} + x[\hat{A}, \hat{B}] + \frac{x^2}{2!}[\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

THEOREM 3.2: Operator Expansion Theorem with Unity C-Number Factor

$$e^{\hat{A}}\hat{B}e^{-\hat{A}} = \hat{B} + [\hat{A}, \hat{B}] + \frac{1}{2!}[\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

This follows from the last theorem by setting $x = 1$.

THEOREM 3.3: Operator Expansion Theorem for a Constant Commutator

If $[\hat{A}, \hat{B}] = c$ where c represents a complex number then Theorem 3.1 provides

$$e^{x\hat{A}}\hat{B}e^{-x\hat{A}} = \hat{B} + cx$$

THEOREM 3.4: Product of Exponentials: Campbell–Baker–Hausdorff Theorem

If $\hat{A}, \hat{B}$ are two operators such that $[\hat{A}, [\hat{A}, \hat{B}]] = 0 = [\hat{B}, [\hat{A}, \hat{B}]]$ then

$$e^{x(\hat{A}+\hat{B})} = e^{x\hat{A}}e^{x\hat{B}}e^{-x^2[\hat{A}, \hat{B}]/2}$$

In particular, for $x = 1$ and $[\hat{A}, \hat{B}] = 0$ we get

$$e^{\hat{A}+\hat{B}} = e^{\hat{A}}e^{\hat{B}}$$

Notice that this is the usual law for adding exponential but it requires the operators to commute.

THEOREM 3.5: A Multiplication of Operators

$$\left[e^{x\hat{A}}\hat{B}e^{-x\hat{A}} \right]^n = e^{x\hat{A}}\hat{B}^ne^{-x\hat{A}}$$

The proof uses the fact that $e^{x\hat{A}}e^{-x\hat{A}} = e^{x\hat{A}-x\hat{A}} = 1$ where the exponents can be combined because they commute (see the Campbell–Baker–Hausdorff theorem—Theorem 3.4).

$$\left[e^{x\hat{A}}\hat{B}e^{-x\hat{A}} \right]^n = \left(e^{x\hat{A}}\hat{B}e^{-x\hat{A}} \right) \left(e^{x\hat{A}}\hat{B}e^{-x\hat{A}} \right) \left(e^{x\hat{A}}\hat{B}e^{-x\hat{A}} \right) \cdots = e^{x\hat{A}}\hat{B}^ne^{-x\hat{A}}$$

3.7 UNITARY OPERATORS AND SIMILARITY TRANSFORMATIONS

Unitary and orthogonal operators map one basis set into another. These operators do not change the length of a vector nor do they change the angle between vectors. While the unitary operators act on abstract Hilbert spaces, the subset of orthogonal operators acts on real Euclidean vectors. The unitary operators preserve the value of the inner product.

3.7.1 ORTHOGONAL ROTATION MATRICES

Orthogonal operators rotate real Euclidean vectors. The word “orthogonal” does not directly concern the inner product between operators but instead refers to the fact that the length of a vector remains unaffected under rotations as well as the angles between vectors. The orthogonal operator can be most conveniently defined through its matrix.

$$\underline{R}^{-1} = \underline{R}^T \quad (3.42)$$

This relation is independent of the basis set chosen for the vector space as it should be since the effect of the “operator” does not depend on the chosen basis set. Recall the definition of the transpose

$$(R^T)_{ab} = R_{ba} \quad \text{or} \quad R_{ab}^T = R_{ba} \quad (3.43)$$

The defining relation in Equation 3.42 can be used to show $\text{Det}(\hat{R}) = 1$

$$1 = \text{Det}(\underline{1}) = \text{Det}(\hat{R}\hat{R}^T) = \text{Det} \hat{R} \text{Det} \hat{R}^T = \text{Det} \hat{R} \text{Det} \hat{R} = (\text{Det} \hat{R})^2$$

and therefore $\text{Det} \hat{R} = 1$ by taking the positive root. The above string of equalities uses the unit operator (unit matrix) defined by $\underline{1} = [\delta_{ab}]$. The discussion shows later that the orthogonal matrix leaves angles and lengths invariant.

Recall that rotations can be viewed as either rotating “vectors” or the “coordinate system.” We take the point of view that operators rotate the vectors as suggested by Figure 3.12. Consider rotating all 2-D vectors by θ (positive when counterclockwise). We find the operator and then the matrix. The rotation operator provides $\hat{R}|1\rangle = |1'\rangle$ and $\hat{R}|2\rangle = |2'\rangle$. Reexpressing $|1'\rangle$ and $|2'\rangle$ in

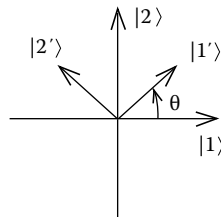


FIGURE 3.12 Rotating the basis vectors and reexpressing them in the original basis set.

terms of the original basis vectors $|1\rangle$ and $|2\rangle$ then provides the matrix elements according to $\hat{R}|1\rangle = R_{11}|1\rangle + R_{21}|2\rangle$ and $\hat{R}|2\rangle = R_{21}|1\rangle + R_{22}|2\rangle$. Figure 3.12 provides

$$\begin{aligned} |1'\rangle &= \hat{R}|1\rangle = \cos\theta|1\rangle + \sin\theta|2\rangle = R_{11}|1\rangle + R_{21}|2\rangle \\ |2'\rangle &= \hat{R}|2\rangle = -\sin\theta|1\rangle + \cos\theta|2\rangle = R_{12}|1\rangle + R_{22}|2\rangle \end{aligned} \quad (3.44)$$

where θ refers to the angle between $|1'\rangle$ and $|1\rangle$. The results can be written as

$$\begin{aligned} \hat{R} &= R_{11}|1\rangle\langle 1| + R_{12}|1\rangle\langle 2| + R_{21}|2\rangle\langle 1| + R_{22}|2\rangle\langle 2| \\ &= \cos\theta|1\rangle\langle 1| - \sin\theta|1\rangle\langle 2| + \sin\theta|2\rangle\langle 1| + \cos\theta|2\rangle\langle 2| \end{aligned} \quad (3.45)$$

Notice that the matrix $\underline{R}$ describes the effect of operating with $\hat{R}$ on the unit vectors. Also notice that the results must be expressed in terms of the original unit vectors, *not* the rotated ones. The operator $\hat{R}$ is most correctly interpreted as associating a new vector $\vec{v}'$ (in the Hilbert space) with the original vector $\vec{v}$. As a note, sometimes people see the word ‘rotation’ and think of an object revolving around an axis. The rotation operators described here do not depend on time. These operators associate a vector in the domain of the operator with another vector making an angle with respect to the first. The angle does not depend on time.

The matrix $\underline{R}$ changes the components of a vector $|v\rangle = x|1\rangle + y|2\rangle$ into $|v'\rangle = x'|1\rangle + y'|2\rangle$ according to

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix} \quad \text{where} \quad \underline{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (3.46)$$

This last relation easily shows $\underline{R}^T \underline{R} = 1$ so that $\underline{R}^{-1} = \underline{R}^T$ as required for an orthogonal operator $\hat{R}$ and matrix $\underline{R}$.

We can now see that the example rotation matrix transforms one basis into another. Equation 3.46 shows that the length of a vector does not change under a rotation by calculating the length

$$\|\vec{v}'\|^2 = (x')^2 + (y')^2 = (x\cos\theta - y\sin\theta)^2 + (x\sin\theta + y\cos\theta)^2 = x^2 + y^2 = \|\vec{v}\|^2$$

Therefore orthogonal matrices do not shrink or expand vectors. The same conclusion can be verified by using Dirac notation

$$\|v'\|^2 = \langle v'|v'\rangle = \langle v|\hat{R}^\dagger \hat{R}|v\rangle = \langle v|\hat{R}^T \hat{R}|v\rangle = \langle v|1|v\rangle = \langle v|v\rangle = \|v\|^2$$

where the fourth term uses the fact that $\hat{R}$ is real. The “rotation” operator $\hat{R}$ does not change the angle between two vectors $|v'\rangle = \hat{R}|v\rangle$ and $|w'\rangle = \hat{R}|w\rangle$. The angle can be defined through the dot product relation $\langle v'|w'\rangle = \vec{v}' \cdot \vec{w}' = v' w' \cos\theta'$.

$$\cos\theta' = \frac{1}{v'w'} \langle v'|w'\rangle = \frac{1}{vw} \langle v|\hat{R}^T \hat{R}|w\rangle = \frac{1}{vw} \langle v|w\rangle = \cos\theta$$

The “rotation” operator $\hat{R}$ is called orthogonal because it does not affect the orthonormality of basis vectors $\{|1\rangle, |2\rangle, \dots\}$ in a real vector space. The set $\{\hat{R}|1\rangle, \hat{R}|2\rangle, \dots\}$ must also be a basis set.

Example 3.30

Write the matrix for the operator that rotates 2-D vectors by 45° counterclockwise. Show that the matrix is orthogonal. The 45° rotation operator provides new unit vectors defined by

$$|1'\rangle = \hat{R}|1\rangle = \frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle \quad \text{and} \quad |2'\rangle = \hat{R}|2\rangle = -\frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle$$

Therefore, the matrix can be written and its transpose must be

$$\underline{R} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad \underline{R}^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Multiplying the two shows $\underline{R}^T \underline{R} = 1$.

Example 3.31

For a 90° vector rotation, the coordinates $x = 1$ and $y = 0$ give the rotation coordinates $x' = 0$ and $y' = 1$ which corresponds to rotating the coordinate axes clockwise (i.e., $\theta < 0$ for the usual definition of an angle).

Example 3.32

Find the new basis vectors under the 2-D rotation. In such a case, we can write

$$\begin{aligned} |1'\rangle &= \cos \theta |1\rangle - \sin \theta |2\rangle \\ |2'\rangle &= \sin \theta |1\rangle + \cos \theta |2\rangle \end{aligned}$$

If needed, we can solve these equations for the unit vectors $|1\rangle$ and $|2\rangle$ and express all the vectors in the Hilbert space in terms of $|1'\rangle$ and $|2'\rangle$

$$\begin{aligned} |1\rangle &= \cos \theta |1'\rangle + \sin \theta |2'\rangle \\ |2\rangle &= -\sin \theta |1'\rangle + \cos \theta |2'\rangle \end{aligned} \tag{3.47}$$

Example 3.33

Find $\vec{r} = 2\tilde{x} + 3\tilde{y}$ in terms of the new basis set using Equation 3.47 with $\theta = 45^\circ$.

$$\vec{r} = 2\left(\frac{1}{\sqrt{2}}|1'\rangle + \frac{1}{\sqrt{2}}|2'\rangle\right) + 3\left(-\frac{1}{\sqrt{2}}|1'\rangle + \frac{1}{\sqrt{2}}|2'\rangle\right) = -\frac{1}{\sqrt{2}}|1'\rangle + \frac{5}{\sqrt{2}}|2'\rangle$$

We have not really rotated $\vec{r}$; we have expressed it in terms of an alternate basis set. If $|1'\rangle$ and $|2'\rangle$ are viewed as rotations of $|1\rangle$ and $|2\rangle$ then we could say that $\vec{r}$ is expressed in the “rotated” basis set.

The “rotation” operator $\hat{R}$ is called orthogonal because it does not affect the orthonormality of basis vectors $\{|1\rangle, |2\rangle, \dots\}$ in a real vector space. The set $\{\hat{R}|1\rangle, \hat{R}|2\rangle, \dots\}$ must also be a basis set. Either basis set works equally well. As will be seen later, we sometimes use a rotated set $\{\hat{R}|a\rangle\}$ because it diagonalizes a matrix.

The set of orthogonal operators is really a subset of the unitary operators.

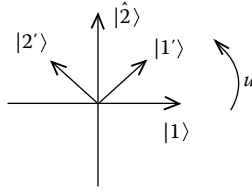


FIGURE 3.13 The unitary operator is determined by the mapping of the basis vectors.

3.7.2 UNITARY TRANSFORMATIONS

A unitary transformation is a “rotation” in the generalized Hilbert space as shown in Figure 3.13. The set of orthogonal operators forms a subset of the unitary operators. A unitary operator “ $\hat{u}$ ” is defined to have the property that

$$\hat{u}^+ = \hat{u}^{-1} \quad \text{or} \quad \hat{u}\hat{u}^+ = 1 = \hat{u}^+\hat{u} \quad (3.48)$$

The unitary operator therefore satisfies $|\det(u)|^2 = 1$ since

$$1 = \text{Det}(\hat{1}) = \text{Det}(\hat{u}\hat{u}^+) = \text{Det}(\hat{u})\text{Det}(\hat{u}^+) = \text{Det}(\hat{u})\text{Det}^*(\hat{u}) = |\text{Det}(\hat{u})|^2$$

which used the property of determinants $\text{Det}(u^T) = \text{Det}(u)$. We can write $\text{Det}(\hat{u}) = e^{i\phi}$. The relation $\hat{u}^+ = \hat{u}^{-1}$ therefore provides the determinant to within a phase factor. We can choose the phase to be zero $\phi = 0$ and thereby require a unitary operator to satisfy $\text{Det}(\hat{u}) = 1$.

The unitary transformations can be thought of as “change of basis operators” similar to the rotation operator $\hat{R}$ in the previous section. That is, if $B_v = \{|a\rangle\}$ forms a basis set then so does $B'_v = \{\hat{u}|a\rangle = |a'\rangle\}$. The operator $\hat{u}$ maps the vector space V into itself $\hat{u}: V \rightarrow V$. Unitary operators preserve the orthonormality relations of the basis set.

$$\langle a'|b'\rangle = (\hat{u}|a\rangle)^+ (\hat{u}|b\rangle) = \langle a|\hat{u}^+\hat{u}|b\rangle = \langle a|1|b\rangle = \langle a|b\rangle = \delta_{ab}$$

As a result, B'_v and B_v are equally good basis sets for the Hilbert space V .

The inverse of the unitary operator $\hat{u}$, $\hat{u}^{-1} = \hat{u}^+$ can be written in matrix notation as

$$\underline{u}^+ = \underline{u}^T{}^* \quad \text{or} \quad (u^+)_{ab} = u_{ba}^* \quad \text{or sometimes} \quad u_{ab}^+ = u_{ba}^*$$

Example 3.34

If $\hat{u} = \sum_{ab} u_{ab}|a\rangle\langle b|$ then $\hat{u}^+$ can be calculated as

$$\hat{u}^+ = \sum_{ab} (u_{ab}|a\rangle\langle b|)^+ = \sum_{ab} (u_{ab})^+ |b\rangle\langle a| = \sum_{ab} u_{ab}^* |b\rangle\langle a|$$

Now notice that u_{ab} represents a single complex number and not the entire matrix so that the dagger can be replaced by the complex conjugate without interchanging the indices.

Example 3.35

Show for the previous example that $\hat{u}^+\hat{u} = 1$

$$\hat{u}^+\hat{u} = \left(\sum_{\alpha\beta} u_{\alpha\beta}^* |\beta\rangle\langle\alpha| \right) \left(\sum_{ab} u_{ab} |a\rangle\langle b| \right) = \sum_{\substack{ab \\ \alpha\beta}} u_{\alpha\beta}^* u_{ab} |\beta\rangle\langle b| \delta_{a\alpha} = \sum_{\substack{ab \\ \beta}} u_{a\beta}^* u_{ab} |\beta\rangle\langle b|$$

We need to work with the product of the unitary matrices.

$$\sum_a u_{a\beta}^* u_{ab} = \sum_a (u^\dagger)_{\beta a} u_{ab} = (u^\dagger u)_{\beta b} = \delta_{\beta b}$$

Notice that we switched the indices when we calculated Hermitian adjoint of the matrix since we are referring to the entire matrix. Substituting this result for the unitary matrices gives us

$$\hat{u}^\dagger \hat{u} = \sum_{\beta b} \delta_{\beta b} |\beta\rangle \langle b| = \sum_b |b\rangle \langle b| = 1$$

3.7.3 VISUALIZING UNITARY TRANSFORMATIONS

Unitary transformations change one basis set into another basis set.

$$B_v = \{|a\rangle\} \rightarrow B'_v = \{\hat{u}|a\rangle = |a'\rangle\}$$

Figure 3.13 shows the effect of the unitary transformation

$$\hat{u}|1\rangle = |1'\rangle \quad \hat{u}|2\rangle = |2'\rangle$$

The operator is defined by its effect on the basis vectors. The two objects $|1'\rangle\langle 1|$ and $|2'\rangle\langle 2|$, which are “basis vectors” for the vector space of operators $\{\hat{u}\}$, perform the following mappings

$$\begin{aligned} |1'\rangle\langle 1| & \text{ maps } |1\rangle \rightarrow |1'\rangle \quad \text{since} \quad [|1'\rangle\langle 1|]|1\rangle = |1'\rangle\langle 1|1\rangle = |1'\rangle \\ |2'\rangle\langle 2| & \text{ maps } |2\rangle \rightarrow |2'\rangle \quad \text{since} \quad [|2'\rangle\langle 2|]|2\rangle = |2'\rangle\langle 2|2\rangle = |2'\rangle \end{aligned}$$

Putting both pieces together gives us a very convenient form for the operator

$$\hat{u} = |1'\rangle\langle 1| + |2'\rangle\langle 2|$$

The operator can be written just by placing vectors next to each other! The operator $\hat{u}$ can be left in the form

$$\hat{u} = \sum_n |n'\rangle\langle n|$$

to handle “rotations” in all directions. Notice that the summation involves only n . This means to sum the following two terms: $|1'\rangle\langle 1|$ and $|2'\rangle\langle 2|$. Of course, to use $\hat{u}$ for actual calculations, either $|n'\rangle$ must be expressed as a sum over $|n\rangle$ or vice versa.

Example 3.36

Consider a 2-D space with basis set $\{|1\rangle, |2\rangle\}$ and a rotation through θ in the counterclockwise direction. Find the rotation operator.

SOLUTION

The solution is

$$\hat{u} = \sum_n |n'\rangle\langle n|$$

where $|n'\rangle$ is the image of $|n\rangle$. To use $\hat{u}$ for actual calculations, $|n'\rangle$ usually should be expressed as a sum over $|n\rangle$. For the 2-D real case, the basis vectors map according to

$$|1'\rangle = \cos \theta |1\rangle + \sin \theta |2\rangle \quad |2'\rangle = -\sin \theta |1\rangle + \cos \theta |2\rangle$$

as shown in the previous section. So that the unitary operator $\hat{u}$ becomes

$$\hat{u} = |1'\rangle\langle 1| + |2'\rangle\langle 2| = \cos \theta |1\rangle\langle 1| - \sin \theta |1\rangle\langle 2| + \sin \theta |2\rangle\langle 1| + \cos \theta |2\rangle\langle 2|$$

Leaving the unitary operator u in terms of $|1'\rangle \langle 1|$ gives a convenient, clear picture of the operator that changes $|n\rangle$ into $|n'\rangle$.

3.7.4 TRACE AND DETERMINANT

The trace is important for calculating averages. Similarity transformations leave the trace and determinant unchanged. That is, trace and determinant operations are invariant with respect to similarity transformations. Consider

$$\hat{A}' = \hat{u}\hat{A}\hat{u}^+ \quad \text{and} \quad \hat{u}: V \rightarrow V$$

The cyclic property of the trace and the fact that $\hat{u}$ is a unitary operator provides

$$\text{Tr}(\hat{A}') = \text{Tr}(\hat{u}\hat{A}\hat{u}^+) = \text{Tr}(\hat{A}\hat{u}^+\hat{u}) = \text{Tr}(\hat{A})$$

since the unitary operator satisfies $\hat{u}^+\hat{u} = 1$. The same calculation can be performed for the determinant

$$\text{Det}(\hat{A}') = \text{Det}(\hat{u}\hat{A}\hat{u}^+) = \text{Det}(\hat{u})\text{Det}(\hat{A})\text{Det}(\hat{u}^+) = \text{Det}(\hat{A})\text{Det}(\hat{u}\hat{u}^+) = \text{Det}(\hat{A})$$

3.7.5 SIMILARITY TRANSFORMATIONS

Assume there exists a linear operator $\hat{O}$ that maps the vector space into itself $\hat{O}: V \rightarrow V$. Assume that the vectors $|v\rangle$ and $|w\rangle$ (not necessarily basis vectors) satisfy an equation of the form:

$$\hat{O}|v\rangle = |w\rangle \tag{3.49}$$

Now suppose that we transform both sides by the unitary transformation $\hat{u}$ and then use the definition of unitary $\hat{u}^+\hat{u} = 1$ to find

$$\hat{u}\hat{O}|v\rangle = \hat{u}|w\rangle \rightarrow \hat{u}\hat{O}\hat{u}^+\hat{u}|v\rangle = \hat{u}|w\rangle$$

Defining $\hat{O}' = \hat{u}\hat{O}\hat{u}^+$ and $|v'\rangle = \hat{u}|v\rangle$, and $|w'\rangle = \hat{u}|w\rangle$ provides

$$\hat{O}'|v'\rangle = |w'\rangle \tag{3.50}$$

which has the same form as the original equation. The difference is that the operator $\hat{O}$ is now expressed in the “rotated basis set” as

$$\hat{O}' = \hat{u}\hat{O}\hat{u}^+ \tag{3.51a}$$

Changing basis vectors also changes the representation of the operator $\hat{O}$ as can easily be seen from the basis expansion of the operator. Basically Equation 3.50 says that the relation that originally held in the original basis has now been transferred to the new basis. Example 3.37 below demonstrates a case for an operator that stretches vectors along the y -direction, which then stretches along the new y -axis after the rotation—the effects of the operator rotate with the system in order that Equation 3.49 should hold in either the original or rotated system.

Transformations as those found in Equation 3.51a are “similarity” transformations. More generally, we write the similarity transformation as

$$\hat{O}' = \hat{S}\hat{O}\hat{S}^{-1} \quad (3.51b)$$

for the general linear transformation $\hat{S}$. Equation 3.51b is equivalent to Equation 3.51a because $\hat{u}$ is unitary $\hat{u}^{-1} = \hat{u}^+$.

The similarity transformation can also be seen to have the form $\hat{O}' = \hat{u}\hat{O}\hat{u}^+$ by using the transformation $\hat{u}$ directly on the vectors in the basis vector expansion. For convenience, assume $\hat{O}: V \rightarrow V$ with $V = \text{Sp } \{|a\rangle\}$. Replacing $|a\rangle$ with $\hat{u}|a\rangle$ and $|b\rangle$ with $\hat{u}|b\rangle$ produces

$$\hat{O} = \sum_{ab} O_{ab} |a\rangle\langle b| \rightarrow \hat{O}' = \sum_{ab} O_{ab} (\hat{u}|a\rangle)(\hat{u}|b\rangle)^+ = \sum_{ab} O_{ab} \hat{u}|a\rangle\langle b|\hat{u}^+ = \hat{u}\hat{O}\hat{u}^+$$

which is the same result as before.

A string of operators can be rewritten using unitary transformation $\hat{u}$

$$(\hat{O}_1\hat{O}_2 + 5\hat{O}_3\hat{O}_4^3)|v\rangle = |w\rangle \rightarrow (\hat{O}'_1\hat{O}'_2 + 5\hat{O}'_3\hat{O}'_4^3)|v'\rangle = |w'\rangle$$

For example, $\hat{O}_4^3$ can be transformed by repeatedly inserting a “1” and applying $1 = \hat{u}^+\hat{u}$ as follows:

$$\hat{u}(\hat{O}_4^3)\hat{u}^+ = \hat{u}(\hat{O}_4\hat{O}_4\hat{O}_4)\hat{u}^+ = \hat{u}(\hat{O}_41\hat{O}_41\hat{O}_4)\hat{u}^+ = \hat{u}\hat{O}_4\hat{u}^+\hat{u}\hat{O}_4\hat{u}^+\hat{u}\hat{O}_4\hat{u}^+ = \hat{O}'_4\hat{O}'_4\hat{O}'_4 = (\hat{O}'_4)^3$$

Example 3.37

Consider the operator $\hat{O} = |1\rangle\langle 1| + 2|2\rangle\langle 2|$ that stretches vectors along the y -direction (axis “2”). Rotate the basis by 90° and discuss the effects on the stretching.

SOLUTION

The operator that rotates by 90° is seen to be given by

$$\hat{u} = |1'\rangle\langle 1| + |2'\rangle\langle 2| = |2\rangle\langle 1| - |1\rangle\langle 2|$$

where the primes indicate the new basis. Then using the matrix isomorphism property (for convenience and practice) we find

$$\hat{u}\hat{O}\hat{u}^+ \equiv \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \equiv 2|1\rangle\langle 1| + |2\rangle\langle 2| = 2|2'\rangle\langle 2'| + |1'\rangle\langle 1'|$$

since the new y -axis points along the old negative x -axis and the new x -axis points along the old y -axis. This relation makes it clear that the rotated operator still stretches along the y -axis but that y -axis is rotated in relation to the old. Figure 3.14 shows how the operator $\hat{O}$ changes vectors which terminate on a unit circle into ones that terminate on an ellipse-like curve. The figure then

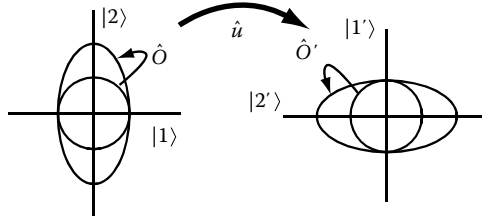


FIGURE 3.14 The operator $\hat{O}$ maps the circle into the ellipse-like curve and $\hat{u}$ rotates.

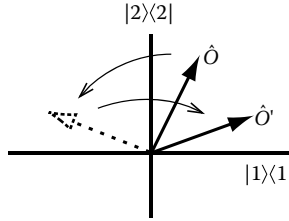


FIGURE 3.15 The effects of the rotation on the operator $\hat{O}$.

shows the results of the similarity transformation. Interestingly, if one represents the operator $\hat{O}$ in its vector space as shown in Figure 3.15 then the rotation moves it by 90° (temporarily including the minus signs) but canceling the negatives changes the operator to the first quadrant. Notice that the operator $\hat{O}$ initially has the larger component along the vertical axis and then after the rotation has the larger component along the horizontal axis (in the original coordinate system).

Example 3.38

Write $\langle v' | \hat{T}' | w' \rangle$ in terms of the objects $|v\rangle$, $\hat{T}$, $|w\rangle$ where $|v'\rangle = \hat{u}|v\rangle$ and $\hat{T}' = \hat{u}\hat{T}\hat{u}^+$ and $|w'\rangle = \hat{u}|w\rangle$. This is done as follows:

$$\langle v' | \hat{T}' | w' \rangle = \langle v | \hat{u}^+ (\hat{u} \hat{T} \hat{u}^+) \hat{u} | w \rangle = \langle v | \hat{T} | w \rangle$$

again $\hat{O}' = \hat{u} \hat{O} \hat{u}^+$ is the representation of the operator $\hat{O}$ using the new basis set $B'_v = \{\hat{u}|a\rangle\}$.

3.7.6 EQUIVALENT AND REDUCIBLE REPRESENTATIONS OF GROUPS

One matrix representation of a group is equivalent to another when the two sets of matrices are related by a similarity transformation. Suppose the two sets of matrices corresponding to each element g of the group G are given by $\{M(g)\}$, $\{M'(g)\}$. One might think of the set of g to be rotations of 120° in the xy -plane or the operations of flipping vectors across the line $x=y$, for example. M and M' might be distinguished in that they originate in different basis sets. If there exists a single transformation S , independent of the particular group element g , such that

$$M'(g) = \underline{S} M(g) \underline{S}^{-1}$$

then the two representations are equivalent. For rotations on a Hilbert space, S would be the unitary transformation. It should be clear, for example, that if the two sets of matrices $\{M(g)\}$, $\{M'(g)\}$ differ only through their basis sets, then they are equivalent.

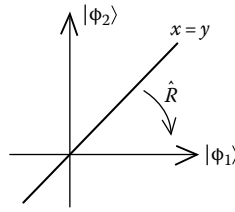


FIGURE 3.16 Rotate the basis through -45° .

Example 3.39

Consider a group of transformations that flip vectors across the line $x = y$. One matrix representation is given by

$$\underline{O} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \underline{O}^{-1} = \underline{O}^+ = \underline{O} \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

If we change basis sets by rotating through -45° as shown in Figure 3.16, the rotation matrix is

$$\underline{R} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

and an equivalent representation can be found by transforming each matrix in the representation using the same $\underline{R}$

$$\underline{O}' = \underline{R}\underline{O}\underline{R}^{-1} = \underline{R}\underline{O}\underline{R}^+ = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (\underline{O}')^{-1} = \underline{O}' \quad I' = I$$

One can see that the original transformation O changed, for example, a vector along the x -axis $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ into one along the y -axis $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and vice versa. In the new representation, vectors parallel to the new x -axis remain unchanged whereas those along the new y -axis map into their negatives $\begin{pmatrix} 0 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ -1 \end{pmatrix}$. The new representation continues to flip across the same line except the description of that line has changed (it is now parallel to the new x -axis) and therefore so has the matrix representing the flipping process. However, the representations are equivalent in that they represent the same flipping process.

3.8 HERMITIAN OPERATORS AND THE EIGENVECTOR EQUATION

The adjoint, self-adjoint, and Hermitian operators play a central role in the study of quantum mechanics and the Sturm–Liouville problem for solving partial differential equations. In quantum mechanics, Hermitian operators represent physically observable quantities (i.e., dynamical variables) such as momentum $\hat{p}$, energy $\hat{\mathcal{H}}$, and electric field. As we shall see later, the eigenvectors of a Hermitian operator form a basis set for the vector space and represent the most fundamental states of the particle. If the particle “occupies” one of these basis states then the result of applying the Hermitian operator to the state produces an eigenvalue, which represents the result of observing (i.e., measuring) the corresponding dynamical variable. The collection of all allowed eigenvalues provides the results for every possible measurement. Besides inducing a basis set, the Hermitian operators have real eigenvalues which makes physical sense since measurements in the laboratory produce real values. Clearly, the Hermitian operator has immense importance to the interpretation of the physical world.

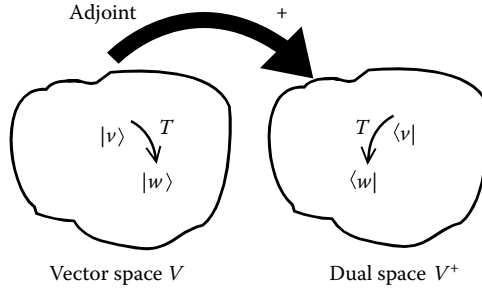


FIGURE 3.17 The vector and dual space.

3.8.1 ADJOINT, SELF-ADJOINT, AND HERMITIAN OPERATORS

Let $\hat{T}: V \rightarrow V$ be a linear transformation defined on a Hilbert space V with basis vectors given by $\{|n\rangle: n = 1, 2, \dots\}$. Let $|f\rangle, |g\rangle$ be two elements in the Hilbert space. We define the adjoint operator $\hat{T}^+$ to be the operator which satisfies

$$\langle g | \hat{T}f \rangle = \langle \hat{T}^+g | f \rangle \quad (3.52)$$

for all functions $|f\rangle$ and $|g\rangle$ in the Hilbert space. Note the use of the alternate notation: $\hat{T}|f\rangle = |\hat{T}f\rangle$.

Previous sections have introduced the notion of the adjoint $\hat{T}^+$ as “somehow” connected with the dual vectors space (Figure 3.17). The definition above suggests a method to calculate an explicit form for $\hat{T}^+$ (as seen later). For now, let us show how the version of the adjoint operator in Chapter 2 relates to the new definition given in Equation 3.52. First consider the term $\langle g | \hat{T}$. Using the adjoint (and the alternate notation), one can write $[\langle g | \hat{T}]^+ = \hat{T}^+|g\rangle = |\hat{T}^+g\rangle$ or, taking the adjoint of both sides, one finds $\langle g | \hat{T} = \langle \hat{T}^+g |$. So therefore, combining these two results, namely $\langle g | \hat{T} = \langle \hat{T}^+g |$ and $\hat{T}|f\rangle = |\hat{T}f\rangle$ we obtain the desired results,

$$\langle \hat{T}^+g | f \rangle = [\langle g | \hat{T}] | f \rangle = \langle g | [\hat{T} | f \rangle] = \langle g | \hat{T}f \rangle$$

So using the previous definition of adjoint, specifically $(\hat{A}\hat{B})^+ = \hat{B}^+\hat{A}^+$, leads to the new definition of adjoint in Equation 3.52.

Conversely, we can show that the new definition for adjoint (Equation 3.52) leads to the relation $(\hat{A}\hat{B})^+ = \hat{B}^+\hat{A}^+$. Consider the relation

$$\langle f | \hat{A}\hat{B}g \rangle = \langle \hat{A}^+f | \hat{B}g \rangle = \langle \hat{B}^+\hat{A}^+f | g \rangle$$

Then by the new definition of adjoint in Equation 3.52, we conclude that the adjoint of $\hat{A}\hat{B}$ but be $\hat{B}^+\hat{A}^+$ as required.

Definition: An operator $\hat{T}$ is self-adjoint or Hermitian if $\hat{T}^+ = \hat{T}$.

Example 3.40

If $\hat{T} = \frac{\partial}{\partial x}$ then find $\hat{T}^+$ for the Hilbert space of differentiable functions that approach zero as $x \rightarrow \pm\infty$. The Hilbert space is

$$\text{HS} = \left\{ f: \frac{\partial f(x)}{\partial x} \text{ exists and } f \rightarrow 0 \text{ as } x \rightarrow \pm\infty \right\}$$

SOLUTION

We want $\hat{T}^+$ such that

$$\langle f | \hat{T}g \rangle = \langle \hat{T}^+ f | g \rangle$$

Start with the quantity on the left

$$\langle f | \hat{T}g \rangle = \int_{-\infty}^{\infty} dx f^*(x) \hat{T}g(x) = \int_{-\infty}^{\infty} dx f^*(x) \frac{\partial}{\partial x} g(x)$$

The procedure usually starts with integration by parts:

$$\langle f | \hat{T}g \rangle = f^*(x)g(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} dx \frac{\partial f^*(x)}{\partial x} g(x)$$

In most cases, the boundary term always gives zero. Notice (to some extent) the Hermitian property of the operators depends on the properties of the Hilbert space. In the present case, the Hilbert space is defined such that $f^*(\infty)g(\infty) - f^*(-\infty)g(-\infty) = 0$; most physically sensible functions drop to zero for very large distances. Next move the minus sign and partial derivative under the complex conjugate to find

$$\langle f | \hat{T}g \rangle = \int_{-\infty}^{\infty} dx \left[-\frac{\partial f(x)}{\partial x} \right]^* g(x) = \langle \hat{T}^+ f | g \rangle$$

Note everything inside the bra $\langle$ must be placed under the complex conjugate $()^*$ in the integral. The operator $\hat{T}^+$ must therefore be $\hat{T}^+ = -\frac{\partial}{\partial x}$ or equivalently

$$\left(\frac{\partial}{\partial x} \right)^+ = -\frac{\partial}{\partial x} \quad (3.53)$$

Example 3.41

Find the adjoint operator

$$\hat{T}^+ = \left(i \frac{\partial}{\partial x} \right)^+$$

for the same set of functions as for Example 3.40 where $i = \sqrt{-1}$.

Method 1: The quick method

$$\left(i \frac{\partial}{\partial x} \right)^+ = (i)^+ \left(\frac{\partial}{\partial x} \right)^+ = (-i) \left(-\frac{\partial}{\partial x} \right) = i \frac{\partial}{\partial x}$$

where the second term comes from Example 3.40.

Method 2:

$$\langle f | \hat{T}g \rangle = \int_{-\infty}^{\infty} dx f^*(x) \left(i \frac{\partial}{\partial x} \right) g(x) = i f^*(x)g(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} dx i \left(\frac{\partial}{\partial x} f^* \right) g(x)$$

Again $f^*(\infty)g(\infty) - f^*(-\infty)g(-\infty) = 0$ and so

$$\langle f | \hat{T} g \rangle = \int_{-\infty}^{\infty} dx \left[i \frac{\partial f(x)}{\partial x} \right]^* g(x) = \langle \hat{T}^+ f | g \rangle$$

Therefore, the adjoint can be identified as

$$\hat{T}^+ = \left(i \frac{\partial}{\partial x} \right)^+ = i \frac{\partial}{\partial x} = \hat{T} \quad (3.54)$$

As a result, both methods show that $\hat{T} = i \frac{\partial}{\partial x}$ is self-adjoint (i.e., Hermitian). For example, the quantum mechanical “momentum operator” which is defined by

$$\hat{p} = \frac{\hbar}{i} \frac{\partial}{\partial x}$$

must be Hermitian; it corresponds to a physical observable.

As an important note, the boundary term $f^*(x)g(x)|_a^b$ (from the partial integration in the inner product) is always arranged to be zero. The method of making it zero depends on the definition of the Hilbert space. A number of different Hilbert spaces can produce a zero surface term. For example, if the function space is defined for $x \in [a, b]$, then the following conditions will work

- (1) $f(a) = f(b) = 0$ $\forall$ for every function f in V . $f \in V$.
- (2) $f(a) = f(b)$ (without being equal to zero) for every f in the space V .

Notice that the property of an operator being Hermitian cannot be entirely separated from the properties of the Hilbert space since the surface terms must be zero.

3.8.2 ADJOINT AND SELF-ADJOINT MATRICES

First, we derive the form of the adjoint using the basis expansion of an operator. In the following, let $|m\rangle$ and $|n\rangle$ (also for i, j) be basis vectors. Take the adjoint of the basis expansion

$$\hat{T} = \sum_{mn} T_{mn} |m\rangle \langle n| \quad \text{to get} \quad \hat{T}^+ = \sum_{mn} T_{mn}^* |n\rangle \langle m|$$

where T_{mn} becomes the complex conjugate since it is only a number. So now

$$\langle i | \hat{T}^+ | j \rangle = \sum_{mn} T_{mn}^* \langle i | n \rangle \langle m | j \rangle = \sum_{mn} T_{mn}^* \delta_{in} \delta_{mj} = T_{ji}^*$$

This last equation shows that the adjoint matrix involves a complex conjugate and has the indices reversed from the matrix $\underline{T}$.

$$(T^+)_{ij} = T_{ji}^* \quad (3.55)$$

Now, we show how the adjoint comes from the basic definition of the adjoint operator. The basic definition of the adjoint can be written as

$$\langle w | \hat{T} v \rangle = \langle \hat{T}^+ w | v \rangle \quad (3.56)$$

To work with $\langle w|\hat{T}v\rangle$ in this definition, we need to use matrix notation for the inner product between two vectors $|w\rangle$ and $|v\rangle$

$$\langle w|v\rangle = \sum_m w_m^* v_m = \underline{w}^+ \underline{v} \quad (3.57)$$

where $\underline{v}$ and $\underline{w}$ are column matrices. The left-hand side of Equation 3.56 can be transformed into the right-hand side as follows:

$$\begin{aligned} \langle w|\hat{T}v\rangle &= \sum_{mn} \{w_m^* \langle m|\} \hat{T} \{v_n|n\rangle\} = \sum_{mn} w_m^* T_{mn} v_n = \sum_{mn} (\underline{T}^T)_{nm} w_m^* v_n \\ &= \sum_{mn} \left[\left(\underline{T}^{*T} \right)_{nm} w_m \right]^* v_n = \left[\underline{T}^{*T} \underline{w} \right]^+ \underline{v} = \langle \hat{T}^+ w|v\rangle \end{aligned}$$

where the “+” in the second to last step comes from requiring that the column vector $\underline{v}^* = (\underline{T}^{*T} \underline{w})^*$ becomes a row vector to multiply into the column vector $\underline{v}$. The adjoint must therefore be $T^+ = T^{*T}$.

Finally, a specific form for a Hermitian matrix can be determined. A matrix is Hermitian provided $T = T^+$. For example, a 2×2 matrix is Hermitian if

$$\underline{T} = \underline{T}^+ \quad \text{so that} \quad \underline{T} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} = T^+$$

For $\underline{T}$ to be Hermitian, require $a = a^*$, $d = d^*$, so that a, b are both real and $b = c^*$. The self-adjoint form of the matrix $\underline{T}$ is then

$$\underline{T} = \begin{bmatrix} a & b \\ b^* & d \end{bmatrix}$$

where both a, d are real.

Example 3.42

For the inner product $\langle w|\hat{T}|v\rangle$, in matrix form, show how the adjoint becomes the transpose and complex conjugate.

SOLUTION

If $\hat{T}$ represents an operator, and $|v\rangle$ and $|w\rangle$ represent two vectors in the Hilbert space with basis set $\{|n\rangle\}$ then

$$\langle w|\hat{T}|v\rangle = \langle w|1\hat{T}1|v\rangle = \sum_{mn} \langle w|m\rangle \langle m|\hat{T}|n\rangle \langle n|v\rangle = \sum_{mn} \langle m|w\rangle^+ \langle m|\hat{T}|n\rangle \langle n|v\rangle \quad (3.58)$$

Equation 3.58 shows how the adjoint comes into play. The components of the vectors $|v\rangle$ and $|w\rangle$ are the collection of complex numbers $\langle m|w\rangle$ and $\langle n|v\rangle$ which can be arranged as the “column vectors.” These “column vectors” really are not vectors at all but instead, a collection of “vector components.”

$$\underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix} \quad \text{and} \quad \underline{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \end{bmatrix}$$

Equation 3.58 shows that the product $\langle w|T|v\rangle$ can be written as

$$\langle w|T|v\rangle = \underline{w}^+ \underline{T} \underline{v} \quad (3.59)$$

where

$$\underline{T} = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1n} \\ T_{21} & T_{22} & \cdots & T_{2n} \\ \vdots & \vdots & & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{bmatrix}$$

Equation 3.59 shows how the inner product can be written as a matrix equation using the adjoint. The adjoint gives

$$\underline{w}^+ = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}^+ = [w_1^* \quad w_2^* \quad \cdots]$$

3.9 RELATION BETWEEN UNITARY AND HERMITIAN OPERATORS

An important exponential relation connects certain unitary operators with other certain Hermitian ones. Those of particular note include rotations in Hilbert space. Interestingly, translations in ordinary 3-D space also appear as rotations in Hilbert space. The unitary operators describe the rotations while the Hermitian ones “generate” those rotations. As will become evident in subsequent chapters, the exponential relation combines conjugate variables such as position–linear momentum, angle–angular momentum, and time–energy. The exponential relation further connects the physical everyday 3-D space with the Hilbert space. A transformation of a quantum mechanical system is associated with the unitary operator in Hilbert space.

3.9.1 RELATION BETWEEN HERMITIAN AND UNITARY OPERATORS

As previously discussed, a Hermitian operator $\hat{H}: V \rightarrow V$ has the property that $\hat{H} = \hat{H}^+$. Unitary operators can be expressed in the form $\hat{u} = e^{i\hat{H}}$ where $\hat{H}$ is a Hermitian operator.

We can show that the operator $\hat{u} = e^{i\hat{H}}$ is unitary by showing $\hat{u}^+ \hat{u} = 1$

$$\hat{u}^+ \hat{u} = (e^{i\hat{H}})^+ (e^{i\hat{H}}) = e^{-i\hat{H}^+} e^{i\hat{H}} = e^{-i\hat{H}} e^{i\hat{H}} = e^0 = 1$$

This is a one-line proof, but a few steps need to be explained in the following steps. One should note that the relation can be extended as $\hat{u} = e^{it\hat{H}}$ when t is a real parameter.

1. A function of an operator $f(\hat{A})$ must be interpreted as a Taylor expansion. We define the “exponential of an operator” to be shorthand notation for a Taylor series expansion in that operator. Recall that the Taylor series expansion of an exponential has the form:

$$e^{ax} = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\partial^n e^{ax}}{\partial x^n} \bigg|_{x=0} x^n = 1 + \frac{\partial}{\partial x} (e^{ax})_{x=0} x + \cdots = 1 + ax + \frac{a^2}{2} x^2 + \cdots$$

In analogy, the exponential of an operator $\hat{H}$ (or equivalently of a matrix $\underline{H}$) can be written as

$$e^{i\underline{H}t} = \underline{1} + (i\underline{H})t + \frac{(i\underline{H})^2}{2}t^2 + \dots \quad \text{so that} \quad e^{i\underline{H}} = \underline{1} + (i\underline{H}) - \frac{\underline{H}^2}{2} + \dots$$

The exponential can now be computed by multiplying matrices on the right-hand side.

2. We wrote

$$e^{-i\hat{H}}e^{i\hat{H}} = e^{i(\hat{H}-\hat{H})} = e^0 = 1$$

As shown in Section 3.6,

$$e^{\hat{A}}e^{\hat{B}} = e^{\hat{A}+\hat{B}}$$

when the commutator of the two operators produces 0, that is,

$$[\hat{A}, \hat{B}] = 0$$

This condition is satisfied because

$$[\hat{H}, \hat{H}] = \hat{H}\hat{H} - \hat{H}\hat{H} = 0$$

Example 3.43

Find the unitary matrix corresponding to $e^{i\underline{H}}$ where

$$\underline{H} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}$$

SOLUTION

First note that the matrix H is Hermitian, i.e., $\underline{H} = \underline{H}^+$

$$\underline{u} = e^{i\underline{H}} = \exp\left\{i\begin{pmatrix} 0.1 & 0 \\ 0 & 0.2 \end{pmatrix}\right\} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + i\begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix} + \frac{i^2}{2!}\begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}^2 + \dots = \begin{bmatrix} e^{i0.1} & 0 \\ 0 & e^{i0.2} \end{bmatrix}$$

Example 3.44

For $\underline{u}$ in Example 3.43, using the unit column vectors

$$\underline{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \underline{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and find the transformed vectors

$$\underline{e}'_1 = \underline{u}\underline{e}_1 \quad \underline{e}'_2 = \underline{u}\underline{e}_2$$

and show that they are orthogonal to each other.

$$\begin{aligned}\underline{e}'_1 &= \begin{pmatrix} e^{i0.1} & 0 \\ 0 & e^{i0.2} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{i0.1} \\ 0 \end{pmatrix} \\ \underline{e}'_2 &= \begin{pmatrix} e^{i0.1} & 0 \\ 0 & e^{i0.2} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ e^{i0.2} \end{pmatrix}\end{aligned}$$

Then

$$\langle e'_1 | e'_2 \rangle \equiv \underline{e}'_1{}^\dagger \underline{e}'_2 = \begin{pmatrix} e^{-i0.1} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ e^{i0.2} \end{pmatrix} = 0$$

3.10 EIGENVECTORS AND EIGENVALUES FOR HERMITIAN OPERATORS

We now show some important theorems. The first theorem shows that Hermitian operators produce real eigenvalues. The importance of this theorem issues from representing all physically observable quantities by Hermitian operators. The result of making a measurement of the observable must produce a real number. For example, for a particle in an eigenstate $|n\rangle$ of the Hermitian energy operator $\hat{\mathcal{H}}$ (i.e., the Hamiltonian), the result for measuring the energy $\hat{\mathcal{H}}|n\rangle = E_n|n\rangle$ produces the real energy E_n . The particle has energy E_n when it occupies state $|n\rangle$. Energy can never be complex (except possibly for some mathematical constructs).

The second theorem shows that the eigenvectors of a Hermitian operator form a basis (we do not prove completeness). This basically says that for every observable in nature, there must always be a Hilbert space large enough to describe all possible results of measuring that observable. The state of the particle or system can be decomposed into the basis vectors.

For boundary value problems, these two theorems say that the Sturm–Liouville equation that has a Hermitian operator always produces a basis set with real eigenvalues. This basis set can be used to expand solutions in an orthonormal expansion as discussed in (books on boundary value problems and partial differential equations).

3.10.1 BASIC THEOREMS FOR HERMITIAN OPERATORS

Before discussing theorems, a few words should be mentioned about notation conventions and about degenerate eigenvalues. We will assume that for each eigenvalue E_n there exists a single corresponding eigenfunction $|\phi_n\rangle$. We customarily label the eigenfunction by either the eigenvalue or by the eigenvalue number as

$$|\phi_n\rangle = |E_n\rangle = |n\rangle$$

Usually, the eigenvalues are listed in the order of increasing value

$$E_1 < E_2 < \cdots$$

The condition of nondegenerate eigenvalues means that for a given eigenvalue, there exists only one eigenvector. The eigenvalues are “degenerate” if for a given eigenvalue, there are multiple eigenvectors.

Nondegenerate	Degenerate
$E_1 \leftrightarrow E_1\rangle$	$E_1 \leftrightarrow E_1\rangle$
$\vdots$	$E_2 \leftrightarrow E_2 1\rangle, E_2 2\rangle$
$E_n \leftrightarrow E_n\rangle$	$E_3 \leftrightarrow E_3\rangle$

The degenerate eigenvectors (which means both states have the same “energy” E_n) actually span a subspace of the full vector space. For example, in the above table, the vectors $|E_2 1\rangle$, $|E_2 2\rangle$ corresponding to the eigenvalue E_2 form a 2-D subspace. Mathematically, we can associate E_2 with any vector in the subspace spanned by $\{|E_2, 1\rangle, |E_2, 2\rangle\}$; however, it is better to choose one vector in the subspace that has significance for a second Hermitian operator (see Theorem 3.9 below and Chapter 5 for more detail). After making the choice, we end up with a nondegenerate case: $|E_1\rangle, |E_2\rangle, |E_3\rangle, \dots$

Physically, the degeneracy can be removed by manipulating the extra degree of freedom represented by the “1” and “2” in $|E_2 1\rangle$, $|E_2 2\rangle$. Sometimes, applying a magnetic field or an electric field will eliminate the degeneracy. As will be seen later, mathematically, we recognize that there exists another Hermitian operator, say $\hat{O}$, that commutes $\hat{H}\hat{O} - \hat{O}\hat{H} = 0$ with the operator $\hat{H}$ so that the eigenvalues of the operator $\hat{O}$ are related to the “1” and “2.”

Now to show that a Hermitian operator $\hat{H}$ has “real” eigenvalues and orthogonal eigenvectors.

THEOREM 3.6: Hermitian Operators Have Real Eigenvalues

If $\hat{H}$ is Hermitian then its eigenvalues are real.

Proof: Assume that the set $\{|n\rangle\}$ contains the eigenvectors corresponding to the eigenvalues $\{E_n\}$ so that the eigenvector equation can be written as $\hat{H}|n\rangle = E_n|n\rangle$. Consider

$$\langle n|\hat{H}|n\rangle = \langle n|E_n|n\rangle = E_n\langle n|n\rangle = E_n$$

where the eigenvectors are assumed to be normalized to unity as $\langle n|n\rangle = 1$. So

$$\langle n|\hat{H}|n\rangle = E_n \tag{3.60}$$

take the adjoint of both sides

$$\langle n|\hat{H}|n\rangle^+ = (E_n)^+$$

Reversing the factors on the left-hand side and changing the “dagger” into a complex conjugate on the right-hand side provides

$$\langle n|\hat{H}^+|n\rangle = E_n^*$$

Using the Hermitian property of the operator $\hat{H} = \hat{H}^+$ we find

$$\langle n|\hat{H}|n\rangle = E_n^* \tag{3.61}$$

Comparing Equations 3.60 and 3.61, we see

$$E_n = E_n^*$$

which means that E_n must be real.

THEOREM 3.7: Hermitian Operators Have Orthogonal Eigenvectors

If $\hat{\mathcal{H}}$ is Hermitian then the eigenvectors corresponding to different eigenvalues are orthogonal.

Proof: Assume $E_m \neq E_n$ and start with two separate eigenvalue equations

$$\begin{array}{ll}
 \hat{\mathcal{H}}|E_m\rangle = E_m|E_m\rangle & \hat{\mathcal{H}}|E_n\rangle = E_n|E_n\rangle \\
 \text{operate with } \langle E_n| & \text{operate with } \langle E_m| \\
 \langle E_n|\hat{\mathcal{H}}|E_m\rangle = E_m\langle E_n|E_m\rangle & \langle E_m|\hat{\mathcal{H}}|E_n\rangle = E_n\langle E_m|E_n\rangle \\
 & \text{Take adjoint of both sides} \\
 & \langle E_n|\hat{\mathcal{H}}|E_m\rangle = E_n\langle E_n|E_m\rangle
 \end{array}$$

where the right-hand column made use of the Hermiticity of the operator $\hat{\mathcal{H}}$ and the reality of the eigenvalues E_n . Now subtract the results of the two columns to find

$$0 = (E_m - E_n)\langle E_n|E_m\rangle$$

We assumed that $E_m - E_n \neq 0$ and therefore $\langle E_n|E_m\rangle = 0$ as required to prove the theorem.

As a result of the last two theorems, the eigenvectors form a complete orthonormal set

$$B = \{|E_n\rangle = |n\rangle\} \quad (3.62)$$

Theorem 3.7 is important for quantum mechanics because it assures us that Hermitian operators, which correspond to physical observables, have eigenvectors that form a basis for the vector space of all physically meaningful wave functions. Therefore, every wave function can be expressed as a linear combination of the eigenvectors.

$$|\psi\rangle = \sum_n \beta_n |n\rangle \quad (3.63)$$

The basis set forms the elementary modes for the physical system. When we make a measurement of the physical observable corresponding to the Hermitian operator, the result will always be one of the eigenvalues and the particle will be found in one of the eigenstates. The full wave function collapses to one of the eigenvectors. The modulus-squared of an expansion coefficient for the wave function $|\beta_n|^2$ provides the probability of the wave function collapsing into a particular eigenstate $P(|\psi\rangle \rightarrow |n\rangle)$.

Next, examine what happens when two Hermitian operators $\hat{A}, \hat{B}$ commute. Each individual Hermitian operator must have a complete set of eigenvectors which means that each Hermitian operator generates a basis set for the vector space. The commutator $[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}$ indicates whether or not the operators commute. The next theorem shows that if the operators commute $[\hat{A}, \hat{B}] = 0$ then the operators $\hat{A}$ and $\hat{B}$ produce the same basis set for the vector space. The vectors space can be either a single space V or a direct product space $V \times W$.

THEOREM 3.8: A Single Basis Set for Commuting Hermitian Operators

Let $\hat{A}, \hat{B}$ be Hermitian operators that commute $[\hat{A}, \hat{B}] = 0$ then there exist eigenvectors $|\xi\rangle$ such that $\hat{A}|\xi\rangle = a_\xi|\xi\rangle$ and $\hat{B}|\xi\rangle = b_\xi|\xi\rangle$.

Proof: Assume that A has a complete set of eigenvectors. Let $|\xi\rangle$ be the eigenvectors of $\hat{A}$ such that

$$\hat{A}|\xi\rangle = a_\xi|\xi\rangle \quad (3.64)$$

Further assume that for each a_ξ there exists only one eigenvector $|\xi\rangle$. Consider

$$\hat{B}\hat{A}|\xi\rangle = \hat{B}a_\xi|\xi\rangle \quad (3.65)$$

But $\hat{B}\hat{A} = \hat{A}\hat{B}$ since $[\hat{A}, \hat{B}] = 0$ and so the right-hand side of this last equation becomes

$$a_\xi(\hat{B}|\xi\rangle) = \hat{B}a_\xi|\xi\rangle = \hat{B}\hat{A}|\xi\rangle = \hat{A}\hat{B}|\xi\rangle = \hat{A}(\hat{B}|\xi\rangle) \quad (3.66a)$$

Therefore, we see that the results of Equation 3.66a

$$\hat{A}(\hat{B}|\xi\rangle) = a_\xi(\hat{B}|\xi\rangle) \quad (3.66b)$$

require $\hat{B}|\xi\rangle$ to be an eigenvector of the operator $\hat{A}$ corresponding to the eigenvalue a_ξ . But there can only be one eigenvector for each eigenvalue. So

$$|\xi\rangle \sim \hat{B}|\xi\rangle$$

or, rearranging this expression and inserting a constant of proportionality b_ξ , we find

$$\hat{B}|\xi\rangle = b_\xi|\xi\rangle$$

This is an eigenvector equation for the operator B ; the eigenvalue is b_ξ .

THEOREM 3.9: Common Eigenvectors and Commuting Operators

As an inverse to Theorem 3.8, if the operators $\hat{A}, \hat{B}$ have a complete set of eigenvectors in common then $[A, B] = 0$.

Proof: First, for convenience, let us represent the common basis set by $|\xi\rangle = |a, b\rangle$ so that

$$\hat{A}|a, b\rangle = a|a, b\rangle \quad \text{and} \quad \hat{B}|a, b\rangle = b|a, b\rangle$$

Let $|v\rangle$ be an element of the direct product space of the eigenvectors for the operators $\hat{A}, \hat{B}$ so that it can be expanded as

$$|v\rangle = \sum_{ab} \beta_{ab} |a, b\rangle$$

then

$$\begin{aligned} AB|v\rangle &= \sum_{ab} \beta_{ab} AB|ab\rangle = \sum_{ab} \beta_{ab} Ab|ab\rangle \sum_{ab} \beta_{ab} ba|ab\rangle \\ &= \sum_{ab} \beta_{ab} aB|ab\rangle = \sum_{ab} \beta_{ab} Ba|ab\rangle \sum_{ab} \beta_{ab} BA|ab\rangle \\ &= BA|v\rangle \end{aligned}$$

This is true for all vectors in the vector space and so $\hat{A}\hat{B} = \hat{B}\hat{A}$.

3.10.2 DIRECT PRODUCT SPACE

Now let us make a comment on direct product spaces. There can be two reasons why the operators commute. First, we know from Section 3.6, if $\hat{B} = f(\hat{A})$ then $[\hat{A}, \hat{B}] = 0$ since we can Taylor expand $\hat{B} = f(\hat{A})$. Second, the operators can commute because they refer to separate Hilbert spaces. For example, $\hat{A}: V \rightarrow V$ while $\hat{B}: W \rightarrow W$ where $V \neq W$.

For the first case, where $\hat{B} = f(\hat{A})$ the operators $\hat{A}, \hat{B}$ cannot be independent. In this case, the basis set $|\xi\rangle$ requires only one parameter say a . For example, consider two Hermitian operators related by $\hat{\mathcal{H}} = \hat{p}^2$ where $\hat{p}$ has eigenvectors $|p\rangle$ then $\hat{\mathcal{H}}$ must satisfy the equation $\hat{\mathcal{H}}|p\rangle = \hat{p}^2|p\rangle = p^2|p\rangle$. Therefore $|p\rangle$ must also be an eigenvector of $\hat{\mathcal{H}}$.

Consider the second case where $\hat{A}, \hat{B}$ refer to different vector spaces $\hat{A}: V \rightarrow V$ and $\hat{B}: W \rightarrow W$ where $V \neq W$. If $|v\rangle \in V$ and $|w\rangle \in W$ then $\hat{B}|v\rangle|w\rangle = |v\rangle\hat{B}|w\rangle$. In other words, the operator $\hat{B}: W \rightarrow W$ does not “see” anything referring to the V space.

What does this imply for the eigenvectors? We could write the eigenvectors given in the previous theorem as

$$|\xi\rangle = |a_\xi, b_\xi\rangle = |a, b\rangle = |a\rangle|b\rangle$$

so long as we keep track of which eigenvector goes with which operator.

$$\begin{aligned}\hat{A}|a\rangle|b\rangle &= (\hat{A}|a\rangle)|b\rangle = a|a\rangle|b\rangle \\ \hat{B}|a\rangle|b\rangle &= |a\rangle(\hat{B}|b\rangle) = b|a\rangle|b\rangle\end{aligned}$$

It should be clear that the set $\{|a\rangle|b\rangle\}$ forms a basis set for the direct product space.

$$\{|a\rangle|b\rangle\} = \{|a\rangle\} \otimes \{|b\rangle\}$$

where the space spanned by the eigenvectors of A and B are $B_A = \{|a\rangle\}$ and $B_B = \{|b\rangle\}$. Notice that we can consider the combined object $|\xi\rangle = |a\rangle|b\rangle$ as either a single basis vector for the direct product space or as two separate vectors for the spaces V, W . If we have an operator $\hat{O} = \hat{A}\hat{B}$ and $[\hat{A}, \hat{B}] = 0$ then the matrix of the operator $\hat{O}$ can be decomposed as the direct product matrix. Generally, in the course of work, commuting operators refer to “different” vector spaces.

3.11 EIGENVECTORS, EIGENVALUES, AND DIAGONAL MATRICES

As will be seen, finding the eigenvectors for an operator is equivalent to making the operator diagonal. The eigenvectors of the operator provide the fundamental modes of the system such as in the study of electromagnetic fields and waves, and in the quantum theory. Our primary motivation is the quantum theory where the eigenvectors and eigenvalues of Hermitian operators provide the allowed “motions” and the possible observed values, respectively. Diagonal operators have the eigenvalues as the diagonal matrix elements, which makes for easy computation. However, one does not always a-priori choose the proper basis set for a vector space that renders an operator of interest diagonal.

After an initial discussion for the motivation of making matrices diagonal, the section then discusses the techniques and theory for making a matrix diagonal.

3.11.1 MOTIVATION FOR DIAGONAL MATRICES

The previous section shows that the set of eigenvectors of the Hermitian operator $\hat{\mathcal{H}}: V \rightarrow V$

$$B_E = \{|E_n\rangle \text{ such that } \hat{\mathcal{H}}|E_n\rangle = E_n|E_n\rangle\}$$

forms a complete set of orthonormal vectors for the Hilbert space. The set of eigenvectors can be used as the basis set for the vector space V . One can see that the set B_E also provides a diagonal form for the Hermitian operator $\hat{\mathcal{H}}: V \rightarrow V$ by starting with the eigenvalue equation

$$\hat{\mathcal{H}}|E_n\rangle = E_n|E_n\rangle \quad (3.67)$$

then operating with $\langle E_m|$ on the left-hand side of Equation 3.67 to find the matrix

$$\underline{\mathcal{H}} = \begin{bmatrix} E_1 & 0 & 0 & \cdots \\ 0 & E_2 & 0 & \cdots \\ 0 & 0 & E_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (3.68)$$

Notice that the eigenvalues appear on the diagonal of the matrix. This last equation is equivalent to expanding the operator as

$$\hat{\mathcal{H}} = \sum_n E_n |E_n\rangle \langle E_n| \quad (3.69)$$

The eigenvectors and eigenvalues for the system. Often one has an initial basis set for which the operator $\hat{\mathcal{H}}$ is not diagonal. In such a case, if one defines a unitary operator $\hat{u}$ that rotates the set of eigenvectors into the initial basis set, then the corresponding rotated operator $\hat{\mathcal{H}}_D = \hat{u} \hat{\mathcal{H}} \hat{u}^\dagger$ will have a diagonal form. One can imagine that the eigenvectors are rotated into the x, y, z (etc.) axes. Then, the eigenvectors form the basis set and the operator $\hat{\mathcal{H}}$ must be diagonal. We will see this in detail in Section 3.11.4.

In quantum theory, the energy represented by the Hamiltonian $\hat{\mathcal{H}}$ forms the primary quantity of interest (we are lucky that the same symbol stands for both Hamiltonian and Hermitian). For example, the wavelength of an emitted photon can be predicted by knowing the difference in energy for two atomic levels. Now if we have a diagonal form for the operator $\hat{\mathcal{H}}$, then not only do we have the simplest form, but we can also determine the energy of a given state at a glance. For example, Equation 3.68 immediately shows that the energy of the state $|E_2\rangle$ is E_2 .

People are also interested in complete sets of commuting operators $\{\hat{\mathcal{H}}, \hat{A}, \hat{B}, \dots\}$. These are the operators that completely described the physical system (as much as possible) and they all have common eigenvectors. So, in general, we prefer to have a basis set for which the operators $\{\hat{\mathcal{H}}, \hat{A}, \hat{B}, \dots\}$ are all diagonal. These operators must all commute in order to have simultaneous eigenvectors. One can easily see this when the operators all have the diagonal form.

The next section examines the main questions.

1. How do we find eigenvectors of a matrix?
2. How do we diagonalize a matrix?
3. What connection does the diagonalization procedure have with the operators.

Example 3.45

If the Hermitian operator $\hat{\mathcal{H}}$ represents energy and it has two eigenvectors $|E_1\rangle, |E_2\rangle$ then find the average energy for the state

$$|f\rangle = \frac{1}{\sqrt{2}}|E_1\rangle + \frac{1}{\sqrt{2}}|E_2\rangle$$

SOLUTION

In this example, the function $|f\rangle$ is decomposed into the sum of two energy eigenstates. The expected energy of a particle in the state $|f\rangle$ becomes

$$\begin{aligned}\langle f|\hat{H}|f\rangle &= \langle f|\hat{H}\left\{\frac{1}{\sqrt{2}}|E_1\rangle + \frac{1}{\sqrt{2}}|E_2\rangle\right\} = \frac{1}{\sqrt{2}}\langle f|\hat{H}|E_1\rangle + \frac{1}{\sqrt{2}}\langle f|\hat{H}|E_2\rangle \\ &= E_1 \frac{1}{\sqrt{2}}\langle f|E_1\rangle + E_2 \frac{1}{\sqrt{2}}\langle f|E_2\rangle = \frac{E_1 + E_2}{2}\end{aligned}$$

This is clearly the correct answer because the state $|f\rangle$ is an equal mixture of the states $|E_1\rangle$ and $|E_2\rangle$; therefore, we expect an equal mixture of the corresponding energies.

3.11.2 EIGENVECTORS AND EIGENVALUES

This section demonstrates the technique for finding the eigenvalues and eigenvectors for a 2×2 matrix. Suppose

$$\underline{H} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$$

Find the vectors $|v\rangle$ that satisfy the operator equation $\hat{H}|v\rangle = \lambda_v|v\rangle$ or in matrix notation

$$\underline{H}\underline{v} = \lambda_v\underline{v}$$

where the eigenvalues λ_v correspond to the eigenvectors $\underline{v}$.

The eigenvectors $\underline{v}$ are specified by the components ξ and η in the vector

$$\underline{v} = \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

The eigenvector equation is

$$\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \lambda \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

or, replacing λ with $\lambda\mathbf{1}$ where

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

these matrix equations can be written as

$$(\hat{H} - \lambda\mathbf{1})|v\rangle = 0 \quad \text{or} \quad \left[\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right] \begin{pmatrix} \xi \\ \eta \end{pmatrix} = 0 \quad (3.70)$$

Now work with the matrix equation. The set of equations for ξ , η , λ has a “nontrivial” solution so long as

$$\det \begin{pmatrix} 0 - \lambda & 2 \\ 2 & 0 - \lambda \end{pmatrix} = 0$$

As a note, if the determinant were not zero then an inverse matrix would exist and the components of the eigenvectors would be $\xi = 0$ and $\eta = 0$. The determinant equation provides two values for the eigenvalue λ (the determinant equation always provides the eigenvalues)

$$\lambda^2 - 4 = 0 \quad \rightarrow \quad \lambda = \pm 2$$

So there are two eigenvalues $\lambda = 2$ and $\lambda = -2$.

Next find the eigenvectors based on Equation 3.70. There are two cases for λ but both eigenvectors are obtained from the first line of Equation 3.70.

Case 1: $\lambda = 2$

The first line of Equation 3.70 gives

$$0\xi + 2\eta = \lambda\xi \quad \text{or equivalently} \quad \eta = \frac{\lambda}{2}\xi = \xi$$

Therefore, the eigenvector corresponding to the eigenvalue $\lambda = 2$ is

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \xi \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Case 2: $\lambda = -2$

Again, the first line of Equation 3.70 gives

$$0\xi + 2\eta = \lambda\xi \quad \text{or} \quad \eta = \frac{\lambda}{2}\xi = -\xi$$

Therefore, the eigenvector corresponding to the eigenvalue $\lambda = -2$ is

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \xi \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

As with all Sturm–Liouville problems, we have an arbitrary constant in each case. We choose the respective values of ξ to normalize the column vectors; i.e., for both Cases 1 and 2,

$$\xi = \frac{1}{\sqrt{2}}$$

The two eigenvectors are then

$$\left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}$$

which correspond to the eigenvalues $+2$ and -2 , respectively.

3.11.3 DIAGONALIZE A MATRIX

To diagonalize an operator $\hat{H}$ (or equivalently, its matrix), apply a similarity transformation to $\hat{H}$ where the similarity transformation represents a change of basis using a unitary operator $\hat{u}$. We know that the operator $\hat{H}$ will be diagonal for a basis consisting of the eigenvectors. Then to make the operator diagonal, one defines a new basis set obtained by rotating the eigenvectors into the original basis vectors. In this manner, the eigenvectors become the new basis set.

The diagonal form of the operator must be

$$\hat{H}_D = \hat{u} \hat{H} \hat{u}^+$$

where the reader should note the usual order of $\hat{u}$ and $\hat{u}^+$. The operator $\hat{u}$ represents a particular transformation that incorporates the eigenvectors of $\hat{H}$. This section shows how to diagonalize the operator $\hat{H}$ by diagonalizing the corresponding matrix $\underline{H}$. So we want a unitary matrix $\underline{u}$ (or equivalently $\underline{u}^+$) that provides the diagonal matrix $\underline{H}_D$. As will be shown, the desired matrix $\underline{u}^+$ has columns consisting of the eigenvectors of the matrix $\underline{H}$

$$\underline{u}^+ = \left[\begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} \cdots \right] \quad \text{where the symbol} \quad \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \quad (3.71)$$

represents the “first” eigenvector written in columnar form, and so on.

Example 3.46

For $\underline{H} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$ in the previous example, find $\underline{u}^+$.

SOLUTION

We found the eigenvectors in Section 3.11.2. The unitary matrix that diagonalizes the matrix $\underline{H}$ must be

$$\underline{u}^+ = \left[\begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} \right] = \left[\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

where $ev - 1$ and $ev - 2$ correspond to the eigenvalues 2, -2, respectively.

Now to prove the claim that the matrix $\underline{H}_D = \underline{u} \underline{H} \underline{u}^+$ is diagonal. In preparation for the demonstration, it is helpful to first show that the unitary change of basis operator satisfies $\underline{u} \underline{u}^+ = 1$.

$$\underline{u} \underline{u}^+ = \begin{bmatrix} (ev - 1)^* \\ (ev - 2)^* \\ \vdots \end{bmatrix} \left[\begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} \cdots \right]$$

The matrix $\underline{u}$ is obtained from $\underline{u}^+$ by simply changing columns into rows (the transpose operation) and remembering to take the complex conjugate. Multiplying the two matrices together provides

$$\underline{u} \underline{u}^+ = \begin{bmatrix} (ev - 1)^* \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} & (ev - 1)^* \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} & \cdots \\ (ev - 2)^* \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} & (ev - 2)^* \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Using the facts that (1) eigenvectors corresponding to different eigenvalues must be orthogonal and (2) eigenvectors corresponding to the same eigenvalue must be normalized, we find the following two matrix relations

$$(ev - 1)^* \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} = \cdots = 0 \quad \text{and} \quad (ev - 1)^* \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} = 1$$

The other entries can be similarly handled. Therefore,

$$\underline{u} \underline{u}^+ = \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \underline{1}$$

as required. The operator $\hat{u}$ is unitary and also satisfies $\underline{u}^+ \underline{u} = \underline{1}$

Now show that the matrix $\underline{H}_D = \underline{u}^+ \underline{H} \underline{u}$ must be diagonal.

$$\underline{u} \underline{H} \underline{u}^+ = \begin{bmatrix} (ev - 1)^* \\ (ev - 2)^* \\ \vdots \end{bmatrix} \underline{H} \begin{bmatrix} \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} & \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} & \cdots \end{bmatrix}$$

The matrix $\underline{H}$ acts on each column vector

$$\begin{pmatrix} e \\ v \\ i \end{pmatrix} \quad \text{to give} \quad \underline{H} \begin{pmatrix} e \\ v \\ i \end{pmatrix} = \lambda_i \begin{pmatrix} e \\ v \\ i \end{pmatrix}$$

so that

$$\underline{u} \underline{H} \underline{u}^+ = \begin{bmatrix} (ev - 1)^* \\ (ev - 2)^* \\ \vdots \end{bmatrix} \left[\underline{H} \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \quad \underline{H} \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} \quad \cdots \right] = \begin{bmatrix} (ev - 1)^* \\ (ev - 2)^* \\ \vdots \end{bmatrix} \left[\lambda_1 \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \quad \lambda_2 \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} \quad \cdots \right]$$

Next, multiplying these two matrices yields

$$\underline{u} \underline{H} \underline{u}^+ = \begin{bmatrix} \lambda_1 (ev - 1)^* \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} & \lambda_2 (ev - 1)^* \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} & \cdots \\ \lambda_1 (ev - 2)^* \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} & \lambda_1 (ev - 2)^* \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

So $\underline{H}_D = \underline{u} \underline{H} \underline{u}^+$ must be diagonal and the diagonal elements must be the eigenvalues. The eigenvalue λ_1 corresponds to the first eigenvector in the matrix $\underline{u}$, and so on.

Example 3.47

Find $\underline{H}_D = \underline{u} \underline{H} \underline{u}^+$ for the previous example.

The previous example gives

$$\underline{u}^+ = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{for} \quad \underline{H} = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}$$

So

$$\underline{H}_D = \underline{u} \underline{H} \underline{u}^+ = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

Notice how the upper left-hand entry +2 is the eigenvalue corresponding to the eigenvector $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ which is the first column in $\underline{u} = \frac{1}{2} \left[\begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right]$.

Example 3.48

Find the set of basis vectors that diagonalizes the matrix

$$\underline{H} = \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}$$

and write the diagonal form of the matrix.

SOLUTION

As before, find the eigenvectors using

$$\underline{H} \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \lambda \begin{pmatrix} \xi \\ \eta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \lambda \begin{pmatrix} \xi \\ \eta \end{pmatrix} \quad (3.72)$$

where $|\nu\rangle = \xi|1\rangle + \eta|2\rangle$ is an eigenvector corresponding to the eigenvalue λ . The eigenvector equation can be written as

$$\begin{pmatrix} 1 - \lambda & i \\ -i & 1 - \lambda \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

For nontrivial ξ, η require

$$\det \begin{pmatrix} 1 - \lambda & i \\ -i & 1 - \lambda \end{pmatrix} = 0$$

which gives the eigenvalues $\lambda = 0, 2$. Next, determine the components of the eigenvectors using the top row of the resultant matrix from Equation 3.72.

$$\xi + i\eta = \lambda\xi$$

Now find η in terms of ξ for each eigenvalue λ .

$\lambda = 0$ gives $\eta = i\xi$ and

$\lambda = 2$ gives $i\eta = \xi$

The eigenvalues and the corresponding eigenvectors must be

$$\lambda_1 = 0 \leftrightarrow \underline{e}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} \quad \lambda_2 = 2 \leftrightarrow \underline{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} i \\ 1 \end{pmatrix}$$

Notice that we choose the arbitrary constant so as to normalize the eigenvectors. Next, diagonalize the matrix $\underline{H}$ using $\underline{H}_D = \underline{u} \underline{H} \underline{u}^+$ where the unitary matrix $\underline{u}^+$ has columns formed by $\underline{e}_1$ and $\underline{e}_2$

$$\underline{u} = \left[\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} i \\ 1 \end{pmatrix} \right] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

Therefore,

$$\underline{H}_D = \underline{u} \underline{H} \underline{u}^+ = \frac{1}{2} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$$

Notice that the order of the eigenvalues on the diagonal corresponds to the order of the column vectors $\underline{e}_1$ and $\underline{e}_2$ in the unitary matrix $\underline{u}^+$.

3.11.4 RELATION BETWEEN A DIAGONAL OPERATOR AND THE CHANGE-OF-BASIS OPERATOR

This section shows why a basis rotation can bring a Hermitian operator into diagonal form. In addition, it shows how the form of the unitary operator follows from the rotation.

Consider a Hilbert space with basis vectors $\{|\phi_i\rangle\}$ and suppose $\hat{H}$ represents a Hermitian operator with eigenvectors $|e^{(i)}\rangle$. The superscript distinguishes the vector from the a th component of the vector $\langle \phi_a | e^{(i)} \rangle = e_a^{(i)}$. The eigenvalue equation has the form:

$$\hat{H}|e^{(i)}\rangle = \lambda_i |e^{(i)}\rangle \quad (3.73)$$

If the basis set were switched from $\{|\phi_i\rangle\}$ to $\{|e^{(i)}\rangle\}$ then $\hat{H}$ would be diagonal in the new basis according to

$$\hat{H} = \sum_i \lambda_i |e^{(i)}\rangle \langle e^{(i)}| \rightarrow \underline{H} = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \\ \vdots & & \ddots \end{bmatrix}$$

Switching the basis vectors is equivalent to rotating them using the unitary operator $\hat{u}$ as illustrated in Figure 3.18. The operator has the form:

$$\hat{u} = \sum_i |\phi_i\rangle \langle e^{(i)}| \quad \text{or} \quad \hat{u}^+ = \sum_i |e^{(i)}\rangle \langle \phi_i| \quad (3.74)$$

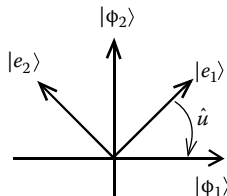


FIGURE 3.18 The operator $\underline{u}$ maps one basis set into another.

What needs to be done to the original Hermitian operator $\hat{H}$ to make it diagonal? One should “rotate” the eigenvectors into the basis set using $\hat{u}$ so that the eigenvectors become the basis set for which the operator $\hat{H}$ is diagonal. This is equivalent to imagining the eigenvectors become the new x -, y -, z -axes (and so on). The rotation of the basis set in Figure 3.18 can be related to a rotation of the operator. Equation 13.73 produces

$$\hat{H}|e^{(i)}\rangle = \lambda_i|e^{(i)}\rangle \rightarrow \hat{H}\hat{u}^+|\phi_i\rangle = \lambda_i\hat{u}^+|\phi_i\rangle \rightarrow \hat{u}\hat{H}\hat{u}^+|\phi_i\rangle = \lambda_i|\phi_i\rangle \quad (3.75)$$

This last results clearly show the new Hermitian operator

$$\hat{H}_D = \hat{u}\hat{H}\hat{u}^+ \quad \hat{H}_D = \sum_i \lambda_i|\phi_i\rangle\langle\phi_i| \quad (3.76)$$

must be diagonal in the original basis set.

Now we demonstrate the matrix form of the unitary operator $\hat{u}^+$ as given by Equation 3.70. The matrix elements of Equation 3.76 can be found in the usual manner.

$$\begin{aligned} \hat{u}^+ &= \sum_i |e^{(i)}\rangle\langle\phi_i| \\ (u^+)_{11} &= \langle\phi_1|\hat{u}^+|\phi_1\rangle = \sum_i \langle\phi_1|e^{(i)}\rangle\langle\phi_i|\phi_1\rangle = e_1^{(1)} \quad \text{and} \quad u_{21}^+ = e_2^{(1)}, \text{ etc.} \end{aligned} \quad (3.77)$$

This procedure shows that the first column of $\underline{u}^+$ consists of the components of the first column vector $\underline{e}^{(1)}$. Similar considerations apply to the other columns.

Finally the new Hermitian operator $\hat{H}_D$ can be shown to be diagonal using the operator $\hat{u}$ as opposed to the procedure for Equation 3.76.

$$\hat{H}_D = \hat{u}\hat{H}\hat{u}^+ = \left[\sum_i |\phi_i\rangle\langle e^{(i)}| \right] \hat{H} \left[\sum_j |\phi_j\rangle\langle e^{(j)}| \right]^+ = \left[\sum_i |\phi_i\rangle\langle e^{(i)}| \right] \hat{H} \left[\sum_j |e^{(j)}\rangle\langle\phi_j| \right] \quad (3.78)$$

The last equality in Equation 3.78 shows how the operator becomes diagonal. The mapping provided by $\hat{u}$ exposes the eigenvectors to the operator $\hat{H}$ for which it is already diagonal. To finish, use Equation 3.73 to find

$$\hat{H}_D = \sum_{i,j} |\phi_i\rangle\langle e^{(i)}|\hat{H}|e^{(j)}\rangle\langle\phi_j| = \sum_{i,j} \lambda_i|\phi_i\rangle\langle e^{(i)}|e^{(j)}\rangle\langle\phi_j| = \sum_i \lambda_i|\phi_i\rangle\langle\phi_i|$$

which clearly has the diagonal form.

3.12 THEOREMS FOR HERMITIAN OPERATORS

Given the importance of Hermitian operators for quantum theory, the present section discusses common theorems for Hermitian operators and provides alternate methods for determining when an operator is Hermitian. Previous sections provide the basic definition of Hermitian operators and show that they have real eigenvalues and orthogonal eigenvectors. The present section builds on this foundation to show how bounded Hermitian operators produce basis sets; that is, the set of eigenvectors is complete in the Hilbert space. As a result of these properties, the Hermitian operator is used to represent physical observables.

3.12.1 COMMON THEOREMS

We now present some basic theorems for operators based, in large part, on the references (in particular, refer to T.D. Lee's book). We will show that Hermitian operators produce complete sets of eigenvectors so that one can confidently use the eigenvectors as a basis set (which is complete and orthonormal).

THEOREM 3.10: A Test for the Zero Operator

If $\hat{O}$ is a linear operator on a Hilbert space then $\hat{O} = 0$ iff (if-and-only-if) $\langle v|\hat{O}|w\rangle = \langle v|\hat{O}w\rangle = 0$ for all vectors (or functions) v, w in the Hilbert Space

Proof: $(\Rightarrow) \hat{O} = 0 \rightarrow \langle v|\hat{O}w\rangle = \langle v|0\rangle = 0$

$(\Leftarrow)$ If $\langle v|\hat{O}w\rangle = 0$ for all f, g in the Hilbert space then take the special case $v = \hat{O}w$ to get $\langle \hat{O}w|\hat{O}w\rangle = 0$. Therefore, by the definition of inner product from Section 3.2, we must have $\hat{O}w = 0$ for every w . Therefore, by definition of the zero operator, we must have $\hat{O} = 0$.

THEOREM 3.11: A Test for the Zero Hermitian Operator

If $\hat{H}$ is a Hermitian linear operator in a Hilbert space, then
 $\hat{H} = 0 \Leftrightarrow \langle v|\hat{H}v\rangle = 0$ for every vector v in the Hilbert space.

Proof: $(\Rightarrow) \hat{H} = 0 \Rightarrow \langle v|\hat{H}v\rangle = \langle v|0\rangle = 0$

$(\Leftarrow)$ We will show two results that hold for all vectors x, y in the Hilbert space, namely

- a. $\langle v|\hat{H}w\rangle + \langle v|\hat{H}w\rangle = 0$ and
- b. $\langle v|\hat{H}w\rangle - \langle w|\hat{H}v\rangle = 0$

For then, by addition, $\langle v|\hat{H}w\rangle = 0$ for all v, w in the Hilbert space. Therefore by Theorem 3.10, we have $\hat{H} = 0$

To show result (a), we will require the starting assumption $\langle v|\hat{H}w\rangle = 0$ for x in the Hilbert space. Note that if v, w are in the Hilbert space, then so is $v + w$. Therefore, by assumption, we must have

$$0 = \langle v + w|\hat{H}(v + w)\rangle = \langle v|\hat{H}v\rangle + \langle v|\hat{H}w\rangle + \langle w|\hat{H}v\rangle + \langle w|\hat{H}w\rangle$$

Also note by assumption $\langle v|\hat{H}v\rangle = \langle w|\hat{H}w\rangle = 0$, so that $\langle v|\hat{H}w\rangle + \langle w|\hat{H}v\rangle = 0$ as required for (a).

To show (b), replace the vector w with the complex vector iw in part (a) to get $\langle v|\hat{H}iw\rangle + \langle iw|\hat{H}v\rangle = 0$. Factoring out the complex i using the complex conjugate implicit in the bra, we find $\langle v|\hat{H}w\rangle - \langle w|\hat{H}v\rangle = 0$.

We replaced w with iw but the complex quantity iw does not have meaning in a "real" Hilbert space. So to show $\langle v|\hat{H}v\rangle = 0 \rightarrow \hat{H} = 0$ for a real Hilbert space, we use result (a) as follows:

$$0 = \langle v|\hat{H}w\rangle + \langle w|\hat{H}v\rangle = \langle v|\hat{H}w\rangle + \langle \hat{H}^+w|v\rangle$$

where the last step follows from the definition of adjoint. Next, using the definition of Hermitian and the fact that the adjoint of the inner product reverses the order and includes a complex conjugate, we find

$$0 = \langle v|\hat{H}w\rangle + \langle \hat{H}w|v\rangle = \langle v|\hat{H}w\rangle + \langle v|\hat{H}w\rangle^* = \langle v|\hat{H}w\rangle + \langle v|\hat{H}w\rangle$$

where the last step follows for real Hilbert spaces. Therefore, as a result, we have $\langle v|\hat{H}w\rangle = 0$ for all v, w in the Hilbert Space. Now use Theorem 3.10 to conclude $\hat{H} = 0$.

THEOREM 3.12: A Test for Hermiticity

A linear operator $\hat{O}$ on a Hilbert space is Hermitian provided $\langle v|\hat{O}v\rangle = \text{Real}$ for all vectors x in the Hilbert space.

Proof: $\langle v|\hat{O}v\rangle$ real means that

$$\langle v|\hat{O}v\rangle = \langle v|\hat{O}v\rangle^* = \langle v|\hat{O}v\rangle^+ = \langle \hat{O}v|v\rangle = \langle v|\hat{O}^+v\rangle$$

therefore, $0 = \langle v|\hat{O} - \hat{O}^+|v\rangle$ for every x in the Hilbert space. The last theorem then gives $\hat{O} = \hat{O}^+$.

3.12.2 BOUNDED HERMITIAN OPERATORS HAVE COMPLETE SETS OF EIGENVECTORS

This section shows that Hermitian operators bounded from below have complete sets of eigenvectors. Therefore, Hermitian operators produce complete sets of orthonormal vectors that can be taken as a basis set for the Hilbert space. The development follows that in T.D. Lee's book listed in the chapter references.

Definition 3.1: Bounded Operator

Let $\hat{H}$ be a Hermitian operator in a Hilbert space with a complete orthonormal set (basis) given by

$$B = \{ |\phi_i\rangle: i = 1, 2, \dots \}$$

The operator $\hat{H}$ is bounded from below if there exists a constant C (note that it will be a real number for a Hermitian operator) such that for all vectors $|f\rangle$ in the Hilbert space

$$\frac{\langle f|\hat{H}|f\rangle}{\langle f|f\rangle} > C \quad (3.79a)$$

The vector $|f\rangle$ in this case is not necessarily normalized. However, note that the vector $|\psi\rangle = |f\rangle/\|f\|$ is normalized to one and that

$$\langle \psi|\hat{H}|\psi\rangle = \frac{\langle f|\hat{H}|f\rangle}{\|f\|^2} = \frac{\langle f|\hat{H}|f\rangle}{\langle f|f\rangle} > C \quad (3.79b)$$

indicates that one can focus on vectors normalized to one (rather than the full vector space) for the bounded property. In effect, one looks to see how the operator affects vectors terminating on the “unit sphere.”

Example 3.49

Suppose $\hat{H}$ is bounded from below, show $-\hat{H}$ is bounded from above.

SOLUTION

$$\frac{\langle f|\hat{H}|f\rangle}{\langle f|f\rangle} > C \Rightarrow \frac{\langle f|-\hat{H}|f\rangle}{\langle f|f\rangle} > C \Rightarrow \frac{\langle f|\hat{H}|f\rangle}{\langle f|f\rangle} < -C$$

Example 3.50

Suppose $\hat{H}$ is Hermitian with eigenvectors $\{|n\rangle$ for $n=0, 1, \dots\}$ and $\hat{H}|n\rangle = E_n|n\rangle$ and $E_0 \leq E_1 \leq E_2 \leq \dots$. Show E_0 must be the lower bound. Assume for this example that the eigenvectors form a basis although we shall show this for some special cases later in this section.

SOLUTION

In view of Equation 3.79b, consider only those vectors normalized to one. Then consider an arbitrary vector $|\psi\rangle$ (normalized to one) that can be expanded in the eigenvectors since we assume that they form a basis.

$$|\psi\rangle = \sum_n \beta_n |n\rangle$$

Equation 3.79b now provides

$$\langle \psi|\hat{H}|\psi\rangle = \sum_n E_n |\beta_n|^2$$

This represents an average and as such the average is always larger than the smallest value going into the average. So we have

$$\langle \psi|\hat{H}|\psi\rangle = \sum_n E_n |\beta_n|^2 \geq E_0$$

“Lower bounded” just means that the average of an operator must always be larger than some number C for all wave functions f , i.e., $\langle f|\hat{H}|f\rangle \geq C$. For energy, it means the average energy for every possible configuration of the system cannot approach $-\infty$.

THEOREM 3.13: The Minimum of Lower Bounded Operators

Let $\hat{H}: V \rightarrow V$ be a Hermitian operator in a vector space V spanned by the eigenvectors $\{|\phi_a\rangle = |E_a\rangle = |a\rangle\}$ where $\hat{H}|E_a\rangle = E_a|E_a\rangle$. Assume that the eigenvalues can be arranged as $E_0 \leq E_1 \leq \dots$ which also orders the eigenvectors. The assumption holds for Hermitian operators since the eigenvalues are real numbers which can always be ordered. The minimum of the ratio

$$E = \frac{\langle f|\hat{H}|f\rangle}{\langle f|f\rangle} \quad (3.80)$$

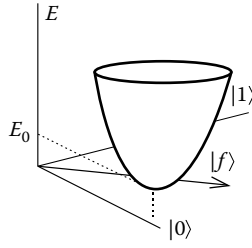


FIGURE 3.19 E is a minimum when f coincides with the zeroth eigenvector.

must be

1. E_0 , if $|f\rangle$ can be any vector in the space spanned by $|E_0\rangle, |E_1\rangle, |E_2\rangle, |E_3\rangle \dots$
2. E_1 , if $|f\rangle$ can be any vector in the subspace spanned by $|E_1\rangle, |E_2\rangle, |E_3\rangle \dots$
3. E_n , if $|f\rangle$ can be any vector in the subspace spanned by $|E_n\rangle, |E_{n+1}\rangle, \dots$

Proof: Let $|f\rangle$ be an arbitrary vector in the Hilbert space. We want the minimum value of E in Equation 3.80. Figure 3.19 suggests that we should look for the vector $|f\rangle$ that makes E a minimum. If the vector $|f\rangle$ points to the minimum of the energy, then a small change in the vector $|f\rangle$, namely $\delta(|f\rangle) = |\delta f\rangle$, must produce a change in energy δE of approximately zero. Note that changing the vector by a small amount produces a new vector given by $|w\rangle = |f\rangle + |\delta f\rangle$ so that $|\delta f\rangle = |w\rangle - |f\rangle$ (we do not use this though).

Let us calculate

$$0 = \delta E = \delta \frac{\langle f | \hat{H} | f \rangle}{\langle f | f \rangle} = \frac{\langle \delta f | \hat{H} | f \rangle + \langle f | \hat{H} | \delta f \rangle}{\langle f | f \rangle} - \frac{\langle f | \hat{H} | f \rangle}{\langle f | f \rangle} \left[\frac{\langle \delta f | f \rangle}{\langle f | f \rangle} + \frac{\langle f | \delta f \rangle}{\langle f | f \rangle} \right]$$

Substituting Equation 3.80 produces

$$\delta E = \frac{\langle \delta f | \hat{H} - E | f \rangle + \langle f | \hat{H} - E | \delta f \rangle}{\langle f | f \rangle} = 0 \quad (3.81)$$

Let us set the small variation of the vector to be $|\delta f\rangle = \varepsilon |f\rangle$ where ε is a small real number (for our purposes, a real quantity will serve the purpose). Equation 3.81 becomes

$$\delta E = \frac{2\varepsilon \langle f | \hat{H} - E | f \rangle}{\langle f | f \rangle} = 0$$

This requires $\hat{H}|f\rangle = E|f\rangle$ which is the eigenvalue equation. We already know the eigenvalues and eigenvectors namely $E_0 \leq E_1 \leq \dots$ and $|E_0\rangle, |E_1\rangle, \dots$. Therefore, E must be one of $E_0 \leq E_1 \leq \dots$. The minimum value of the ratio must be

$$E = \frac{\langle f | \hat{H} | f \rangle}{\langle f | f \rangle} \geq \frac{\langle E_n | \hat{H} | E_n \rangle}{\langle E_n | E_n \rangle} = E_n \geq E_0$$

which proves part 1 of the theorem.

The second part of the theorem is identical except the vector space does not include $|E_0\rangle$ and so we do not include E_0 as a lower bound to the sequence.

Definition 3.2: Completeness for an Infinite Discrete Set

A set of basis vectors $\{|n\rangle\}$ is complete if for any vector $|v\rangle$ there exists constants β_n such that if

$$|v\rangle = \sum_{n=0}^q \beta_n |n\rangle + |R_q\rangle \quad (3.82a)$$

where $|R_q\rangle$ represents a remainder vector (i.e., small difference vector) and q is an integer, then

$$\lim_{q \rightarrow \infty} \langle R_q | R_q \rangle = 0 \quad (3.82b)$$

This definition applies to either an infinite or finite set of discrete basis vectors. It requires the summation over the basis vectors to converge to the arbitrary vector in the space. The convergence then requires the remainder (i.e., its length) to approach zero.

THEOREM 3.14: Hermiticity and Completeness

If a Hermitian operator $\hat{H}$ is bounded from below (but not from above) then the set of (normalized) eigenvectors $\{|n\rangle\}$ forms a complete basis.

Proof: The operator $\hat{H}$ satisfies Theorem 3.13 where $\hat{H}|n\rangle = E_n|n\rangle$ and the eigenvalues are arranged so that $E_0 < E_1 < \dots < E_q < \dots$. The eigenvectors are normalized and satisfy the orthonormality condition $\langle m | n \rangle = \delta_{mn}$.

Let $|f\rangle$ be an arbitrary vector in the Hilbert space as required in the definition for the lower bound. As usual, let $\beta_n = \langle n | f \rangle$; however, $|f\rangle$ is not a priori normalized to one. The remainder vector (a.k.a. error vector) becomes $|R_q\rangle = |f\rangle - \sum_{n=0}^q \beta_n |n\rangle$. The theorem is proven by showing $\langle R_q | R_q \rangle \rightarrow 0$ as $q \rightarrow \infty$.

To start, one can verify that $\langle n | R_q \rangle = 0$ for $n \leq q$ and so Theorem 3.13 provides

$$\frac{\langle R_q | \hat{H} | R_q \rangle}{\langle R_q | R_q \rangle} \geq E_{q+1} \geq E_q \quad (3.83a)$$

Given that $\hat{H}$ does not have an upper bound, we must have $E_q \rightarrow \infty$ as $q \rightarrow \infty$. The infinite limit also requires there to exist an integer Q such that for all $q > Q$, one must find $E_q > 0$ and therefore

$$\langle R_q | R_q \rangle \leq \frac{\langle R_q | \hat{H} | R_q \rangle}{E_q} \quad (3.83b)$$

Using Equation 3.82a, one can easily show

$$\langle R_q | \hat{H} | R_q \rangle = \langle f | \hat{H} | f \rangle - \sum_{n=0}^q |\beta_n|^2 E_n \quad (3.84)$$

Combining the last two equations provides

$$\langle R_q | R_q \rangle \leq \frac{\langle f | \hat{H} | f \rangle}{E_q} - \frac{\sum_{n=0}^q |\beta_n|^2 E_n}{E_q} \quad (3.85)$$

The second summation in Equation 3.85 is a nonnegative number (the second term is negative when the minus sign is included) which can be seen as follows. The summation can be divided into two parts

$$\begin{aligned} \frac{\sum_{n=0}^{q>Q} |\beta_n|^2 E_n}{E_q} &= \frac{\sum_{n=0}^Q |\beta_n|^2 E_n}{E_q} + \frac{\sum_{n=Q+1}^q |\beta_n|^2 E_n}{E_q} \\ &= \frac{\sum_{n=0}^Q |\beta_n|^2 E_n}{E_q} + \left[|\beta_{Q+1}|^2 \frac{E_{Q+1}}{E_q} + \dots + |\beta_q|^2 \right] \end{aligned} \quad (3.86)$$

The first term in Equation 3.85 approaches zero as $E_q \rightarrow \infty$ since $\langle f | \hat{H} | f \rangle$ is a fixed. In the resulting Equation 3.86, the first term is negative but can be made arbitrarily small by choosing $q > Q'$ so that the second term in brackets dominates and the resulting full expression must then be nonnegative for $q > \text{Max}\{Q, Q'\}$.

Consequently, Equation 3.85 becomes

$$0 \leq \langle R_q | R_q \rangle \leq \frac{\langle f | \hat{H} | f \rangle}{E_q} \quad (3.87)$$

And finally, taking $E_q \rightarrow \infty$ as $q \rightarrow \infty$, we find the desired results of $\langle R_q | R_q \rangle \rightarrow 0$ as $q \rightarrow \infty$.

Example 3.51

Show for a Hilbert space of twice differentiable functions that the operator $\hat{H} = \hat{p}^2 + V(x)$ must produce a complete basis set where $\hat{p} = -i\partial_x$. Assume V is real and that $V(x) > 0$ for convenience.

SOLUTION

We must show that $\hat{H}$ is (1) Hermitian, (2) bounded from below, and (3) not bounded from above.

1. One can easily see that $\hat{H}$ is Hermitian since $V^+ = V^* = V$ and $(\hat{p}^2)^+ = (\hat{p})^+(\hat{p})^+ = \hat{p}^2$ as shown in previous sections in the present chapter.
2. $V(x)$ is bounded below since $\langle f | V | f \rangle = \int dx f^*(x) V f(x) = \int dx V(x) |f(x)|^2 \geq 0$ since V is nonnegative. Also, $\langle f | \hat{p}^2 | f \rangle = (\langle f | \hat{p} \rangle \langle \hat{p} | f \rangle) = (\langle f | \hat{p}^+ \rangle \langle \hat{p} | f \rangle) = (\hat{p} | f \rangle)^+ (\hat{p} | f \rangle) = \|\hat{p} | f \rangle\|^2 \geq 0$. Combining the results shows that $\hat{H}$ must be bounded from below.
3. Unbounded from above can be seen by choosing a family of functions $f(x) = e^{-x^2/\lambda^2}$ in the space so that $\langle f | \hat{p}^2 | f \rangle \sim e^{-x^2/\lambda^2} \frac{2}{\lambda^2} (2x^2 + 1)$ which approaches infinity for $\lambda \rightarrow 0$.

3.12.3 DERIVATION OF THE HEISENBERG UNCERTAINTY RELATION

Later chapters will discuss in detail how commuting Hermitian operators correspond to dynamical variables that can be simultaneously and precisely measured. This means that the measurement of one does not affect the measurement of the other and that in principle, repeated measurements produce the same values. In such a case, there is not any dispersion among the measured values which then requires the standard deviation to be zero. However, when two Hermitian operators do not commute, the measurement of the dynamical variable corresponding to one necessarily interferes with the measurement of the other. In this case, one does not find identical values with repeated sets of measurements and therefore finds nonzero standard deviation (Heisenberg uncertainty relation). We now consider mathematical statements leading to the Heisenberg uncertainty relation.

THEOREM 3.15

If two Hermitian operators $\hat{A}, \hat{B}$ commute then there exists a simultaneous set of basis functions $|a, b\rangle = |a\rangle|b\rangle$ such that

$$\hat{A}|a, b\rangle = a|a, b\rangle \quad \text{and} \quad \hat{B}|a, b\rangle = b|a, b\rangle \quad (3.88)$$

and vice versa.

Proof: Suppose $|\phi\rangle$ represents an eigenvector of $\hat{A}$ such that $\hat{A}|\phi\rangle = a|\phi\rangle$ so that one can also write $|\phi\rangle = |a\rangle$ if desired. Next show that $\hat{B}|\phi\rangle$ is an eigenvector of $\hat{A}$ corresponding to the eigenvalue of a .

$$\hat{A}(\hat{B}|\phi\rangle) = \hat{A}\hat{B}|\phi\rangle = \hat{B}\hat{A}|\phi\rangle = \hat{B}a|\phi\rangle = a(\hat{B}|\phi\rangle)$$

where the third term made use of the fact that the operators commute. By our naming convention for eigenvectors, we can write

$$\hat{B}|\phi\rangle \sim |a\rangle \quad (3.89)$$

since $\hat{B}|\phi\rangle$ is a vector corresponding to the eigenvalue a of the operator $\hat{A}$. However, we have another name for the vector $|\phi\rangle$, namely $|\phi\rangle = |a\rangle$. Now Equation 3.89 can be written as

$$\hat{B}|\phi\rangle \sim |\phi\rangle \quad \text{or} \quad \hat{B}|a\rangle \sim |a\rangle \quad (3.90a)$$

Therefore $|\phi\rangle = |a\rangle$ must also be an eigenvector of the operator $\hat{B}$. Suppose the eigenvalue is b then one finds

$$\hat{B}|\phi\rangle = b|\phi\rangle \quad \text{or} \quad \hat{B}|a\rangle = b|a\rangle \quad (3.90b)$$

According to our naming convention, the vector $|\phi\rangle$ can also be written in a manner to include an indication of the eigenvalue b as

$$|\phi\rangle = |a, b\rangle \quad (3.91)$$

We now show that two noncommuting Hermitian operators must always produce an uncertainty relation.

THEOREM 3.16

If two operators $\hat{A}, \hat{B}$ are Hermitian and satisfy the commutation relation $[\hat{A}, \hat{B}] = i\hat{C}$ then the observed values a, b of the operators must satisfy a Heisenberg uncertainty relation of the form $\sigma_a \sigma_b \geq \frac{1}{2} |\langle \hat{C} \rangle|$.

Proof: Consider the “real, positive number” defined by

$$\xi = \langle (\hat{A} + i\lambda\hat{B})\psi | (\hat{A} + i\lambda\hat{B})\psi \rangle \quad (3.92)$$

which we know to be a real and positive since the inner product provides the length of the vector. The vector, in this case, is defined by

$$|(\hat{A} + i\lambda\hat{B})\psi\rangle = (\hat{A} + i\lambda\hat{B})|\psi\rangle$$

We assume that λ is a real parameter. Now working with the number ξ and using the definition of adjoint, namely

$$\langle\hat{O}f|g\rangle = \langle f|\hat{O}^+g\rangle$$

Equation 3.92 provides

$$\begin{aligned}\xi &= \langle\psi|(\hat{A} + i\lambda\hat{B})^+(\hat{A} + i\lambda\hat{B})\psi\rangle = \langle\psi|(\hat{A}^+ - i\lambda\hat{B}^+)(\hat{A} + i\lambda\hat{B})\psi\rangle \\ &= \langle\psi|(\hat{A} - i\lambda\hat{B})(\hat{A} + i\lambda\hat{B})\psi\rangle\end{aligned}$$

where the last step uses the Hermiticity of the operators $\hat{A}, \hat{B}$. Multiply the operator terms and suppress reference to the function (for convenience) to obtain

$$\xi = \langle\hat{A}^2\rangle - \lambda\langle\hat{C}\rangle + \lambda^2\langle\hat{B}^2\rangle \geq 0$$

which must hold for all values of the parameter λ . The minimum value of the positive real number ξ is found by differentiating with respect to the parameter λ .

$$\frac{\partial\xi}{\partial\lambda} = 0 \rightarrow \lambda = \frac{\langle\hat{C}\rangle}{2\langle\hat{B}^2\rangle}$$

The minimum value of the positive real number ξ must be

$$\xi_{\min} = \langle\hat{A}^2\rangle - \frac{1}{4}\frac{\langle\hat{C}\rangle^2}{\langle\hat{B}^2\rangle} \geq 0$$

Multiplying through by $\langle\hat{B}^2\rangle$ to find

$$\langle\hat{A}^2\rangle\langle\hat{B}^2\rangle \geq \frac{1}{4}\langle\hat{C}\rangle^2 \quad (3.93)$$

We could have assumed the quantities $\langle\hat{A}\rangle = \langle\hat{B}\rangle = 0$ and we would have been finished at this point. However, the commutator $[\hat{A}, \hat{B}] = i\hat{C}$ holds for the two Hermitian operators defined by

$$\hat{A} \rightarrow \hat{A} - \langle\hat{A}\rangle \quad \hat{B} \rightarrow \hat{B} - \langle\hat{B}\rangle$$

As a result, Equation 3.93 becomes

$$\langle(\hat{A} - \langle\hat{A}\rangle)^2\rangle\langle(\hat{B} - \langle\hat{B}\rangle)^2\rangle \geq \frac{1}{4}\langle\hat{C}\rangle^2$$

However, the terms in the angular brackets are related to the standard deviations σ_a, σ_b , respectively. We obtained the proof to the theorem by taking the square root of the previous expression

$$\sigma_a\sigma_b \geq \frac{1}{2}|\langle\hat{C}\rangle|$$

Notice that this Heisenberg uncertainty relation involves the absolute value of the expectation value of the operator C . By its definition, the operator C must be Hermitian and its expectation value must be real.

3.13 RAISING–LOWERING AND CREATION–ANNIHILATION OPERATORS

Raising and lowering operators are especially associated with quantum mechanics but they can also be used in boundary value problems. The raising operators map one basis vector to the next in the sequence while the lowering operator has the reverse effect. For the quantum mechanics, the raising operator essentially adds one quantum of energy while the lowering operator removes one quantum of energy. Sometimes these operators are also called promotion and demotion operators. Modern physics, chemistry, and electrical engineering also make use of the closely related creation and annihilation operators. The creation and annihilation operators are used to create a particle from the vacuum, and to destroy a particle and return it to the vacuum. The first few sections of discussion center on the raising and lowering operators (Figure 3.20).

3.13.1 DEFINITION OF THE LADDER OPERATORS

For this discussion, we assume that Hilbert space is spanned by the basis set

$$\{\phi_1 = |1\rangle, \phi_2 = |2\rangle, \dots\}$$

The set might arise as the set of eigenvectors for a Hermitian operator. In this case, we assume that the eigenvalues are arranged in ascending order

$$\lambda_1 < \lambda_2 < \dots$$

Notice that the ascending order of eigenvalues induces a natural order for the eigenvectors as shown by the numbers “1, 2, ...” in the ket symbols. For the Hamiltonian, the eigenvalues would be the allowed energies for the system and are therefore arranged from lowest energy to highest energy.

Raising and lowering operators (denoted by $\hat{a}^+$ and $\hat{a}$) map the Hilbert space V into itself (never $V \rightarrow W$). We will focus on a special set of ladder operators, namely those for the harmonic oscillator. These have special normalization. The raising operator $\hat{a}^+$ is defined by

$$\hat{a}^+|n\rangle = \sqrt{n+1}|n+1\rangle \quad (3.94a)$$

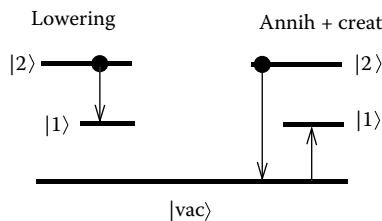


FIGURE 3.20 A comparison of the “lowering–raising” operation with the “creation–annihilation” operation as used in the quantum theory.

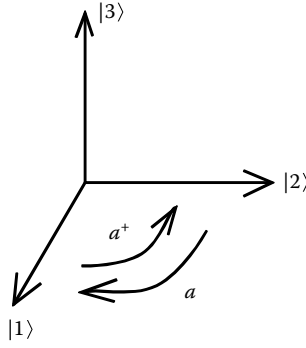


FIGURE 3.21 Ladder operators map basis vectors into adjacent basis vectors.

whereas, the lowering operator $\hat{a}$ is defined by

$$\hat{a}|n\rangle = \sqrt{n}|n-1\rangle \quad (3.94b)$$

where $\hat{a}|0\rangle = 0$.

In general, ladder operators can have any normalization. It is only necessary to map one basis vector into another as in

$$\hat{a}|n\rangle = C_1|n-1\rangle \quad \hat{a}^+|n\rangle = C_2|n+1\rangle \quad (3.95)$$

Here, $\hat{a}^+$ is the adjoint of the lowering operator $\hat{a}$ but the operators are not Hermitian so that $\hat{a}^+ \neq \hat{a}$. Notice that the “lowest” eigenvector is $|1\rangle$ so that $\hat{a}|1\rangle = 0$ (Figure 3.21). Chapter 5 will show general consideration to deduce Equations 3.94 from 3.95.

3.13.2 MATRIX AND BASIS-VECTOR REPRESENTATIONS OF THE RAISING AND LOWERING OPERATORS

Previous sections show that representing operators and vectors in terms of matrices provides a convenient computational tool. It is no longer necessary to refer to the explicit differential and functional forms of the operators and vectors. Now let us find the matrix representations of the raising and lowering operators $\hat{a}, \hat{a}^+$ associated with the Harmonic oscillator (to be discussed in Chapter 5).

Let the vector space V be spanned by the basis set

$$B_V = \{\phi_0 = |0\rangle, \phi_1 = |1\rangle, \dots\}$$

The matrix of an operator is obtained from the basic definition

$$\hat{T}|j\rangle = \sum_i T_{ij}|i\rangle \quad \text{for } |i\rangle, |j\rangle \in B_V$$

so that the matrix elements are

$$T_{ij} = \langle i|\hat{T}|j\rangle$$

For $\hat{a}$ then

$$\hat{a}^+|j\rangle = \sqrt{j+1}|j+1\rangle$$

so that

$$(\underline{a}^+)_{ij} = \langle i | \hat{a}^+ | j \rangle = \sqrt{j+1} \langle i | j+1 \rangle = \sqrt{j+1} \delta_{i,j+1}$$

Therefore, the matrix is

$$\underline{a}^+ = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ \sqrt{1} & 0 & 0 & 0 & \\ 0 & \sqrt{2} & 0 & 0 & \\ 0 & 0 & \sqrt{3} & 0 & \\ \vdots & & & \ddots & \end{bmatrix}$$

Note that the index i, j starts at 0, 0! Similarly,

$$a_{ij} = \langle i | \hat{a} | j \rangle = \sqrt{j} \langle i | j-1 \rangle = \sqrt{j} \delta_{i,j-1}$$

which has the matrix

$$\underline{a} = \begin{bmatrix} 0 & \sqrt{1} & 0 & 0 & \cdots \\ 0 & 0 & \sqrt{2} & 0 & \\ 0 & 0 & 0 & \sqrt{3} & \\ 0 & 0 & 0 & 0 & \ddots \\ \vdots & & & & \end{bmatrix}$$

Example 5.52

Using Equation 3.95 with $C_1 = C_2 = 1$, find $\underline{a}^+$ operating on the column vector for the first basis function

$$\underline{\phi}_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}$$

SOLUTION

$$\underline{a}^+ \underline{\phi}_0 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix} = \underline{\phi}_1$$

Next, let us find the basis vector expansion of the raising and lowering operators for the Harmonic oscillator. As usual, start with the definition of the matrix element

$$\hat{a}^+ |j\rangle = \sqrt{j+1} |j+1\rangle$$

Multiply both sides by $\langle j |$ on the right, and sum over j to get

$$\hat{a}^+ \sum_{j=0}^{\infty} |j\rangle \langle j| = \sum_{j=0}^{\infty} \sqrt{j+1} |j+1\rangle \langle j|$$

Therefore, the closure relation provides

$$\hat{a}^+ = \sum_{j=0}^{\infty} \sqrt{j+1} |j+1\rangle \langle j| \quad (3.96)$$

which shows explicitly that $\hat{a}^+$ maps the basis vector $|j\rangle$ into the next one in the sequence $|j+1\rangle$.

The adjoint of Equation 3.96 provides

$$\hat{a} = \sum_{j=0}^{\infty} \sqrt{j+1} |j\rangle \langle j+1|$$

which can be rewritten by setting $i=j+1$

$$\hat{a} = \sum_{i=1}^{\infty} \sqrt{i} |i-1\rangle \langle i|$$

Example 3.53

What is the basis-vector representation of $\hat{N} = \hat{a}^+ \hat{a}$ for the Harmonic oscillator?

SOLUTION

$$\begin{aligned} \hat{N} = \hat{a}^+ \hat{a} &= \left(\sum_{m=0}^{\infty} \sqrt{m+1} |m+1\rangle \langle m| \right) \left(\sum_{n=0}^{\infty} \sqrt{n} |n-1\rangle \langle n| \right) = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} \sqrt{m+1} \sqrt{n} |m+1\rangle \langle m| n-1\rangle \langle n| \\ &= \sum_{n=1}^{\infty} \sqrt{(n-1)+1} \sqrt{n} |(n-1)+1\rangle \langle n| = \sum_{n=1}^{\infty} n |n\rangle \langle n| \end{aligned}$$

where the second line follows because $\langle m|n-1\rangle = \delta_{m, n-1} \rightarrow m=n-1$.

The sum can start at $n=0$ to find

$$\hat{N} = \hat{a}^+ \hat{a} = \sum_{n=0}^{\infty} n |n\rangle \langle n|$$

Notice that the eigenvalues of the number operator $\hat{a}^+ \hat{a}$ appear as the expansion coefficients and note that $\hat{a}^+ \hat{a}$ is diagonal.

3.13.3 RAISING AND LOWERING OPERATORS FOR DIRECT PRODUCT SPACE

Let V and W be two vector spaces spanned by the basis sets $B_V = \{|\phi_1\rangle, |\phi_2\rangle, \dots\}$ and $B_W = \{|\psi_1\rangle, |\psi_2\rangle, \dots\}$, respectively. The direct product space is spanned by

$$B = B_V \otimes B_W = \{|\phi_1\rangle|\psi_1\rangle, |\phi_1\rangle|\psi_2\rangle, \dots, |\phi_2\rangle|\psi_1\rangle, |\phi_2\rangle|\psi_2\rangle, \dots\}$$

If $\hat{a}, \hat{a}^+$ and $\hat{b}, \hat{b}^+$ are “Harmonic oscillator” lowering and raising operators for the vector spaces V and W , respectively, then combinations of the form $\hat{a}^+ \hat{b}$ act on the product space. For example,

$$\hat{a}^+ \hat{b} |\phi_3 \psi_5\rangle = \hat{a}^+ |\phi_3\rangle \hat{b} |\psi_5\rangle = \sqrt{3+1} |\phi_4\rangle \sqrt{5} |\psi_4\rangle = \sqrt{20} |\phi_4 \psi_4\rangle$$

These direct product space operators provide a convenient method of calculating the so-called Clebsch–Gordon coefficients as will be seen in Chapter 5.

3.14 TRANSLATION OPERATORS

Common mathematical operations such as rotating or translating coordinates are handled by operators in the quantum theory. Previous sections show that states transform by the application of a single unitary operator whereas “operators” transform through a similarity transformation. The translation through the spatial coordinate x provides a standard example. Every operation in physical space has a corresponding operation in the Hilbert space.

3.14.1 EXPONENTIAL FORM OF THE TRANSLATION OPERATOR

Let $\hat{x}$ and $\hat{p}$ be the position operator and an operator defined in terms of a derivative

$$\hat{p} = \frac{1}{i} \frac{\partial}{\partial x}$$

which is the “position” representation of $\hat{p}$ and $i = \sqrt{-1}$. The position representation of $\hat{x}$ is x . The operator $\hat{p}$ is Hermitian (note that $\hat{p}$ is the momentum operator from quantum theory except the $\hbar$ has been left out of the definition given above). The coordinate kets satisfy $\hat{x}|x\rangle = x|x\rangle$ and the operators satisfy $[\hat{x}, \hat{p}] = [\hat{x}\hat{p} - \hat{p}\hat{x}] = i$ as can be easily verified

$$[\hat{x}, \hat{p}]f(x) = [\hat{x}\hat{p} - \hat{p}\hat{x}]f(x) = x \frac{1}{i} \frac{\partial}{\partial x} f - \frac{1}{i} \frac{\partial}{\partial x} (xf) = x \frac{1}{i} \frac{\partial}{\partial x} f - \frac{x}{i} \frac{\partial f}{\partial x} - \frac{1}{i} f = if$$

comparing both sides, we see that the “operator” equation $[\hat{x}, \hat{p}] = i$ holds. The commutator being nonzero defines the so-called conjugate variables. The translation operator uses products of the conjugate variables. The operator $\hat{p}$ is sometimes called the generator of translations. The Hamiltonian is the generator of translations in time.

This section shows that the exponential $\hat{T}(\eta) = e^{-i\eta\hat{p}}$ translates the coordinate system according to

$$\hat{T}(\eta)f(x) = e^{-i\eta\hat{p}}f(x) = f(x - \eta)$$

where $\hat{p} = \frac{1}{i} \frac{\partial}{\partial x}$ and η is real for this case. The proof starts by working with a small displacement ξ_k (Figure 3.22) and calculating the Taylor expansion about the point x

$$f(x - \xi_k) \cong f(x) - \frac{\partial f(x)}{\partial x} \xi_k + \cdots = \left(1 - \xi_k \frac{\partial}{\partial x} + \cdots\right) f(x)$$

Substituting the operator for the derivative

$$\hat{p} = \frac{1}{i} \frac{\partial}{\partial x}$$

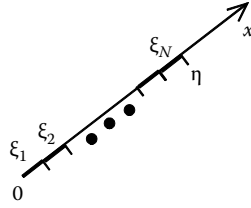


FIGURE 3.22 The total translation is divided into smaller translations.

gives

$$f(x - \xi_k) = \left(1 - \xi_k \frac{\partial}{\partial x} + \dots\right) f(x) = (1 - i\xi_k \hat{p} + \dots) f(x) = \exp(-i\xi_k \hat{p}) f(x)$$

Now, by repeated application of the infinitesimal translation operator, we can build up the entire length η

$$f(x + \eta) = \prod_k \exp(-i\xi_k \hat{p}) f(x) = \exp\left(-\sum_k i\xi_k \hat{p}\right) f(x) = \exp(-i\eta \hat{p}) f(x)$$

So the exponential with the operator $\hat{p}$ provides a translation according to

$$\hat{T}(\eta) f(x) = e^{-i\eta \hat{p}} f(x) = f(x - \eta)$$

Note that the translation operator is unitary $\hat{T}^+ = \hat{T}^{-1}$ for η real since $\hat{p}$ is Hermitian. It is easy to show $\hat{T}^+(-\eta) = \hat{T}(\eta)$. The operator $\hat{p}$ is the generator of translations. In the quantum theory, the momentum conjugate to the displacement direction generates the translation according to

$$\hat{T}(\eta) f(x) = e^{-i\eta \hat{p}/\hbar} f(x) = f(x - \eta) \quad \text{where} \quad \hat{p} = \frac{\hbar}{i} \frac{\partial}{\partial x}$$

Notice the extra factor of $\hbar$.

3.14.2 TRANSLATION OF THE POSITION OPERATOR

The position-coordinate ket $|x\rangle$ is an eigenvector of the position operator $\hat{x}$

$$\hat{x}|x\rangle = x|x\rangle$$

We can show that

$$\hat{T}(\eta) \hat{x} \hat{T}^+(\eta) = \hat{x} - \eta$$

where $\hat{T}(\eta) = e^{-i\eta \hat{p}}$ by using the operator expansion theorem in Section 3.6 with $\eta \rightarrow -\eta$

$$e^{-\eta \hat{A}} \hat{B} e^{\eta \hat{A}} = \hat{B} - \frac{\eta}{1!} [\hat{A}, \hat{B}] + \frac{\eta^2}{2!} [\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

Using $\hat{A} = i\hat{p}$ and the commutation relations $[\hat{x}, \hat{p}] = i$, we find

$$e^{-i\eta\hat{p}}\hat{x}e^{i\eta\hat{p}} = \hat{x} - \frac{\eta}{1!}[\hat{p}, \hat{x}] + \frac{\eta^2}{2!}[\hat{p}, [\hat{p}, \hat{x}]] + \cdots = \hat{x} - \eta$$

3.14.3 TRANSLATION OF THE POSITION-COORDINATE KET

The position-coordinate ket $|x\rangle$ is an eigenvector of the position operator $\hat{x}$

$$\hat{x}|x\rangle = x|x\rangle$$

What position-coordinate ket $|\phi\rangle$ is an eigenvector of the translated operator

$$\hat{T}(\eta)\hat{x}\hat{T}^+(\eta) = \hat{x} - \eta$$

that is, what is the state $|\phi\rangle = \hat{T}(\eta)|x\rangle$? The eigenvector equation for the translated operator $\hat{x}_T = \hat{T}\hat{x}\hat{T}^+$ is

$$\hat{x}_T|\phi\rangle = [\hat{T}(\eta)\hat{x}\hat{T}^+(\eta)][\hat{T}(\eta)|x\rangle] = \hat{T}(\eta)\hat{x}|x\rangle = x\hat{T}(\eta)|x\rangle = x|\phi\rangle$$

The second to last step follows because x is a number and not an operator in the present context. To continue, we know the translated operator is $\hat{x}_T = \hat{x} - \eta$ and therefore the previous equation provides

$$x|\phi\rangle = \hat{x}_T|\phi\rangle = (\hat{x} - \eta)|\phi\rangle = (\phi - \eta)|\phi\rangle$$

Comparing both sides, we see $\phi = x + \eta$ which therefore shows that the translated position vector is

$$|\phi\rangle = \hat{T}(\eta)|x\rangle = |x + \eta\rangle$$

3.14.4 EXAMPLE USING THE DIRAC DELTA FUNCTION

Show that

$$|\phi\rangle = \hat{T}(\eta)|x_o\rangle = |x_o + \eta\rangle$$

using the fact that the position-ket represents the Dirac delta function in Hilbert space

$$|x_o\rangle \equiv |\delta(\bullet - x_o)\rangle$$

where “ $\bullet$ ” represents the missing variable. If “ x ” is a coordinate on the x -axis then

$$\langle x|x_o\rangle = \delta(x - x_o)$$

As a side note, which will become more clear in Chapter 5, the operator $\hat{p}$ can be shown to have the x -representation of $\hat{p} \leftrightarrow \frac{1}{i}\frac{\partial}{\partial x} = -i\partial_x$ using the eigenvector equation $\hat{p}|p\rangle = p|p\rangle$ and assuming a plane wave representation of $|p\rangle$ as e^{ipx} . Then

$$\hat{p}|p\rangle = p|p\rangle \rightarrow \langle x|\hat{p}|p\rangle = p\langle x|p\rangle$$

Inserting the resolution of unity for the x -coordinate representation,

$$\int d\xi \langle x|\hat{p}|\xi\rangle \langle \xi|p\rangle = p\langle x|p\rangle \rightarrow \int d\xi \langle x|\hat{p}|\xi\rangle e^{ip\xi} = pe^{ipx}$$

Comparing the two sides, one can see that

$$\langle x|\hat{p}|\xi\rangle = \delta(\xi - x)(-i\partial_x)$$

will satisfy the previous equation. We symbolize the x -coordinate representation of $\hat{p}$ as $\hat{p}(x) = -i\partial_x$.

Returning to the problem at hand, the translation operator in the x -representation provides

$$\langle x|\hat{T}(\eta)|x_o\rangle = e^{-i\eta\hat{p}(x)}\langle x|x_o\rangle = e^{-i\eta\hat{p}(x)}\delta(x - x_o) = \delta(x - \eta - x_o) = \langle x|x_o + \eta\rangle$$

Evidently

$$\hat{T}(\eta)|x_o\rangle = |x_o + \eta\rangle$$

3.14.5 RELATION AMONG HILBERT SPACE AND THE 1-D TRANSLATION, AND LIE GROUP

The operator that translates functions and operators through 1-D displacements η in Euclidean space takes the form of a unitary operator $\hat{T}(\eta) = e^{-i\eta\hat{p}}$ in the abstract Hilbert space. In particular, representing the coordinates along the displacement direction by axes one realizes that the unitary operator is actually a rotation operator that maps one coordinate into another according to $\hat{T}(\eta): |x\rangle \rightarrow |x - \eta\rangle$. That is, the map changes x into $x - \eta$. Also notice the translations occur in one dimension whereas the Hilbert space has uncountably many dimensions.

One can show that the set of translations $\{\hat{T}(\eta) = e^{-i\eta\hat{p}}\}$ forms a group with η as a continuous parameter. That is, each group element corresponds to a different value of η . One might consider $\hat{p}$ as a basis vector for an operator vector space so that $\eta\hat{p}$ represents all of the vectors in the space— $\{\eta\hat{p}\}$ forms a vector space. The term “generator” of the group $\{\hat{T}(\eta) = e^{-i\eta\hat{p}}\}$ can refer to either a product $\eta\hat{p}$ or to a basis vector $\hat{p}$. Notice that all group elements smoothly connect to the identity by varying the parameter η . One often refers to the group as a “Lie Group.” For more than one generator, the commutation relations essentially determine the structure of the group.

3.14.6 TRANSLATION OPERATORS IN THREE DIMENSIONS

Representing the displacement by $\vec{\eta} = \tilde{x}\eta_x + \tilde{y}\eta_y + \tilde{z}\eta_z$ and the generator consists of three parts $\vec{p} = \tilde{x}\hat{p}_x + \tilde{y}\hat{p}_y + \tilde{z}\hat{p}_z$ where for the coordinate representation, we have $\hat{p}_x = -i\partial_x$, $\hat{p}_y = -i\partial_y$, $\hat{p}_z = -i\partial_z$. The representation of the group becomes $\{\hat{T}(\eta) = e^{-i\vec{\eta}\cdot\vec{p}}\}$ which consists of unitary operators (Figure 3.22).

3.15 FUNCTIONS IN ROTATED COORDINATES

This section shows how the form of a function changes under rotations. It then demonstrates a rotation operator.

3.15.1 ROTATING FUNCTIONS

If we know a function $f(x, y)$ in one set of coordinates (x, y) then what is the function $f'(x', y')$ for coordinates (x', y') that are rotated through an angle θ with respect to the first set (x, y) .

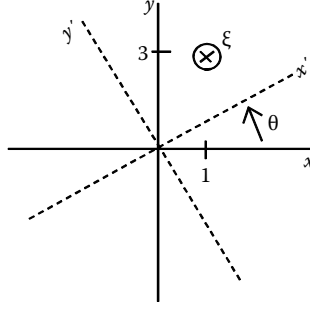


FIGURE 3.23 Rotated coordinates.

Consider a point ξ in space as indicated in the picture. The single point can be described by the primed or unprimed coordinate system. The key fact is that the equations linking the two coordinate systems describe the single point ξ . The equations for coordinate rotations are

$$\underline{r}' = \underline{R} \underline{r} \quad (3.97)$$

where

$$\underline{r}' = \begin{pmatrix} x' \\ y' \end{pmatrix} \quad \underline{R} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \quad \underline{r} = \begin{pmatrix} x \\ y \end{pmatrix} \quad (3.98)$$

and $\underline{r}'$ and $\underline{r}$ represent the single point ξ . Notice that the matrix differs by a minus sign from that discussed in Section 3.7.1 since Figure 3.23 relates one set of coordinates to another whereas Section 3.7.1 rotates vectors.

A functional value z associated with the point ξ is the same value regardless of the reference frame. Therefore, we require

$$z = f'(x', y') = f(x, y) \quad (3.99)$$

since (x', y') and (x, y) specify the same point ξ and the function, which does not rotate, must have a single value at the point ξ . We can write the last equation using Equation 3.97 as

$$f'(x', y') = f(x, y) = f(\underline{R}^{-1} \underline{r}') \quad (3.100)$$

where for the depicted 2-D rotation

$$\underline{R}^{-1} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

Example 3.54

Suppose the value associated with the point $\underline{r} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ is 10, that is $f(1, 3) = 10$ what is $f'(x' = 3, y' = -1)$ for $\theta = \pi/2$?

SOLUTION

Using Equation 3.100, we find

$$f'(3, -1) = f[\underline{R}^{-1} \underline{r}'] = f \left[\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right] = f \left[\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right] = f(1, 3) = 10$$

3.15.2 ROTATION OPERATOR

The unitary operator

$$\hat{R} = e^{i\vec{\alpha} \cdot \vec{L}/\hbar} \quad (3.101)$$

maps a function into another that corresponds to rotated position coordinates. Here, $\hat{L} = L_x\hat{x} + L_y\hat{y} + L_z\hat{z}$ is the generator of rotations (later called the angular momentum operator) and $\vec{\alpha} = \alpha_x\hat{x} + \alpha_y\hat{y} + \alpha_z\hat{z}$ gives a rotation angle. The constant $\hbar$, which is related to Planck's constant by $\hbar = h/(2\pi)$, has been included so that $\hat{L}$ also represent the angular momentum in the quantum theory; for nonquantum applications, one can make the replacement $\hbar \rightarrow 1$ for simplicity and convenience. For example, α_z is the rotation angle around the $\hat{z}$ -axis and L_z is the generator for the group of rotations about the z -axis as well as the z -component of angular moment. In the 3-D case, $|\vec{\alpha}|$ is the rotation angle about the unit axis $\vec{\alpha}/|\vec{\alpha}|$. In many cases, it suffices to consider rotations around the z -axis by a judicious choice of coordinate systems.

Consider the simple case of a rotation about the $\hat{z}$ -axis.

$$\hat{R} = e^{i\theta_0 \hat{L}_z/\hbar} \quad (3.102)$$

where the generator L_z has the form $\hat{L}_z = -i\hbar \partial/\partial\theta$. The nonzero commutator $[\theta, \hat{L}_z] = i\hbar$ indicates that the rotation operator uses products of conjugate variables similar to the translation operator. Consider a function $\psi(r, \theta) \equiv \psi(\theta)$ and calculate a *new* function corresponding to the old one evaluated at $\theta \rightarrow \theta + \varepsilon$. The Taylor expansion gives

$$\psi'(\theta) = \psi(\theta + \varepsilon) = \psi(\theta) + \frac{\varepsilon}{1!} \frac{\partial}{\partial\theta} \psi(\theta) + \frac{\varepsilon^2}{2!} \frac{\partial^2}{\partial\theta^2} \psi(\theta) + \dots = \left[\hat{1} + \frac{\varepsilon}{1!} \frac{\partial}{\partial\theta} + \frac{\varepsilon^2}{2!} \frac{\partial^2}{\partial\theta^2} + \dots \right] \psi(\theta) = e^{\varepsilon \partial_\theta} \psi(\theta)$$

where $\partial_\theta = \partial/\partial\theta$. We can rearrange the exponential in terms of the z -component of the angular momentum $L_z = \frac{\hbar}{i} \frac{\partial}{\partial\theta}$ to find $\hat{R}(\varepsilon) = e^{\varepsilon \partial_\theta} = e^{i\varepsilon L_z/\hbar}$ where $\hbar$ symbolizes a constant for the quantum theory (Planck's constant $h = 2\pi\hbar$). Repeatedly applying the operator produces the rotation

$$\hat{R}(\theta_0) = e^{i\theta_0 L_z/\hbar} \quad (\text{Coordinate rotation}) \quad (3.103a)$$

$$\psi'(\theta) = \psi(\theta + \theta_0) = \hat{R}(\theta_0) \psi(\theta) \quad (\text{Coordinate rotation}) \quad (3.103b)$$

Figure 3.24 shows that the rotation moves the function in the direction of a negative angle or rotates the coordinates in the positive direction.

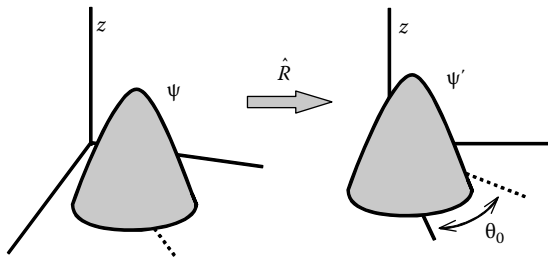


FIGURE 3.24 Rotating the function through an angle.

If we replace $\theta_o \rightarrow -\theta_o$ then the rotation would be in the opposite sense. An appropriated definition for the rotation operator and the rotated functions becomes

$$\hat{R}(\theta_o) = e^{-i\theta_o L_z/\hbar} \quad (\text{Function rotation}) \quad (3.104a)$$

$$\psi'(\theta) = \psi(\theta - \theta_o) = \hat{R}(\theta_o) \psi(\theta) \quad (\text{Function rotation}) \quad (3.104b)$$

Equations 3.104 represent the active point of view.

3.15.3 RECTANGULAR COORDINATES FOR THE GENERATOR OF ROTATIONS ABOUT Z

We can easily show the rectangular-coordinate form for the generator of rotation around the z -axis given by $L_z = \frac{\hbar}{i} \frac{\partial}{\partial \theta}$. The rectangular and polar forms are related by

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta$$

Therefore,

$$\frac{\partial}{\partial \theta} \psi(x, y) = \frac{\partial \psi}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial \psi}{\partial y} \frac{\partial y}{\partial \theta} = -r \sin \theta \frac{\partial \psi}{\partial x} + r \cos \theta \frac{\partial \psi}{\partial y} = \left(x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} \right) \psi = \frac{i}{\hbar} L_z \psi$$

from which one concludes the rectangular form

$$L_z = xp_y - yp_x = \frac{\hbar}{i} (x\hat{p}_y - y\hat{p}_x)$$

One can likewise deduce the full set of generators by cyclic permutation of the subscripts

$$xp_y - yp_x = \frac{\hbar}{i} (x\hat{p}_y - y\hat{p}_x) = L_z \quad (3.105a)$$

$$yp_z - zp_y = \frac{\hbar}{i} (y\hat{p}_z - z\hat{p}_y) = L_x \quad (3.105b)$$

$$zp_x - xp_z = \frac{\hbar}{i} (z\hat{p}_x - x\hat{p}_z) = L_y \quad (3.105c)$$

Owing to the fact that L represents the angular momentum, as will become more apparent in Chapter 5, these relations can also be written as

$$\vec{L} = \vec{r} \times \vec{p} = \begin{vmatrix} \tilde{x} & \tilde{y} & \tilde{z} \\ x & y & z \\ p_x & p_y & p_z \end{vmatrix} \quad (3.106)$$

The antisymmetric tensor ε_{ijk} can be used to provide a more convenient and compact notation

$$L_i = \sum_{jk} \varepsilon_{ijk} x_j p_k \quad (3.107)$$

3.15.4 ROTATION OF THE POSITION OPERATOR

The position operator can be written in rotated form. Denote the position operator by $\hat{r} = \hat{x}\tilde{x} + \hat{y}\tilde{y} + \hat{z}\tilde{z}$ where $\tilde{x}, \tilde{y}, \tilde{z}$ represent the usual Euclidean unit vectors. The position operator

provides the relation $\hat{r}|\vec{r}_o\rangle = \vec{r}_o|\vec{r}_o\rangle$. Now consider a rotation of a function. The relation between the new and old functions gives $\langle\vec{r}|\psi'\rangle = \langle\vec{r}|\hat{R}|\psi\rangle \equiv \langle\vec{r}'|\psi\rangle$. We therefore conclude that $|\vec{r}'\rangle = \hat{R}^+|\vec{r}\rangle$. For example, the coordinate ket might represent the wave function for a particle localized at the particular point $\vec{r}$. We see that the operator rotates the location in the positive angle direction. We can also see that the position operator must satisfy the relation

$$\hat{r}|\vec{r}'\rangle = \vec{r}'|\vec{r}'\rangle \rightarrow \hat{r}\hat{R}^+|\vec{r}\rangle = \vec{r}'\hat{R}^+|\vec{r}\rangle \rightarrow \hat{R}\hat{r}\hat{R}^+|\vec{r}\rangle = \vec{r}'|\vec{r}\rangle \rightarrow \hat{r}' = \hat{R}\hat{r}\hat{R}^+$$

which gives the rotated form of the position operator. We can also show

$$\langle\vec{r}|\psi'\rangle = \langle\vec{r}'|\psi\rangle \rightarrow \psi'(\vec{r}) = \psi(\vec{r}') \rightarrow \hat{R}\psi(\vec{r}) = \psi(\vec{r}') = \psi(\vec{r}^{-1}\vec{r})$$

where R is the corresponding operator for Euclidean vectors. This shows that for every operation in coordinate space, there must correspond an operation in Hilbert space. The angle represents the coordinate space while the angular momentum represents the Hilbert space operation.

3.15.5 STRUCTURE CONSTANTS AND LIE GROUPS

A ‘‘Lie’’ group has elements that depend on a continuous parameter. The translation group provides one example. The set of rotations described by $\hat{R} = e^{i\vec{\alpha}\cdot\vec{L}/\hbar}$ provides another as L_x, L_y, L_z symbolize generators for the group, and the $\alpha_x, \alpha_y, \alpha_z$ provide the continuous parameters. In general, the group elements in a Lie group have the form (for a compact group) of a unitary operator (when α_n real and $\hat{G}_n$ Hermitian)

$$\text{Exp}\{i\alpha_n\hat{G}_n\} \quad (\text{Einstein sum convention}) \quad (3.108)$$

As a note, repeated adjacent indices have an implied summation so Equation 3.108 should be read as $\text{Exp}\{i\alpha_n\hat{G}_n\} = \text{Exp}\{i\sum_n \alpha_n\hat{G}_n\}$. The remainder of the equations for the present section will use the Einstein summation convention. The set $\{\hat{G}_n\}$ consists of the generators that also form a basis for the vector space of operators. The collection $\{\alpha_n\hat{G}_n\}$ for all possible values of α_n give all of the elements of the vector space. One should note that the Lie group is necessarily different from the vectors space.

Given that the product of two group elements such as $e^{i\alpha_n\hat{G}_n}$ and $e^{i\beta_n\hat{G}_n}$ must produce a third element in the group such as $e^{i\delta_n\hat{G}_n}$

$$e^{i\delta_n\hat{G}_n} = e^{i\alpha_n\hat{G}_n}e^{i\beta_n\hat{G}_n} \quad (3.109)$$

For example, using the operator expansion theorem

$$\hat{O} = e^{x\hat{A}}\hat{B}e^{-x\hat{A}} = \hat{B} + x[\hat{A}, \hat{B}] + \frac{x^2}{2!}[\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

on $(e^{i\lambda\hat{G}_b}e^{+i\lambda\hat{G}_a}e^{-i\lambda\hat{G}_b})e^{-i\lambda\hat{G}_a}$ to find

$$(e^{i\lambda\hat{G}_b}e^{+i\lambda\hat{G}_a}e^{-i\lambda\hat{G}_b})e^{-i\lambda\hat{G}_a} = 1 + i\lambda[\hat{G}_b, e^{+i\lambda\hat{G}_a}] + \dots \sim 1 + \frac{\lambda^2}{2}[\hat{G}_a, \hat{G}_b] + \dots$$

For the product to yield another group element, one requires

$$[\hat{G}_a, \hat{G}_b] = if_{abc}\hat{G}_c \quad (3.110)$$

where the structure constants f_{abc} determine the multiplication laws of the group elements (refer to the H. Georgi book in the references for more information).

Returning to Equation 3.109, one can find the relation between the δ_c , α_a , β_b by expanding all exponentials to linear order and then keeping only linear terms at the end.

$$1 + i\delta_c \hat{G}_c - \frac{1}{2}(\delta_c \hat{G}_c)^2 = 1 + i(\alpha_a \hat{G}_a + \beta_b \hat{G}_b) - \frac{1}{2}(\alpha_a \hat{G}_a + \beta_b \hat{G}_b)^2 - \frac{1}{2}[\alpha_a \hat{G}_a, \beta_b \hat{G}_b] \quad (3.111)$$

The commutator terms were added to complete the square for the third term on the right (since the generators do not necessarily commute for calculating the square). Drop the squared terms in this last equation and use the following relation incorporating the structure constants

$$-\frac{1}{2}[\alpha_a \hat{G}_a, \beta_b \hat{G}_b] = -\frac{1}{2}\alpha_a \beta_b [\hat{G}_a, \hat{G}_b] = -\frac{1}{2}\alpha_a \beta_b f_{abc} \hat{G}_c$$

to find

$$1 + i\delta_c \hat{G}_c = 1 + i(\alpha_a \hat{G}_a + \beta_b \hat{G}_b) - \frac{1}{2}\alpha_a \beta_b f_{abc} \hat{G}_c \quad (3.112)$$

Notice that repeated indices imply a summation over all basis vectors $\hat{G}_n$. Therefore, compare each side for a particular n to find

$$\delta_n = \alpha_n + \beta_n - \frac{1}{2}\alpha_a \beta_b f_{abn} \quad (3.113)$$

where the Einstein summation convention applies to repeated indices in the product.

3.15.6 STRUCTURE CONSTANTS FOR THE ROTATION LIE GROUP

Commutation relations form the corner stone of quantum theory in order to determine complete sets of observables and the possible states for particles (systems) that will be found upon observation. Equation 3.105 shows the three generators for the group which also represent the angular momentum operators. One can easily demonstrate the following commutation relations

$$[\hat{L}_x, \hat{L}_y] = i\hbar \hat{L}_z \quad [\hat{L}_y, \hat{L}_z] = i\hbar \hat{L}_x \quad [\hat{L}_z, \hat{L}_x] = i\hbar \hat{L}_y \quad (3.114)$$

as will be shown in Chapter 5 using the fundamental relations between position and momentum $[x, p_x] = i\hbar$, $[y, p_y] = i\hbar$, and $[z, p_z] = i\hbar$. The relations in Equation 3.114 can be summarized by

$$[\hat{L}_i, \hat{L}_j] = i\hbar \varepsilon_{ijk} \hat{L}_k \quad (3.115)$$

where the completely antisymmetric tensor has the form

$$\varepsilon_{ijk\dots} = \begin{cases} +1 & \text{even permutations of } 1, 2, 3, \dots \\ -1 & \text{odd permutation of } 1, 2, 3, \dots \\ 0 & \text{if any of } i = j = k \text{ holds} \end{cases} \quad (3.116)$$

For example $\varepsilon_{132} = -1$, $\varepsilon_{312} = +1$, and $\varepsilon_{133} = 0$. Comparing Equations 3.115 and 3.110 shows the structure constants have the form $f_{ijk} = \hbar \varepsilon_{ijk}$.

3.16 DYADIC NOTATION

This section develops the dyadic notation for the second rank tensor. Studies in solid state sometimes use dyadic quantities to describe the effective mass of an electron or hole. Dyads also find usage in studies of electromagnetism for nonisotropic quantities such as might be the case for the dependence of molecular polarization on electric field. The dyad is equivalent to 2-D matrices (second rank tensor) but makes use of a convenient vector notation.

3.16.1 NOTATION

For semiconductors with nonisotropic effective mass, for example, the general formulas relating the acceleration $\vec{a}$ of a particle to the applied force $\vec{F}$ have the form

$$\vec{F} = \overleftrightarrow{m} \cdot \vec{a} \quad (3.117)$$

where the dyadic quantity $\overleftrightarrow{m}$ represents the effective mass. This equation represents the case when the applied force produces an acceleration in a direction other than parallel to the force (see the discussion in Chapter 7).

3.16.2 EQUIVALENCE BETWEEN THE DYAD AND THE MATRIX

A dyad can be written in terms of components, for example,

$$\overleftrightarrow{A} = \sum_{ij} A_{ij} \tilde{e}_i \tilde{e}_j \quad (3.118)$$

where the unit vector $\tilde{e}_i$ can be one of the basis vectors $\{\tilde{x}, \tilde{y}, \tilde{z}\}$ for a 3-D space, and the $\tilde{e}_i \tilde{e}_j$ symbol places the unit vectors next to each other without an operator separating them.

Example 3.55

Find $\overleftrightarrow{A} \cdot \vec{v}$ for $\overleftrightarrow{A} = 1\tilde{e}_1\tilde{e}_1 + 2\tilde{e}_3\tilde{e}_2 + 3\tilde{e}_2\tilde{e}_3$ and $\vec{v} = 4\tilde{e}_1 + 5\tilde{e}_2 + 6\tilde{e}_3$

SOLUTION

$$\overleftrightarrow{A} \cdot \vec{v} = (1\tilde{e}_1\tilde{e}_1 + 2\tilde{e}_3\tilde{e}_2 + 3\tilde{e}_2\tilde{e}_3) \cdot (4\tilde{e}_1 + 5\tilde{e}_2 + 6\tilde{e}_3) = 4\tilde{e}_1 + 10\tilde{e}_3 + 18\tilde{e}_2 = 4\tilde{x} + 18\tilde{y} + 10\tilde{z}$$

The coefficients in Equation 3.118 can be arranged in a matrix. This means that a 3×3 matrix provides an alternate representation of the second rank tensor and the dyad. The matrix elements can easily be seen to be

$$\tilde{e}_a \cdot \overleftrightarrow{A} \cdot \tilde{e}_b = \sum_{ij} A_{ij} \tilde{e}_a \cdot \tilde{e}_i \tilde{e}_j \cdot \tilde{e}_b = \sum_{ij} A_{ij} \delta_{ai} \delta_{jb} = A_{ab} \quad (3.119)$$

The procedure should remind you of Dirac notation for the matrix.

The unit dyad can be written as

$$\overleftrightarrow{1} = \sum_i \tilde{e}_i \tilde{e}_i \quad (3.120)$$

Applying the definition of the matrix elements in Equation 3.119 shows the unit dyad produces the unit matrix.

Example 3.56

Show that if $\overleftrightarrow{1} = \overleftrightarrow{A}$ then $A_{ab} = \delta_{ab}$

SOLUTION

Operate with $\tilde{e}_a$ on the left and $\tilde{e}_b$ on the right to find

$$\tilde{e}_a \cdot \overleftrightarrow{1} \cdot \tilde{e}_b = \tilde{e}_a \cdot \overleftrightarrow{A} \cdot \tilde{e}_b = \tilde{e}_a \cdot \left(\sum_{ij} A_{ij} \tilde{e}_i \tilde{e}_j \right) \cdot \tilde{e}_b = \sum_{ij} A_{ij} \delta_{ai} \delta_{jb} = A_{ab}$$

Using Equation 3.120, one can see $\tilde{e}_a \cdot \overleftrightarrow{1} \cdot \tilde{e}_b = \delta_{ab}$. So we have $A_{ab} = \delta_{ab}$

Now let us discuss the inverse of a dyad. Suppose

$$\overleftrightarrow{1} = \overleftrightarrow{A} \cdot \overleftrightarrow{B} \quad (3.121)$$

then we can show that $\overleftrightarrow{B} = \overleftrightarrow{A}^{-1}$ where $\overleftrightarrow{A} = \sum_{ii'} A_{ii'} \tilde{e}_i \tilde{e}_{i'}$ and $\overleftrightarrow{B} = \sum_{jj'} B_{jj'} \tilde{e}_j \tilde{e}_{j'}$. Operating on the left of Equation 3.121 with $\tilde{e}_a$ and on the right by $\tilde{e}_b$ produces

$$\delta_{ab} = \tilde{e}_a \cdot \left(\sum_{ii'} A_{ii'} \tilde{e}_i \tilde{e}_{i'} \cdot \sum_{jj'} B_{jj'} \tilde{e}_j \tilde{e}_{j'} \right) \cdot \tilde{e}_b = \sum_{ii'jj'} A_{ii'} B_{jj'} \tilde{e}_a \cdot \tilde{e}_i \tilde{e}_{i'} \cdot \tilde{e}_j \tilde{e}_{j'} \cdot \tilde{e}_b$$

The dot products produce Kronecker delta functions.

$$\delta_{ab} = \sum_{ii'jj'} A_{ii'} B_{jj'} \delta_{ai} \delta_{i'j} \delta_{j'b} = \sum_j A_{aj} B_{jb}$$

which shows the matrices $\underline{A}$ and $\underline{B}$ must be inverses.

3.17 REVIEW EXERCISES

- 3.1 Show that the “null space” of a linear operator $\hat{T}$ defined by $N = \{|\nu\rangle: \hat{T}|\nu\rangle = 0\}$ forms a vector space. The proof can be simplified by noting the set N is contained in a vector space V .
- 3.2 Show that the inverse of a linear operator $\hat{T}$ does not exist when the null space $N = \{|\nu\rangle: \hat{T}|\nu\rangle = 0\}$ has more than one element.
- 3.3 Let $\hat{T}: V \rightarrow W$ be an “onto” linear operator. Let $V = \text{Sp}\{|\phi_i\rangle: i = 1, \dots, n_v\}$ and $W = \text{Sp}\{|\psi_i\rangle: i = 1, \dots, n_w\}$. Show that

$$\text{Dim}(V) = \text{Dim}(W) + \text{Dim}(N)$$

where $N = \text{null space } N = \{ |v\rangle : \hat{T}|v\rangle = 0 \}$. Hint: Let $|1\rangle, \dots, |n\rangle$ be the basis for N . Let $|1\rangle, \dots, |n\rangle, |n+1\rangle, \dots, |p\rangle$ be the basis for V . Use the definition of linearly independent. Note that $0 = \hat{T} \sum_{i=n+1}^p c_i |i\rangle$ requires $\sum_{i=n+1}^p c_i |i\rangle$ be in the null space. The null space has only $\vec{0}$ in common with $\text{Sp}\{|n+1\rangle, \dots, |p\rangle\}$.

- 3.4** For vector spaces V and W and linear operator $\hat{T}: V \rightarrow W = \text{Range}(\hat{T})$, show that every vector $|w\rangle$ must have multiple preimages in V when the null space $N = \{ |v\rangle : \hat{T}|v\rangle = 0 \}$ has multiple elements. Conclude that the inverse of $\hat{T}$ does not exist.
Hint: Suppose $|w\rangle \in W$, $|w\rangle \neq \vec{0}$ and $\hat{T}|v\rangle = |w\rangle$. Examine $N + \{ |v\rangle \}$ where N represents the null space.
- 3.5** Suppose an operator $\hat{T}: V \rightarrow W' = \text{Range}(\hat{T})$ where W' is contained in the vector space W . Prove or disprove that W' is a vector space.
- 3.6** If x is an element of a group G , prove that the inverse element x^{-1} is unique.
- 3.7** Find the multiplication table for a group with exactly three elements. Note that $gg = g$ would require $g = e = \text{identity}$ and there would only be one element in the group.
- 3.8** Consider a 2-D vector space $V = \text{Sp}\{ \phi_1, \phi_2 \}$ and the operator $\hat{T}: V \rightarrow V$ defined by

$$\hat{T}|\phi_1\rangle = \frac{1}{\sqrt{2}}|\phi_1\rangle + \frac{1}{\sqrt{2}}|\phi_2\rangle \quad \hat{T}|\phi_2\rangle = -\frac{1}{\sqrt{2}}|\phi_1\rangle + \frac{1}{\sqrt{2}}|\phi_2\rangle$$

Show that the operator does not change the length of an arbitrary vector

$$|v\rangle = a|\phi_1\rangle + b|\phi_2\rangle$$

- 3.9** Under what conditions is the function $y = mx + b$ linear according to the definitions in Section 3.1?
- 3.10** Prove or disprove which of the following operators are linear for a vector space of differentiable functions $\{f(x, y, z)\}$.
- $\hat{T} = d/dx$ (derivative).
 - $\hat{T} = x \frac{d}{dx}$.
 - $\hat{T} = \nabla$ (gradient).
 - $\hat{T}f = \left(\frac{df}{dx}\right)^2$.
 - The dot product between real vectors. What do you suppose bilinear means?
- 3.11** Write a linear operator that doubles the angle between a vector and the horizontal axis.
- 3.12** Prove the Levi-Civita formula for the determinant of a 2×2 matrix. Repeat the procedure for a 3×3 matrix.
- 3.13** Show in general that $\text{Det}(c\hat{A}) = c^n \text{Det}(\hat{A})$ where $\hat{A}: V \rightarrow V$, $n = \text{Dim}(V)$, and c is a complex number. Hint: It is easiest to work with the antisymmetric tensor for this purpose.
- 3.14** Show that the row expansion method to evaluate a 3×3 determinant

$$\begin{vmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{vmatrix} = T_{11} \begin{vmatrix} T_{22} & T_{23} \\ T_{32} & T_{33} \end{vmatrix} - T_{12} \begin{vmatrix} T_{21} & T_{23} \\ T_{31} & T_{33} \end{vmatrix} + T_{13} \begin{vmatrix} T_{21} & T_{22} \\ T_{31} & T_{32} \end{vmatrix}$$

follows from the basic definition of the determinant using the Levi-Civita formula.

- 3.15** Using the Levi-Civita formula for evaluating a determinant, show that expanding a determinant along row i and column j produces terms with $(-1)^{i+j}$ as factors.
- 3.16** Find the inverse of the following matrix using row operations

$$\underline{M} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

3.17 Show the following relations

$$\text{Det}(\hat{A}\hat{B}) = \text{Det}(\hat{A})\text{Det}(\hat{B}) \quad \text{and} \quad \text{Det}(\hat{A}\hat{B}\hat{C}) = \text{Det}(\hat{A})\text{Det}(\hat{B})\text{Det}(\hat{C})$$

You can use the first relation to prove the second one.

3.18 Show that for a 2×2 matrix, the inverse does not exist when the determinant is zero.

3.19 Show that a 2×2 determinant is zero when one row is a constant multiplied by the other row.

3.20 Show the $\det(T)$ is independent of the particular basis chosen for the vector space. Hint: Use the unitary operator and a similarity transformation to change T , then use the results of previous problems.

3.21 Assume $\hat{A}, \hat{B}$ operate on a single vector space $V = \text{Sp}\{|1\rangle, |2\rangle, \dots\}$. Show $\text{Tr}(\hat{A}\hat{B}) = \text{Tr}(\hat{B}\hat{A})$ by inserting the closure relation.

3.22 Show the relation $\text{Tr}(\hat{A}\hat{B}\hat{C}) = \text{Tr}(\hat{B}\hat{C}\hat{A}) = \text{Tr}(\hat{C}\hat{A}\hat{B})$ assuming $\hat{A}, \hat{B}, \hat{C}$ all operate on $V = \text{Sp}\{|1\rangle, |2\rangle, \dots\}$ for simplicity.

3.23 Show that the trace of the operator $\hat{T}$ is “independent” of the chosen basis set. Hint: Use a unitary operator to change basis and also use the closure relation.

3.24 Show that the set of operators forms a group with respect to operator addition.

3.25 Show that the relation between operators and matrices $\hat{T} = \sum_{ab} T_{ab} |\phi_a\rangle\langle\psi_b|$ where $V_1 = \text{Sp}\{|\phi_a\rangle\}$, $V_2 = \text{Sp}\{|\psi_a\rangle\}$ forms an isomorphism.

3.26 Consider the group formed from rotations in the x - y plane $\{R_\theta \text{ for } \theta = 0^\circ, 120^\circ, 240^\circ\}$ where R_θ refers to a rotation through an angle θ . The following table shows the multiplicative results. Find the matrix representation for the operators in the following table.

Mult	R_0	R_{120}	R_{240}
R_0	R_0	R_{120}	R_{240}
R_{120}	R_{120}	R_{240}	R_0
R_{240}	R_{240}	R_0	R_{120}

3.27 Show that the isomorphism between operators and matrices inherent in

$$\hat{T} = \sum_{ab} T_{ab} |\phi_a\rangle\langle\psi_b|$$

dictates the form of matrix addition.

3.28 Show that if $\hat{A}, \hat{B}$ are linear operators then so is $\hat{A}\hat{B}$

3.29 Show

$$\hat{O} = e^{x\hat{A}} \hat{B} e^{-x\hat{A}} = \hat{B} + x[\hat{A}, \hat{B}] + \frac{x^2}{2!} [\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

by making a Taylor expansion of both $e^{x\hat{A}}$ and $e^{-x\hat{A}}$.

3.30 Let $\{|\phi_1\rangle, |\phi_2\rangle\}$ be a basis set. Write the following operator in matrix notation

$$\hat{L}_1 = |\phi_1\rangle\langle\phi_1| + 2|\phi_1\rangle\langle\phi_2| + 3|\phi_2\rangle\langle\phi_2|$$

3.31 Let $\{|\phi_1\rangle, |\phi_2\rangle\}$ be a basis set. Write the following operator in matrix notation

$$\hat{L}_2 = |\phi_1\rangle\langle\phi_1| + (1+2j)|\phi_1\rangle\langle\phi_2| + (1-2j)|\phi_2\rangle\langle\phi_1| + 3|\phi_2\rangle\langle\phi_2|$$

3.32 Prove that the mapping of the basis vector by an operator uniquely determines the operator.

- 3.33** Using the operators in Problems 3.30 and 3.31 determine if $\hat{O}_1 = \hat{L}_1 \hat{L}_2$ is the same as $\hat{O}_2 = \hat{L}_2 \hat{L}_1$. Write the matrices for $\hat{O}_1$ and $\hat{O}_2$.
- 3.34** A Hilbert space V has basis $\{|\phi_1\rangle, |\phi_2\rangle\}$. Assume that the linear operator $\hat{L}: V \rightarrow V$ has the matrix $\underline{L} = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}$. Write the operator in the form $\hat{L} = \sum_{ij} L_{ij} |\phi_i\rangle \langle \phi_j|$.
- 3.35** Write an operator $\hat{L}: V \rightarrow V$ in the form $\hat{L} = \sum L_{ab} |\phi_a\rangle \langle \phi_b|$ when $\hat{L}$ maps the basis set $\{|\phi_1\rangle, |\phi_2\rangle\}$ into the basis set $\{|\psi_1\rangle, |\psi_2\rangle\}$ according to the rule $\hat{L}|\phi_1\rangle = |\psi_1\rangle$ and $\hat{L}|\phi_2\rangle = |\psi_2\rangle$. Assume that the two sets of basis vectors are related as follows:

$$|\psi_1\rangle = \frac{1}{\sqrt{3}}|\phi_1\rangle + \sqrt{\frac{2}{3}}|\phi_2\rangle \quad \text{and} \quad |\psi_2\rangle = -\sqrt{\frac{2}{3}}|\phi_1\rangle + \frac{1}{\sqrt{3}}|\phi_2\rangle$$

- 3.36** Let $\{|\phi_1\rangle, |\phi_2\rangle\}$ be a basis set. Write the following operator in matrix notation

$$\hat{L} = |\phi_1\rangle \langle \phi_1| + 2|\phi_1\rangle \langle \phi_2| + 3|\phi_2\rangle \langle \phi_2|$$

- 3.37** Let $\{|\phi_1\rangle, |\phi_2\rangle\}$ be a basis set. Write the following operator in matrix notation

$$\hat{L} = |\phi_1\rangle \langle \phi_1| + (1 + 2j)|\phi_1\rangle \langle \phi_2| + (1 - 2j)|\phi_2\rangle \langle \phi_1| + 3|\phi_2\rangle \langle \phi_2|$$

- 3.38** Suppose $\hat{H} = \sum_n E_n |n\rangle \langle n|$ where $E_n \neq 0$ for all n . What value of c_n in $\hat{O} = \sum_n C_n |n\rangle \langle n|$ makes $\hat{O}$ the inverse of $\hat{H}$ so that $\hat{H}\hat{O} = 1 = \hat{O}\hat{H}$.
- 3.39** If $\hat{H} = 1|\phi_1\rangle \langle \phi_1| + 2|\phi_2\rangle \langle \phi_2|$ and $|\psi(0)\rangle = 0.86|\phi_1\rangle + 0.51|\phi_2\rangle$ is the wave function for an electron at $t = 0$. Find the average energy $\langle \psi(0) | \hat{H} | \psi(0) \rangle$.
- 3.40** Prove that the required property $\langle \hat{A} | \alpha \hat{B} + \beta \hat{C} \rangle = \alpha \langle \hat{A} | \hat{B} \rangle + \beta \langle \hat{A} | \hat{C} \rangle$ for $\langle \hat{A} | \hat{B} \rangle = \text{Tr} \hat{A}^\dagger \hat{B}$ to be an inner product. Use $L = \{\hat{T}: V \rightarrow V\}$.
- 3.41** Prove $\langle \hat{A} | \hat{A} \rangle = 0$ if and only if $\hat{A} = 0$ for $\hat{A} \in L = \{\hat{T}: V \rightarrow V\}$, the set of linear operators. Hint: Consider the expansion of an operator in a basis set.
- 3.42** Show that the set of linear operators $\{\hat{T}: V \rightarrow W\}$ mapping the vector space V into the vector space W forms a vector space.
- 3.43** Show that the proposed inner product $\langle \hat{A} | \hat{B} \rangle = \text{Trace}(\hat{A}^\dagger \hat{B})$ defined on the set of operators $L = \{\hat{T}: V \rightarrow W\}$ satisfies the three properties for inner product given in Section 2.1.
- 3.44** Determine if the quantity $\langle \hat{L}_1 | \hat{L}_2 \rangle = \frac{\text{Tr}\{\hat{L}_1^\dagger \hat{L}_2\}}{\text{Dim}(V)}$ satisfies the requirements for an inner product where $L_1, L_2: V \rightarrow V$. Prove it one way or the other.
- 3.45** Suppose $V = \text{Sp}\{|1\rangle, |2\rangle, \dots, |n\rangle\}$ and $\hat{L}: V \rightarrow V$ according to

$$\hat{L}|1\rangle = |\phi_1\rangle \quad \text{and} \quad \hat{L}|2\rangle = |\phi_2\rangle, \text{ etc.}$$

where $|\phi_1\rangle, |\phi_2\rangle$ are not necessarily orthogonal. Use the inner product $\langle \hat{L}_1 | \hat{L}_2 \rangle = \frac{\text{Tr}\{\hat{L}_1^\dagger \hat{L}_2\}}{\text{Dim}(V)}$ to show $\hat{L}$ has unit length so long as $|\phi_1\rangle, |\phi_2\rangle$ have unit length. Hint: First write $\hat{L} = |\hat{\phi}_1\rangle \langle 1| + \dots$, then calculate $\hat{L}^\dagger \hat{L}$ having terms such as $|1\rangle \langle 1| \phi_1 | \phi_1 + \dots$, and then calculate the trace.

- 3.46** (A) Find the “length” of a unitary operator $\hat{u}: V \rightarrow V$ where $\text{Dim}(V) = N$. That is, calculate $\|\hat{u}\|^2 = \langle \hat{u} | \hat{u} \rangle = \text{Tr}(\hat{u}^\dagger \hat{u}) / \text{Dim}(V)$. It is probably easiest to use matrices after taking the trace. (B) Find the length of an operator that doubles the length of every vector in an $N = 2$ vector space. (C) Find the length of the operator defined by $\hat{O}|v\rangle = c|v\rangle$.

3.47 Show that the “length” of an operator $\hat{O}$ according to the trace formula

$$\langle \hat{L}_1 | \hat{L}_2 \rangle = \text{Tr}(\hat{L}_1^\dagger \hat{L}_2)$$

is equivalent to finding the length of the vector $\hat{O}|v\rangle$ where $|v\rangle = \sum_n v_n |n\rangle$ and $|v_n|^2 = 1$. What is the length of $|v\rangle$? What is the length of $\hat{O}|v\rangle$? How do these two lengths compare with the length of the operator?

3.48 For a finite dimension space V , show that if one uses the definition for inner product as

$$\langle \hat{L}_1 | \hat{L}_2 \rangle = \frac{\text{Tr}\{\hat{L}_1^\dagger \hat{L}_2\}}{\text{Dim}(V)}$$

then the “length” of an operator $\hat{O}$ according to this trace formula is equivalent to the length of the vector $\hat{O}|v\rangle$ where $|v\rangle = \sum_n v_n |n\rangle$ must have unit length and $|v_m|^2 = |v_n|^2$ for all components designated by m, n .

3.49 Consider a direct product space $V \otimes V$ where $V = \text{Sp}\{|1\rangle, |2\rangle\}$ with only two basis vectors. If

$$\hat{O} = 1|11\rangle\langle 11| + 2|11\rangle\langle 12| + 3|11\rangle\langle 21| + 4|12\rangle\langle 11| + 5|21\rangle\langle 11| + 6|21\rangle\langle 21| + 7|22\rangle\langle 22|$$

Then find the conventional matrix to describe this linear operator.

3.50 Find the trace of the following operator

$$\hat{O} = 1|11\rangle\langle 11| + 2|11\rangle\langle 12| + 3|11\rangle\langle 21| + 4|12\rangle\langle 11| + 5|21\rangle\langle 11| + 6|21\rangle\langle 21| + 7|22\rangle\langle 22|$$

3.51 Consider two operators $\hat{O}$ on the vector space V . If the $\hat{U}$ diagonalizes the operator $\hat{O}$, can you find an operator that diagonalizes $\hat{O} \otimes \hat{O}$? Prove it one way or the other.

3.52 If $\hat{O} = 1|11\rangle\langle 11| + 2|11\rangle\langle 12| + 3|11\rangle\langle 21| + 4|12\rangle\langle 11| + 5|21\rangle\langle 11| + 6|21\rangle\langle 21| + 7|22\rangle\langle 22|$ then find $\hat{O}\{|11\rangle + 2|12\rangle\}$ using both operator and matrix notation.

3.53 Prove properties 1–7 for the commutator given in Section 3.6.

3.54 If $\hat{A}^2 = \hat{A}$ and $[\hat{A}, \hat{B}] = 1$ then show $[e^{i\hat{A}x}, \hat{B}] = e^{ix} - 1$

3.55 Find $\sin \underline{A}$ where $\underline{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Hint: Use the Taylor expansion.

3.56 Find $\sin \underline{A}$ where $\underline{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. Hint: Find a matrix $\underline{u}$ such that $\underline{u}\underline{A}\underline{u}^\dagger = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \underline{A}_D$ where λ_i represents the eigenvalues. Taylor expand $\sin \underline{A}$. Calculate $\hat{u}[\sin \underline{A}]\hat{u}^\dagger$.

3.57 Consider a 3-D coordinate system. Write the matrix that rotates 45° about the x -axis.

3.58 Suppose an operator rotates vectors by $\theta = 30^\circ$. Write the operator in the form $\sum_{a,b} c_{ab}|a\rangle\langle b|$ and write the matrix.

3.59 Consider a rotated basis set $|n'\rangle = \hat{u}|n\rangle$. Show that the closure relation in the primed system leads to the closure relation in the unprimed system.

$$1 = \sum |n'\rangle\langle n'| \rightarrow 1 = \sum |n\rangle\langle n|$$

3.60 Find a condition on c that makes the following matrix Hermitian

$$\begin{bmatrix} 1 & c \\ c & -1 \end{bmatrix}$$

3.61 Find a condition on a that makes the following operator Hermitian

$$\hat{L} = |1\rangle\langle 1| + aj|1\rangle\langle 2| + aj^*|2\rangle\langle 1| + |2\rangle\langle 2|$$

where $j = \sqrt{-1}$.

3.62 Show that the trace of a Hermitian operator $\hat{H}$ must be the sum of the eigenvalues λ_i given by $\text{Trace } \hat{H} = \sum \lambda_i$. Hint: Let $\{|n\rangle\}$ be the basis for the space V where $\hat{H}: V \rightarrow V$. Let $\hat{u}$ be the unitary operator that diagonalizes the operator.

$$\text{Tr } \hat{H} = \sum_n \langle \varphi_n | \hat{H} | \varphi_n \rangle = \sum_n \langle \varphi_n | \hat{u}^\dagger \hat{H} \hat{u} | \varphi_n \rangle = \sum_n \langle \phi_n | \hat{H}_D | \phi_n \rangle = \text{Tr } \hat{H}_D$$

The eigenvalues must be on the diagonal of $\hat{H}_D$

$$\hat{H}_D | \phi_n \rangle = \lambda_n | \phi_n \rangle \rightarrow (\underline{H}_D)_{ab} = \langle \phi_a | \hat{H}_D | \phi_b \rangle = \lambda_b \delta_{ab}$$

3.63 Show that the determinant of the operator in the previous problem must be the product of eigenvalues.

3.64 Assume that $\underline{H}$ is a 3×3 matrix and the columnar form of the three eigenvectors have the form

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

Show by matrix multiplication the following:

$$\underline{H} \begin{pmatrix} e \\ v \\ i \end{pmatrix} = \lambda_i \begin{pmatrix} e \\ v \\ i \end{pmatrix} \quad \text{and} \quad \underline{u}^\dagger \underline{H} \underline{u} = \begin{bmatrix} (ev-1)^* \\ (ev-2)^* \\ \vdots \end{bmatrix} \left[\lambda_1 \begin{pmatrix} e \\ v \\ 1 \end{pmatrix} \quad \lambda_2 \begin{pmatrix} e \\ v \\ 2 \end{pmatrix} \quad \cdots \right]$$

where $\underline{u}^\dagger$ has three columns consisting of the three eigenvectors.

3.65 Find the eigenvectors for Hermitian matrix

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

and then show how to make it diagonal.

3.66 Find the eigenvectors for the non-Hermitian matrix and then diagonalize it.

$$\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

3.67 Use the definition of adjoint $\langle f | \hat{L} g \rangle = \langle \hat{L}^\dagger f | g \rangle$ for $\hat{L} = a(d/dx)$ to show that $\hat{L}^\dagger = -\hat{L}$ requires a to be purely real. Assume that the Hilbert space consists of functions $f(x)$ such that $f(\pm\infty) = 0$.

3.68 Use the definition of adjoint $\langle f | \hat{L} g \rangle = \langle \hat{L}^\dagger f | g \rangle$ for $\hat{L} = a(d/dx)$ to show that $\hat{L}^\dagger = \hat{L}$ requires a to be purely imaginary. Assume that the Hilbert space consists of functions $f(x)$ such that $f(\pm\infty) = 0$.

- 3.69** If $\hat{L} = \partial^2/\partial x^2$ then find $\hat{L}^+$ by partial integration. Assume a Hilbert space of differentiable functions $\{\psi(x)\}$ such that $\psi(x \rightarrow \pm\infty) = 0$.
- 3.70** Show $(\hat{A}\hat{B})^+ = \hat{B}^+\hat{A}^+$ using $\langle f|\hat{T}g\rangle = \langle \hat{T}^+f|g\rangle$.
- 3.71** Without multiplying the matrices, find the adjoint of the following matrix equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} g \\ h \end{bmatrix}$$

- 3.72** Suppose $\hat{O} = \hat{O}^{(V)}\hat{O}^{(W)}$ where $V = \text{Sp}\{|\phi_a\rangle\}$ and $W = \text{Sp}\{|\psi_a\rangle\}$. Show

$$\underline{O}_{ab,cd} = \langle \phi_a | \hat{O}^{(V)} | \phi_c \rangle \langle \psi_b | \hat{O}^{(W)} | \psi_d \rangle$$

- 3.73** For the basis vector expansion of $|\Psi\rangle = \sum_{ab} \beta_{ab} |\phi_a \psi_b\rangle$ in the tensor product space $V \otimes W$ with $V = \text{Sp}\{|\phi_i\rangle\}$ and $W = \text{Sp}\{|\psi_j\rangle\}$, show that the expansion coefficients must be $\beta_{ab} = \langle \phi_a \psi_b | \Psi \rangle$ and the closure relation has the form $\sum_{ab} |\phi_a \psi_b\rangle \langle \phi_a \psi_b| = \hat{1}$.
- 3.74** For a vector space V spanned by $\{|1\rangle, |2\rangle\}$ with $\hat{u}$ an orthogonal rotation by 45° and $\hat{T} = |1\rangle\langle 1| + 2|2\rangle\langle 2|$, find $\hat{T}$ in the new basis set. Hint: Find $\hat{u}$ by visual inspection and write in terms of the original basis.
- 3.75** Show $[\theta, \hat{L}_z] = i\hbar$.
- 3.76** Prove the operator expansion theorem

$$\hat{O} = e^{x\hat{A}}\hat{B}e^{-x\hat{A}} = \hat{B} + x[\hat{A}, \hat{B}] + \frac{x^2}{2!}[\hat{A}, [\hat{A}, \hat{B}]] + \dots$$

by expanding the exponentials and collecting terms.

- 3.77** Show $\text{Tr}(\hat{O}) = \sum_{i,j} T_{ij}$ when $\hat{O} = \sum_{ij} |j\rangle\langle i| \hat{T}$
- 3.78** Show that a linear operator $\hat{A}$ can always be written as $\hat{A} = \hat{A}_1 + i\hat{A}_2$ where $(\hat{A}_1)_{mn}$ and $(\hat{A}_2)_{mn}$ are both real for all basis vectors. Hint: Consider the basis vector expansion of the operator.
- 3.79** Show that if $\hat{A}$ is a Hermitian linear operator, then all of its elements must be real. That is, each element $\underline{A}_{mn}$ is real. Hint: Consider the previous problem.
- 3.80** Show that if $\hat{A}$ is an anti-Hermitian linear operator $\hat{A}^+ = -\hat{A}$, then all of its elements must be pure imaginary. That is, each element $\underline{A}_{mn}$ is pure imaginary. Hint: Consider the previous problem.
- 3.81** Show that the set of translations $\{\hat{T}(\eta) = e^{-i\eta\hat{p}}\}$ forms a group with η as a continuous parameter.

REFERENCES AND FURTHER READING

Classics and standard

1. Dirac P.A.M., *The Principles of Quantum Mechanics*, 4th ed., Oxford University Press, Oxford (1978).
2. Von Neumann J., *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, NJ (1996).
3. Byron F.W. and Fuller R.W., *Mathematics of Classical and Quantum Physics*, Dover Publications, New York (1970).
4. Von Neuman J., *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, NJ (1996).
5. Schwinger J., *Quantum Kinematics and Dynamics*, W.A. Benjamin Inc., New York (1970).
6. Lee T.D., *Particle Physics and Introduction to Field Theory*, Harwood Academic Publishers, New York (1981).
7. Cushing J.T., *Applied Analytical Mathematics for Physical Scientists*, John Wiley & Sons, Inc., New York (1975).

Introductory

8. Krause E.F., *Introduction to Linear Algebra*, Holt, Rinehart & Winston, New York (1970).
9. Bronson R., *Matrix Methods, An Introduction*, Academic Press, New York (1970).

Involved

10. Loomis L.H. and Sternberg S., *Advanced Calculus*, Addison-Wesley Publishing Co., Reading, MA (1968).
11. Stakgold I., *Green's Functions and Boundary Value Problems*, 2nd ed., John Wiley & Sons, New York (1998).

A variety

12. Dennerly P. and Krzywicki A., *Mathematics for Physicists*, Dover Publications, Mineola, NY (1995).
13. Schechter M., *Operator Methods in Quantum Mechanics*, Dover Publications, Mineola, NJ (2002).
- Lawden D.F., *The Mathematical Principles of Quantum Mechanics*, Dover Publications, Mineola, NJ (1995).
14. Akhiezer N.I. and Glazman I.M., *Theory of Linear Operators in Hilbert Space*, Dover Publications, Mineola, NJ (1993).

Group theory

15. Rose J.S., *A Course on Group Theory*, Dover Publications, Mineola, NJ (1994).
16. Barnard T. and Neill H., *Mathematical Groups*, Teach Yourself Books of Hodder & Stoughton, Reading, MA (1996).
17. Weyl H., *The Theory of Groups and Quantum Mechanics*, Dover Publications, Mineola, NJ (1950).

Lie algebras

18. Georgi, H., *Lie Algebras in Particle Physics*, Benjamin/Cummings Publishing Company, Reading, MA (1982).

Miscellaneous

19. Dahlquist G. and Bjorck A., *Numerical Methods*, Dover Publications, Mineola, NJ (2003).

4 Fundamentals of Classical Mechanics

The classical mechanics, founded on Newton's laws, forms a cornerstone for a variety of fields especially physics and engineering. Elementary studies focus on the force and vector quantities. A conceptually simpler approach formulates the dynamics in terms of energy. The principle of least action provides Lagrange's and Hamilton's equations that substitute for Newton's laws and describe the motion of a particle or system. The quantum theory modifies a number of physical assumptions and again poses the dynamics in terms of a Hamiltonian. In fact, the quantum mechanical Hamiltonian comes from the classical one by substituting operators for the classical dynamical variables.

The present chapter summarizes the concepts for the Lagrangian and Hamiltonian form of classical mechanics starting with generalized coordinates and constraints. The Lagrange and Hamilton formulations make extensive use of generalized coordinates—especially for the study of fields. The chapter shows how minimizing the action leads to the Lagrangian and then to the Hamiltonian. Similar to the study of the commutator in quantum mechanics, the classical mechanics gives rise to the Poisson brackets. The quantum commutator and Poisson brackets perform similar functions in each of their respective theories. The chapter shows how the Lagrangian of discrete coordinates can be generalized to the continuous case and then how it produces the Schrödinger wave equation. Section 4.8 discusses Einstein's special relativity as an introduction to the modern point of view of space-time. The remainder of the book will be primarily interested in the notation used in the special relativity.

4.1 CONSTRAINTS AND GENERALIZED COORDINATES

The Lagrangian and Hamiltonian formulation of classical mechanics provide simple techniques for deriving equations of motion using energy relations. Rather than being concerned with complicated vector relations, these alternate formulations allow one to use the scalar quantities of kinetic and potential energy. In classical mechanics, the Hamiltonian consists of the sum of kinetic T and potential energy V and can be represented by $H = T + V$. On the other hand, the Lagrangian has the form $L = T - V$. However, the two functionals H and L are related to each other by a Legendre transformation. These functionals provide a gateway to quantizing systems of multiple particles and fields (such as electromagnetic fields).

The Lagrangian (L) and Hamiltonian (H) are functionals of generalized coordinates. Generalized coordinates consist of any set of independent variables that describe the object (or objects) under scrutiny. For example, the rectangular coordinates x, y, z or the cylindrical coordinates r, θ, φ provide generalized coordinates for an unconstrained point particle (i.e., for a particle free to move in three dimensions). These generalized coordinates depend on time when they describe a point on a moving object.

4.1.1 CONSTRAINTS

Constraints represent a priori knowledge of a physical system. They reduce the total number of degrees of freedom available to the system. For example, Figure 4.1 shows a collection of masses

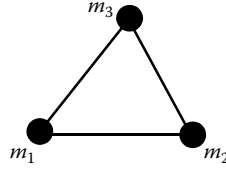


FIGURE 4.1 Three masses connected by rigid rods.

interconnected by rigid (massless) rods. These rods constrain the distance between the masses and therefore reduce the number of degrees of freedom; however the whole system (of three masses) can translate or rotate. As another example, the walls of a container also impose constraints on a system. In this case, the constraints are important only when the molecules in the container make contact with the walls. For quantum theory, constraints are quite nonphysical since small particles experience forces and not constraints. For example, electrostatic forces (and not rigid rods) hold atoms in a lattice. Sometimes constraints appear in the quantum description to simplify problems. Evidently, constraints are mostly important for macroscopic classical systems.

4.1.2 GENERALIZED COORDINATES

Suppose a generalized set of coordinates $S_q = \{q_1, q_2, \dots, q_k\}$ describes the position of N point particles. A single point particle has exactly 3 degrees of freedom corresponding to the three translational directions. Without constraints, N particles have $k = 3N$ degrees of freedom. Position vectors normally describe the location of the N particles

$$\begin{aligned}\vec{r}_1 &= \vec{r}_1(q_1, \dots, q_k, t) \\ &\vdots \\ \vec{r}_N &= \vec{r}_N(q_1, \dots, q_k, t)\end{aligned}\tag{4.1}$$

For example, the $\{q_i\}$ might be spherical coordinates. The q_i are independent of each other in this case. Constraints reduce the degrees of freedom so that $k < 3N$; that is, the constraints eliminate $3N - k$ degrees of freedom. As a note, we make use of the generalized coordinates especially for fields, which do not use the notion of constraints.

Example 4.1: A Pulley System Connecting Two Point Particles

Assume a massless pulley (Figure 4.2). Normally two point masses would have 6 degrees of freedom. Confining the masses to a two-dimensional (2-D) plane reduces the degrees of freedom to 4. Allowing only vertical motion for the two masses reduces the degrees of freedom to 2. The string requires the masses to move together and reduces the number of degrees of freedom to 1. The motion of both masses can be described by either the generalized coordinate $q_1 = \theta$ or the position of m_1 above a horizontal reference plane. The single generalized coordinate describes the position vectors $\vec{r}_1, \vec{r}_2$ for the masses.

Configuration space consists of the collection of the k generalized coordinates $\{q_1, q_2, \dots, q_k\}$ where each coordinate can take on a range of values. These generalized coordinates are especially important for the Lagrange formulation of dynamics. We can define generalized velocities by

$$\{\dot{q}_1, \dot{q}_2, \dots, \dot{q}_k\}\tag{4.2}$$

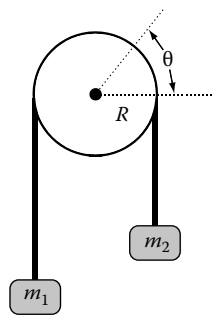


FIGURE 4.2 Two masses connected by a string passing over a pulley.

However, they are not independent of the generalized coordinates for the Lagrange formulation. That is, the variations δq , $\delta \dot{q}$ depend on each other.

The generalized coordinates discussed so far constitute a discrete set whereby the coordinates are in one-to-one correspondence with a finite subset of the integers. It is possible for the set to be infinite. A continuous set of coordinates would have elements in 1–1 correspondence with a “continuous” subset of the “real” numbers. The distinction is important for a number of topics especially field theory.

Let us discuss a picture for the generalized coordinates and velocities especially important for field theories. We already know how to picture the position of particles in space for the case of x -, y -, z -coordinates. So instead, let us take an example that illustrates the distinction between indices and generalized coordinates. Let us start with a collection of atoms arranged along a one-dimensional (1-D) line oriented along the x -direction. Assume the number of atoms is k . As illustrated in the top portion of Figure 4.3, the atoms have equilibrium positions represented by the coordinates x_i . Given one atom for each equilibrium position x_i , the atoms can be labeled by either the respective equilibrium position x_i or by the number i . The bottom portion of the figure shows the situation for the atoms displaced along the vertical direction. In this case, the generalized coordinates label the displacement from equilibrium. For the 1-D case shown, the generalized coordinates can be written equally well as either q_i or $q(x_i)$ so that $q_i = q(x_i) = q_{x_i}$. In this case, we think of x_i or i as indices to label a particular point in space or atom. More generally for 3-D motion, each atom would have three generalized coordinates and three generalized velocities.

Mathematically, the displacements could be randomly assigned. It is only when the dynamics (Newton’s laws, etc.) are applied to the problem that the displacements become correlated. That is, the flow of energy from one atom to the next influences the motions in a predictable manner. Mathematically, without dynamics (i.e., Newton’s laws), atom #1 can be moved to position q_1 and

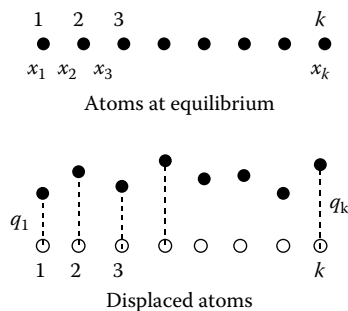


FIGURE 4.3 Example of generalized coordinates for atoms in a lattice.

atom #2 to position q_2 without there being any reason for choosing those two positions. The position of either atom can be independently assigned. This notion of independent translations leads to an alternate formulation of Newton's laws.

Let us return briefly to Figure 4.3 and discuss its importance to field theories. Let us focus on electromagnetics. When we write the electric field, for example, as $\vec{E}(x, t)$ we think of x as an index labeling a particular point along a line in space. We might $\vec{E}$ as a displacement of "something" at point x . The displacement can vary with time. There must be three generalized coordinates at the point x . The three generalized coordinates are the three vector components of $\vec{E}$. So $\vec{E}$ really represents three displacements at the point x and not just one. Also notice that the indices x form a continuum rather than the discrete set indicated in Figure 4.3.

4.1.3 PHASE SPACE COORDINATES

A system, which can consist of a single or multiple particles, evolves in time when it follows a curve in phase space. Phase space consists of the generalized coordinates and conjugate momentum

$$\{q_1, q_2, \dots, q_k, p_1, p_2, \dots, p_k\} \quad (4.3)$$

all of which are assumed to be independent of one another. The momentum p_i is conjugate to the coordinate q_i because it describes the momentum of the particle corresponding to the direction q_i . Assigning particular values to the $2k$ -coordinates in phase space specifies the "state of the system." The phase space coordinates are used primarily with the Hamiltonian of the system.

The Hamilton formulation of dynamics uses phase space coordinates.

$$\{q_1, q_2, \dots, q_k, p_1, p_2, \dots, p_k\} \quad (4.4)$$

Each member of the set of the phase space coordinates has the same level of importance as any other member so that one cannot be more fundamental than another. For example, a point particle can be independently given position coordinates x, y, z and momentum coordinates $\{p_x, p_y, p_z\}$. This means that the particle can be assigned a random position and a random velocity. Given that the phase space coordinates are all independent, we can also vary the coordinates in an independent manner; that is, the variations $\delta q, \delta p$ must be independent of one another. The term "configuration space" applies to the coordinates $\{q_1, q_2, \dots, q_k\}$ and the term "phase space" applies to the full set of coordinates $\{q_1, q_2, \dots, q_k, p_1, p_2, \dots, p_k\}$. Essentially, in the absence of dynamics, position and momentum can be arbitrarily assigned to each particle.

Example 4.2

The momentum p_x describes the momentum of a particle along the x -direction.

Example 4.3

Consider the pulley system shown in Figure 4.2. The momentum conjugate to the generalized coordinate θ is the total angular momentum along the axis of the pulley.

4.2 ACTION, LAGRANGIAN, AND LAGRANGE'S EQUATION

The notion that nature follows a "law of least action" has a long history starting around 200 BC. The law of reflection in optics as well as the law of refraction can be derived from the notion that light travels from one point to another in space by following a path that requires the shortest amount

of time to traverse. In the 1700s, the law was reformulated to mean that the dynamics of mechanical systems minimize the action, which is defined as Et where E is the energy and t is the time. In the 1800s, Hamilton stated the most general form: a dynamical system will follow a path that minimizes the action defined as the time integral of the Lagrangian $L = T - V$, which is the difference between the kinetic energy (T) and potential energy (V). The Hamiltonian (the energy of a system) is related to the Lagrangian. Today, the Lagrangian and Hamilton play central roles in quantum theory. The Schrödinger equation can be found from the classical Hamiltonian by replacing the classical dynamical variables with operators. The Feynman path integral provides a beautiful formulation of the quantum principle by incorporating the integral over all possible paths for the action.

This section first shows the origin of the Lagrangian in Newton's laws and then develops Hamilton's principle. The two developments although producing the same results have some significant philosophical differences. Newton's laws involve forces that act on an object. These forces are external to the object. Hamilton's principle however, discusses the dynamics in terms of quantities possessed by the object (kinetic and potential energy).

4.2.1 ORIGIN OF THE LAGRANGIAN IN NEWTON'S EQUATIONS

Forces acting on the constituent particles of a system control the dynamics of the system. However, we often do not know the forces of constraint until after solving the problem. The present section discusses the "derivation" of the Lagrangian based directly on Newton's laws. As such, it is quite rigorous and does not require the added assumption of the law of least action. The reader can read Section 4.2.2 for the more intuitive approach and then return to the present section. D'Alembert (and Bernoulli) divided the forces $F = F^{(a)} + F^{(c)}$ into "applied forces" $F^{(a)}$ and "constraint forces" $F^{(c)}$ and assumed that the virtual work $\delta W^{(c)} = \vec{F}^{(c)} \cdot \delta \vec{r}_i = 0$ done by forces of constraint must be zero since the forces act perpendicular to the direction of motion. D'Alembert's principle means that Newton's second law $F = ma$ takes the form

$$\sum_i \left(\vec{F}_i^{(a)} - \frac{d\vec{P}_i}{dt} \right) \cdot \delta \vec{r}_i = 0 \quad (4.5)$$

where

$\delta \vec{r}_i$ is the virtual displacements of the i th particle

$\vec{P}_i = m\vec{\dot{r}}_i$ is the momentum

The virtual displacements must be consistent with the constraints. Virtual displacements do not involve time so that the spatial force distribution does not change. The virtual displacements $\delta \vec{r}_i$ cannot be independent of each other without incorporating the equations of constraint; they must be rewritten in terms of q_i and δq_i . We can use

$$\delta \vec{r}_i = \sum_j \frac{\partial \vec{r}_i}{\partial q_j} \delta q_j$$

where time is not included. Often generalized forces Q_i are defined for Equation 4.5

$$\delta W = \sum_i \vec{F}_i^{(a)} \cdot \delta \vec{r}_i = \sum_{i,j} \vec{F}_i \cdot \frac{\partial \vec{r}_i}{\partial q_j} \delta q_j = 0 \quad \rightarrow \quad Q_j = \sum_i \vec{F}_i \cdot \frac{\partial \vec{r}_i}{\partial q_j} \quad (4.6)$$

So that the work has a form similar to the usual vector definition as

$$\delta W = \sum_i \vec{F}_i^{(a)} \cdot \delta \vec{r}_i = \sum_j Q_j \delta q_j$$

Continuing, Lagrange's equation can be found by manipulating the momentum term in Equation 4.5

$$\sum_i \frac{d\vec{P}_i}{dt} \cdot \delta\vec{r}_i = \sum_i m_i \ddot{\vec{r}}_i \cdot \delta\vec{r}_i = \sum_{i,j} m_i \ddot{\vec{r}}_i \cdot \frac{\partial \vec{r}_i}{\partial q_j} \delta q_j = \sum_{i,j} \left[\frac{\partial}{\partial t} \left(m_i \dot{\vec{r}}_i \cdot \frac{\partial \vec{r}_i}{\partial q_j} \right) - m_i \dot{\vec{r}}_i \cdot \frac{\partial \dot{\vec{r}}_i}{\partial q_j} \right] \delta q_j$$

where

$$\vec{v}_i = \frac{d\vec{r}_i}{dt} = \frac{d}{dt} \vec{r}_i(q_1, \dots, q_k, t) = \sum_j \frac{\partial \vec{r}_i}{\partial q_j} \frac{dq_j}{dt} + \frac{\partial \vec{r}_i}{\partial t} = \sum_j \frac{\partial \vec{r}_i}{\partial q_j} \dot{q}_j + \frac{\partial \vec{r}_i}{\partial t} \quad \text{so that} \quad \frac{\partial \vec{v}_i}{\partial \dot{q}_j} = \frac{\partial \vec{r}_i}{\partial q_j}$$

Therefore we find that the momentum term becomes

$$\begin{aligned} \sum_i \frac{d\vec{P}_i}{dt} \cdot \delta\vec{r}_i &= \sum_{i,j} \left[\frac{\partial}{\partial t} \left(m_i \vec{v}_i \cdot \frac{\partial \vec{v}_i}{\partial \dot{q}_j} \right) - m_i \vec{v}_i \cdot \frac{\partial \vec{v}_i}{\partial q_j} \right] \delta q_j \\ &= \sum_j \left[\frac{\partial}{\partial t} \frac{\partial}{\partial \dot{q}_j} \left(\sum_i \frac{1}{2} m_i \vec{v}_i^2 \right) - \frac{\partial}{\partial q_j} \left(\sum_i \frac{1}{2} m_i \vec{v}_i^2 \right) \right] \delta q_j \end{aligned} \quad (4.7)$$

The quantity

$$T = \sum_i \frac{1}{2} m_i \vec{v}_i^2$$

is the kinetic energy and it is a function of only $\dot{q}_j$ for $j = 1, \dots, k$.

The generalized forces can be included by combining Equations 4.5 and 4.6

$$\sum_i \frac{d\vec{P}_i}{dt} \cdot \delta\vec{r}_i = \sum_i \vec{F}_i^{(a)} \cdot \delta\vec{r}_i = \sum_j Q_j \delta q_j$$

or, substituting for the summation over momentum we have

$$\sum_j \left[\frac{\partial}{\partial t} \frac{\partial T}{\partial \dot{q}_j} - \frac{\partial T}{\partial q_j} \right] \delta q_j = \sum_j Q_j \delta q_j$$

Using the fact that the δq_j are independent,

$$\frac{\partial}{\partial t} \frac{\partial T}{\partial \dot{q}_j} - \frac{\partial T}{\partial q_j} = Q_j$$

which is a form of Lagrange's equation. Divide the forces acting on each particle into conservative and nonconservative forces. The forces can be written in terms of a potential $V(q_1, q_2, \dots, q_k)$ and the nonconservative forces $Q_j^{(nc)}$ as

$$Q_j = -\frac{\partial V}{\partial q_j} + Q_j^{(nc)} \quad (4.8)$$

The potential V does not depend on $\dot{q}_j$ and the kinetic energy T does not depend on q_j so that the Lagrangian $L = T - V$ satisfies Lagrange's equation

$$\frac{\partial}{\partial t} \frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = Q_j^{(nc)} \quad (4.9)$$

We have shown in this section that Lagrange's equation and the Lagrangian are a result of the instantaneous state of the system, instantaneous forces, and the virtual displacements. As is well known, the Lagrangian and Lagrange's equation are most commonly obtained by the calculus of variations along with an integral principle as shown in Section 4.2.2.

4.2.2 LAGRANGE'S EQUATION FROM A VARIATIONAL PRINCIPLE

The forces acting on a system of particles control the dynamics of the system. However, forces of constraint are often not known until after the problem is solved. D'Alembert (and Bernoulli) divided the forces into applied and constraint forces, and assumed that the virtual work done by forces of constraint is zero (since the forces are assumed to act perpendicular to the direction of motion). The resulting derivation results in the Lagrange formulation of mechanics. Lagrange's equations provide an alternative formulation of Newton's laws. This section discusses the typical variational method of obtaining the Lagrangian and Lagrange's equations. The method is particularly easy to generalize for systems consisting of continuous sets of coordinates (i.e., field theory).

Hamilton's principle produces Lagrange's equation for conservative systems. Of all the possible paths in configuration space that a system could follow between two fixed points $1 = (q_1^{(1)}, q_2^{(1)}, \dots, q_k^{(1)})$ and $2 = (q_1^{(2)}, q_2^{(2)}, \dots, q_k^{(2)})$, the path that it actually follows makes the following action integral an extremum (preferably a minimum) (Figure 4.4).

$$I = \int_1^2 dt L(q_1, q_2, \dots, q_k, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_k, t) \quad (4.10)$$

The Lagrangian L is a functional of the kinetic energy T and potential energy V according to $L = T - V$ for particles. The procedure assumes fixed endpoints but this can be generalized to variable endpoints. To minimize the notation, let $q_i, \dot{q}_i$ represent the entire collection of points in $\{q_1, q_2, \dots, q_k, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_k\}$.

To find the extremum of the action integral (with fixed end points)

$$I = \int_1^2 dt L(q_i, \dot{q}_i, t)$$

define a new path in configuration space for each generalized coordinate q_i by

$$q_i'(t) = q_i(t) + \delta q_i$$

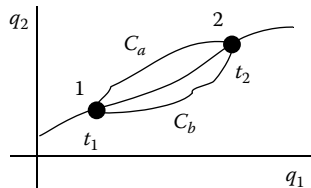


FIGURE 4.4 Three paths connecting fixed end points.

where the time t parameterizes the curve in configuration space. Assume q_i extremizes the integral I . We can find the functional form of each $q_i(t)$ by requiring the variation of the integral around q_i to vanish as follows.

$$0 = \delta I = \int_1^2 dt \sum_i \left[\frac{\partial L(q_i, \dot{q}_i, t)}{\partial q_i} \delta q_i + \frac{\partial L(q_i, \dot{q}_i, t)}{\partial \dot{q}_i} \delta \dot{q}_i \right] \quad (4.11)$$

Partially integrate the second term using the fact that $\delta q_i(t_1) = 0 = \delta q_i(t_2)$ to find

$$0 = \delta I = \int_1^2 dt \sum_i \left[\frac{\partial L(q_i, \dot{q}_i, t)}{\partial q_i} - \frac{d}{dt} \frac{\partial L(q_i, \dot{q}_i, t)}{\partial \dot{q}_i} \right] \delta q_i$$

The small variations δq_i are assumed to be independent so that

$$\frac{\partial L}{\partial q_i} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = 0 \quad \text{for } i = 1, 2, \dots \quad (4.12)$$

where $L = T - V$.

The canonical momentum can be defined as

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \quad (4.13)$$

where p_i denotes the momentum conjugate to the coordinate q_i . The canonical momentum does not always agree with the typical momentum mv for a particle. The canonical momentum for an EM field interacting with a particle consists of the particle and field momentum.

Example 4.4

Consider a single particle of mass m constrained to move vertically along the y -direction and acted upon by the gravitational force $F = -mg$ (see Figure 4.2)

$$T = \frac{1}{2} m(\dot{y})^2 \quad V = mgy \quad L = T - V = \frac{1}{2} m(\dot{y})^2 - mgy$$

Lagrange's equation

$$\frac{\partial L}{\partial y} - \frac{d}{dt} \frac{\partial L}{\partial \dot{y}} = 0$$

gives Newton's second law for a gravitational force $-mg - m\ddot{y} = 0$ with the derivatives

$$\frac{\partial \dot{y}}{\partial y} = 0 = \frac{\partial y}{\partial \dot{y}}$$

since y and $\dot{y}$ are taken to be independent arguments of the Lagrangian. As a result, the equation of motion for the particle becomes $\ddot{y} = -g$ which gives the usual functional form of the height as $y = -\frac{g}{2}t^2 + v_0 t + y_0$.

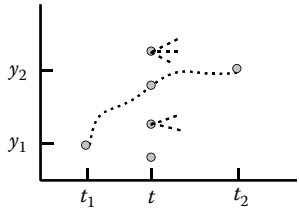


FIGURE 4.5 The function is determined by its value and slope at each point.

How can $y, \dot{y}$ be independent when they appear to be connected by $\dot{y} = dy/dt$? This relation assumes that the function y is already defined. Let us start with the step of defining the function y . At any value t , we can arbitrarily assign a value y and a value $\dot{y}$. The only requirement is that the function y must have fixed endpoints y_1 and y_2 . These boundary conditions restrict only two points out of an uncountable infinite number. Figure 4.5 illustrates the concept. Notice that the value t can be assigned a large number of values of y and $\dot{y}$ without affecting the endpoints. Therefore, there can be many curves connecting points $A = (t_1, y_1)$ and $B = (t_2, y_2)$. The equation $\dot{y} = dy/dt$ gives a procedure for calculating the slope $\dot{y}$ only after we know the function y in some interval. For example, suppose we discuss the motion of a line of atoms so that the independent variables are $\{y, \dot{y}\}$ where $\dot{y}$ is the velocity. We can arbitrarily assign a displacement and a speed at each point x . It is only after we solve Newton’s equations that we know how the speed and position at those points are interrelated.

Example 4.5

Find the equations of motion for the pulley system shown in Figure 4.6. Assume the pulley is massless, and $m_2 > m_1$ and that $y_1(t) = 0, y_2(t) = h$. The kinetic energy is $T = \frac{1}{2}m_1\dot{y}_1^2 + \frac{1}{2}m_2\dot{y}_2^2$ and $V = m_1gy_1 + m_2gy_2$. The remaining 2 degrees of freedom y_1, y_2 can be reduced to one since $y_2 = h - y_1$. We therefore have

$$T = \frac{1}{2} (m_1 + m_2)\dot{y}_1^2 \quad V = m_1gy_1 + m_2g(h - y_1)$$

Lagrange’s equation

$$\frac{\partial L}{\partial y_1} - \frac{d}{dt} \frac{\partial L}{\partial \dot{y}_1} = 0 \quad \text{produces} \quad \ddot{y}_1 = \frac{(m_1 - m_2)g}{(m_1 + m_2)}$$

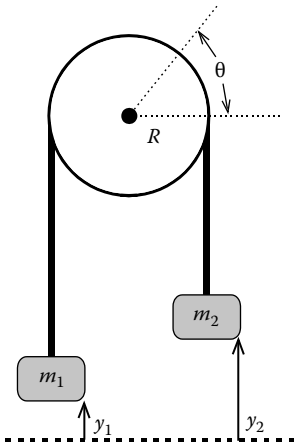


FIGURE 4.6 Pulley system.

4.3 HAMILTONIAN

The Hamiltonian represents the total energy of a system. The quantum theory derives its Hamiltonian from the classical one by substituting operators for the classical dynamical variables.

4.3.1 HAMILTONIAN FROM THE LAGRANGIAN

Consider a closed, conservative system so that the Lagrangian L does not explicitly depend on time. The total energy and the total number of particles remain constant (in time) for a closed system. We define a conservative system to be one for which all of the forces can be derived from a potential. We do not consider any equations of constraint for quantum mechanics and field theory. Differentiating the Lagrangian provides

$$\frac{dL}{dt} = \sum_i \left(\frac{\partial L}{\partial q_i} \frac{dq_i}{dt} + \frac{\partial L}{\partial \dot{q}_i} \frac{d\dot{q}_i}{dt} \right) + \frac{\partial L}{\partial t} \quad (4.14)$$

The last term is zero by assumption

$$\frac{\partial L}{\partial t} = 0$$

Substitute Lagrange's equation

$$\frac{\partial L}{\partial q_i} = \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i}$$

to find

$$\frac{dL}{dt} = \sum_i \left[\left(\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} \right) \dot{q}_i + \frac{\partial L}{\partial \dot{q}_i} \frac{d\dot{q}_i}{dt} \right] = \sum_i \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \dot{q}_i \right) \quad (4.15)$$

Using the definition for the conjugate momentum given by

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \quad (4.16)$$

Equation 4.15 becomes

$$\frac{d}{dt} \left[\sum_i \dot{q}_i p_i - L \right] = 0$$

The Hamiltonian H is defined to be

$$H = \sum_i \dot{q}_i p_i - L \quad (4.17)$$

which is the total energy of the system in this case. *Important point:* We consider H to be a function of q_i, p_i whereas we consider L to be a function of $q_i, \dot{q}_i$.

4.3.2 HAMILTON'S CANONICAL EQUATIONS

The Hamiltonian leads to Hamilton's canonical equations

$$\dot{q}_j = \frac{\partial H}{\partial p_j} \quad \dot{p}_j = -\frac{\partial H}{\partial q_j} \quad (4.18)$$

These equations allow us to find equations of motion from the Hamiltonian. We will see for the quantum theory that the operator form of the q_j and p_j must satisfy commutation relations. The classical equivalent of the commutation relations appears in Section 4.4 on the Poisson brackets.

Hamilton's canonical equations (Equation 4.18) can now be demonstrated. Starting with Equation 4.17 we can write

$$\frac{\partial H}{\partial p_j} = \frac{\partial}{\partial p_j} \left[\sum_i \dot{q}_i p_i - L \right] = \dot{q}_j - \frac{\partial L}{\partial p_j} \quad (4.19)$$

Next noting that L depends on q_i , $\dot{q}_i$ and not p_i , we find

$$\frac{\partial H}{\partial p_j} = \dot{q}_j \quad (4.20)$$

which proves the first of Hamilton's equations. We can demonstrate the second of Hamilton's equations by using Lagrange's equation and the canonical momentum

$$\frac{\partial L}{\partial q_j} = \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_j} \quad p_j = \frac{\partial L}{\partial \dot{q}_j} \quad (4.21)$$

from the previous section. We find

$$\frac{\partial H}{\partial q_j} = \frac{\partial}{\partial q_j} \left[\sum_i \dot{q}_i p_i - L \right] = 0 - \frac{\partial L}{\partial q_j} = -\frac{d}{dt} \frac{\partial L}{\partial \dot{q}_j} = -\frac{d}{dt} p_j = -\dot{p}_j$$

Example 4.6

Find H and $\dot{q}_i$, $\dot{p}_i$ for a particle of mass m at a height y in a gravitational field.

SOLUTION

The Lagrangian has the form

$$L = T - V = \frac{1}{2} m (\dot{y})^2 - mgy$$

The Hamiltonian H can be written as a function of the coordinate and its conjugate momentum. The relation for the canonical momentum for the Lagrangian

$$p = \frac{\partial L}{\partial \dot{y}} = m\dot{y}$$

allows H to be written as

$$H = \dot{y}p - L = \frac{p}{m}p - \left[\frac{1}{2}m\left(\frac{p}{m}\right)^2 - mgy \right] = \frac{p^2}{2m} + mgy$$

and then

$$\dot{y} = \frac{\partial H}{\partial p} = \frac{p}{m} \quad \dot{p} = -\frac{\partial H}{\partial y} = mg$$

The Hamiltonian H can be seen to be the sum of the kinetic and potential energy $T + V$ by calculating

$$H = \sum_i \dot{q}_i p_i - L$$

with $L = T - V$ and using a general quadratic form for the kinetic energy

$$T = \sum_{i,j} a_{ij} \dot{q}_i \dot{q}_j \quad \text{where } a_{ij} = a_{ji}$$

The canonical momentum is

$$p_m = \frac{\partial L}{\partial \dot{q}_m} = 2 \sum_i a_{im} \dot{q}_i$$

Therefore,

$$\begin{aligned} H &= \sum_m \dot{q}_m p_m - L = \sum_m \dot{q}_m 2 \sum_i a_{im} \dot{q}_i - (T - V) = 2 \sum_{mi} a_{im} \dot{q}_i \dot{q}_m - T + V \\ &= 2T - T + V \\ &= T + V \end{aligned}$$

Example 4.7

For the pulley system in Figure 4.7, find the Hamiltonian and Newton's equations of motion. Assume the pulley is massless and h represents the maximum height difference between m_1 and m_2 .

SOLUTION

The potential energy is

$$T = \frac{1}{2} (m_1 + m_2) \dot{y}_1^2 \quad V = m_1 g y_1 + m_2 g (h - y_1)$$

The Hamiltonian must be a function of momentum and not velocity. The Lagrangian $L = T - V$ gives the canonical momentum

$$p_1 = \frac{\partial L}{\partial \dot{y}_1} = \frac{\partial}{\partial \dot{y}_1} \frac{1}{2} (m_1 + m_2) \dot{y}_1^2 = M \dot{y}_1$$

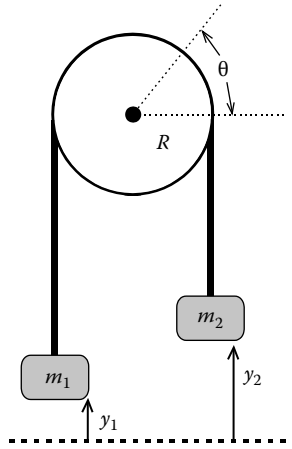


FIGURE 4.7 The pulley system.

where $M = m_1 + m_2$. The kinetic energy can be rewritten as $T = \frac{1}{2}(m_1 + m_2)\dot{y}_1^2 = p_1^2/2M$.

The Hamiltonian can be written as

$$H = \dot{q}_1 p_1 - L = \frac{p_1}{M} p_1 - (T - V) = \frac{p_1^2}{M} - \frac{p_1^2}{2M} + m_1 g y_1 + m_2 g (h - y_1) = \frac{p_1^2}{2M} + g y_1 (m_1 - m_2) + m_2 g h$$

Newton's equation of motion provides the rate of change of motion. The Hamiltonian gives the time rate of change of momentum as

$$\dot{p}_1 = -\frac{\partial H}{\partial q_1} = -g(m_1 - m_2)$$

which can be rewritten as a second-order differential equation if desired. Notice how the momentum $p_1 = M\dot{y}_1$ represents a type of total momentum but not the usual vector sum that might be written as $p_{\text{vect}} = (-m_2 + m_1)\dot{y}_1$.

4.4 POISSON BRACKETS

The Poisson brackets provide an alternative method to determine the time evolution of a system. Poisson brackets directly suggest commutation relations in the quantum theory and a procedure for canonical quantization; however strictly speaking, one cannot derive the quantum theory from the classical one. The utility of the Poisson brackets includes deducing “constants” of the motion as conserved quantities.

4.4.1 DEFINITION OF THE POISSON BRACKET AND RELATION TO THE COMMUTATOR

We first define the Poisson brackets using the “[...]” similar to the commutator discussed in Chapter 3. However, the classical Poisson brackets involve derivatives of functions where as the quantum mechanical commutators do not have this general form.

Definition: Let $A = A(q_i, p_i)$ $B = B(q_i, p_i)$ be two differentiable functions of the generalized coordinates and momentum. We define the Poisson brackets by

$$[A, B] = \sum_i \left[\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial B}{\partial q_i} \frac{\partial A}{\partial p_i} \right]$$

Sometimes we subscript the brackets with q, p

$$[A, B] = [A, B]_{q,p}$$

The Poisson bracket and commutator appear similar (when one ignores the fact that Poisson brackets have derivatives) and provide somewhat similar formulations for the dynamics of a system. In the quantum theory, operators replace the classical dynamical variables (e.g., p 's and q 's). In fact, one starting method for finding the quantum Hamilton consists of determining the classical Hamiltonian, then substituting operators for the classical dynamical variables, and then specifying the commutators for those operators. Chapter 5 will show how the Heisenberg quantum picture is the closest analog to classical mechanics because the operators carry the system dynamics. In quantum theory, the commutation relations give time derivatives of operators where recall that the commutator is defined by $[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}$ with $\hat{A}, \hat{B}$ as operators. In the classical theories, the Hamiltonian uses functions for the dynamical variables (such as momentum p) and the quantum theory replaces the functions with operators (such as $\hat{p}$). Both the commutation relations and Poisson brackets determine the evolution of the dynamical variables.

4.4.2 BASIC PROPERTIES FOR THE POISSON BRACKET

Some basic properties can be proved from the basic definition of the Poisson brackets.

1. Let A, B be functions of the phase space coordinates q, p and let c be a number then

$$[A, A] = 0 \quad [A, B] = -[B, A] \quad [A, c] = 0$$

2. Let A, B, C be differentiable functions of the phase space coordinates q, p then

$$[A + B, C] = [A, C] + [B, C] \quad [A, BC] = [A, B]C + B[A, C]$$

3. The time evolution of the dynamical variable A (for example) can be calculated by

$$\frac{dA}{dt} = [A, H] + \frac{\partial A}{\partial t}$$

Proof:

$$\frac{dA}{dt} = \sum_i \left[\frac{\partial A}{\partial q_i} \frac{dq_i}{dt} + \frac{dp_i}{dt} \frac{\partial A}{\partial p_i} \right] + \frac{\partial A}{\partial t}$$

We include the partial with respect to time in case A explicitly depends on time. Substituting the two relations for the rate of change of position and momentum

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

the Poisson brackets become

$$\frac{dA}{dt} = \sum_i \left[\frac{\partial A}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial H}{\partial q_i} \right] + \frac{\partial A}{\partial t} = [A, H] + \frac{\partial A}{\partial t}$$

Although the order of multiplication $AH=HA$ does not matter in classical theory, the order must be maintained in quantum theory. In quantum theory, the order of two operators can only be switched by using the commutation relations.

$$4. \dot{q}_m = [q_m, H] \quad \dot{p}_m = [p_m, H]$$

Proof: Consider the first one for example

$$[q_m, H] = \sum_i \left[\frac{\partial q_m}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial q_i} \frac{\partial q_m}{\partial p_i} \right] = \sum_i \left[\delta_{im} \frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial q_i} 0 \right] = \frac{\partial H}{\partial p_m} = \dot{q}_m$$

$$5. [q_i, q_j] = 0 \quad [p_i, p_j] = 0 \quad [q_i, p_j] = \delta_{ij}$$

These properties are all very similar to those that arise in the quantum theory.

4.4.3 CONSTANTS OF THE MOTION AND CONSERVED QUANTITIES

One can show that a dynamical variable that commutes (in the sense of the Poisson bracket) with the Hamiltonian corresponds to a conserved quantity. The conservation can be seen from Property #3 in Section 4.2.2

$$\frac{dA}{dt} = [A, H] + \frac{\partial A}{\partial t}$$

when $\partial_t A = 0$ and $[A, H] = 0$ so that $A = \text{constant}$. The use of $\partial_t A = 0$ indicates that the dynamical variable A only has time dependence through the canonical phase space coordinates. Several examples are in order.

Example 4.8: Conservation of Energy

Assume a closed system whereby energy does not enter the system under consideration (i.e., the system described by the Hamiltonian) so that $\partial_t H = 0$. Then Property #3 in Section 4.2.2 provides $H = \text{constant}$ since the order of derivatives in the Poisson brackets $[H, H]$ does not matter.

Example 4.9: Conservation of Momentum

Starting with, for example, Property #4 in Section 4.2.2, $\dot{p}_m = [p_m, H]$, then a zero Poisson bracket produces $p_m = \text{constant}$.

Example 4.10: Cyclic Coordinates

If a Hamiltonian does not depend on a coordinate q_m (the definition of cyclic coordinate) then the conjugate momentum p_m must be conserved. This can be seen from either Hamilton's relations or from the Poisson brackets. Hamilton's relation provides $\dot{p}_m = -\partial H / \partial q_m = 0$ so that $p_m = \text{constant}$. The fact that q_m is cyclic produces a zero Poisson bracket in Property #3 above and thereby leads to the same results.

Example 4.11: Equation of Motion

$$\text{Suppose } H = \frac{p^2}{2m} + \frac{k}{2}x^2 \quad \text{find } [p, H]$$

SOLUTION

$\dot{p} = [p, H] = \frac{\partial p}{\partial x} \frac{\partial H}{\partial p} - \frac{\partial H}{\partial x} \frac{\partial p}{\partial p} = 0 - kx$ which is Newton's second law for the motion of a mass on a spring.

4.5 LAGRANGIAN AND NORMAL COORDINATES FOR A DISCRETE ARRAY OF PARTICLES

The motion of an array of particles provides an example for the Lagrangian as well as the use of normal modes, which among other topics, have applications to phonons. The generalized coordinates for an array of particles describes the displacement of a particle from its equilibrium position. For phonons, the system consists of atoms capable of moving about their equilibrium point. A generalized coordinate in this case describes the displacement of an atom from its equilibrium position. However, the solution to the equation of motion for each atom consists of a Fourier summation over the eigenfrequencies. The motion of each atom does not exhibit the simplest case of translational motion since it does not necessarily exhibit a single oscillation frequency nor does it show the collective behavior of the particles in the array. The normal coordinates provide an example of a coordinate transformation and explicitly show how the oscillation modes of all the atoms can be decoupled.

4.5.1 LAGRANGIAN AND EQUATIONS OF MOTION

Consider a linear array of atoms of mass m linked to nearest neighbors by a quadratic potential. Figure 4.8 shows that x_n labels the equilibrium position of atom $\#n$ and the generalized coordinate u_n represents the displacement of atom $\#n$ from its equilibrium position. Atom $\#n$ exists in an electrostatic "potential well" V_n created by its immediate neighbors. The potential energy depends on the separation between the atoms (rather than on the indices x_n) since the atoms give rise to the potential. One often assumes only nearest neighbor atoms, namely $\#(n-1)$ and $\#(n+1)$, directly exert forces on atom $\#n$ through the electrostatic potential. The displacement of atom $\#n$ from equilibrium is represented by u_n as shown in Figure 4.8. The potential for atom $\#n$ has the Taylor expansion

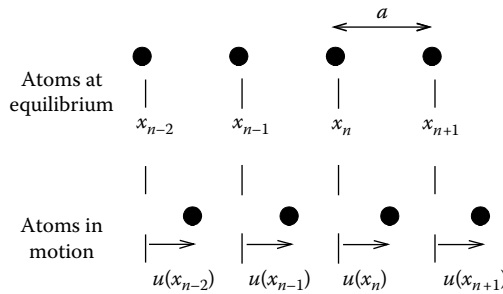


FIGURE 4.8 Top: Atoms at their equilibrium positions. Bottom: Atoms displaced from their equilibrium positions.

$$V_n(u_n + x_n) \cong V(x_n) + \left. \frac{dV_n}{du_n} \right|_{x_n} u_n + \frac{1}{2} \left. \frac{d^2 V_n}{du_n^2} \right|_{x_n} u_n^2 + \cdots \quad (4.22)$$

The equilibrium point x_n corresponds to zero slope and therefore the term linear in the Taylor expansion must be zero. Equation 4.22 has a form similar to that for a linear array of atoms with mass m interconnected by springs with spring constant β_m . The validity of the spring model can be seen as follows. The quadratic term has the form $\beta x^2/2$ which arises from the linear force of the form $F = -\beta x$ similar to Hook's law for springs but with x replaced by u_n and the parameter β as the spring constant. Therefore we identify the spring constant as $\beta \sim \left. \frac{d^2 V}{du_n^2} \right|_{x_n}$ which arises from the quadratic approximation for the electrostatic potential. The first term $V(x_n)$ can be taken as zero by shifting the zero of energy. The term with the first derivative of V is likewise zero since the potential is evaluated at equilibrium. The potential energy term in the Lagrangian is a result of stretching the spring from equilibrium by an amount $u_n - u_{n-1}$. All springs must be included.

The Lagrangian L consists of the difference between total kinetic and potential energy.

$$L = T - V = \sum_{m=1}^{N+1} \frac{1}{2} m \dot{u}_m^2 - \sum_{m=1}^{N+1} \frac{\beta_m}{2} (u_m - u_{m-1})^2 \quad (4.23)$$

where the coupling constant β_m couples atom $\#m$ with its nearest neighbor $\#(m-1)$. Assume that atom $\#0$ and $\#(N+1)$ have fixed positions (i.e., fixed endpoint boundary conditions). The terms involving $m=0$ and $m=N+1$ do not contribute to the summation as these atoms remain fixed in place. Sometimes, one assumes a single type of atom comprises the linear array and therefore there exists only one coupling constant $\beta_m = \beta$. However, we do not make this assumption.

Lagrange's equations take the usual form

$$\frac{\partial L}{\partial u_n} - \frac{d}{dt} \frac{\partial L}{\partial \dot{u}_n} = 0 \quad (4.24)$$

Using the fact that the generalized coordinates and velocities are independent

$$\frac{\partial u_m}{\partial u_n} = \delta_{mn} \quad \frac{\partial \dot{u}_m}{\partial u_n} = 0 \quad \frac{\partial \dot{u}_m}{\partial \dot{u}_n} = \delta_{mn}$$

the equation of motion for atom $\#n$ becomes

$$m \ddot{u}_n + (\beta_{n+1} + \beta_n) u_n - \beta_{n+1} u_{n+1} - \beta_n u_{n-1} = 0 \quad (4.25)$$

4.5.2 TRANSFORMATION TO NORMAL COORDINATES

The coordinates u_n focus on the motion of each individual atom $\#n$. The interaction of atom $\#n$ with other atoms produces a complicated motion for atom $\#n$ consisting of multiple Fourier components (i.e., multiple frequencies and uncorrelated phases or amplitudes of oscillation). On the other hand, the normal coordinates describe a collective motion with a single frequency. The focus shifts from a single atom to a spatially extended sinusoidal wave on the crystal. Each atom participating in the oscillation has the same oscillation frequency as every other. The normal modes can be Fourier summed to provide the general wave in the crystal. The phonon normally refers to the smallest quantum of energy for the amplitude of the normal mode. In this sense, the phonon energy must be distributed across all of the atoms participating in the collective motion to form the normal modes; that is, the phonon is not associated with any single atom. The present section illustrates the

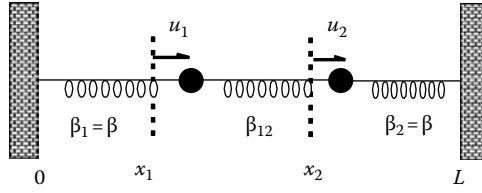


FIGURE 4.9 Longitudinal vibration of masses m coupled by springs.

difference between the motion of single atoms and those participating in the collective motion for the normal modes.

A simple demonstration of normal modes uses two atoms as shown in Figure 4.9 (see Marion's book on *Classical Dynamics* for more details). Notice the middle coupling constant differs from that at either end. The equations of motion (Equation 4.25) provide the results

$$\begin{aligned} m\ddot{u}_1 + (\beta + \beta_{12})u_1 - \beta_{12}u_2 &= 0 \\ m\ddot{u}_2 + (\beta + \beta_{12})u_2 - \beta_{12}u_1 &= 0 \end{aligned} \quad (4.26)$$

The fundamental solutions have the form $e^{i\omega t}$. For our case, there will be two independent positive angular frequencies ω_1 and ω_2 (for real sinusoidal solutions) and four frequencies counting the negative values for complex exponential solutions (which must be combined to give the sinusoidal solutions). We start with the angular frequency variable ω and find the specific angular frequencies ω_1 and ω_2 .

$$\begin{aligned} u_1(t) &\sim B_1 e^{i\omega t} \\ u_2(t) &\sim B_2 e^{i\omega t} \end{aligned} \quad \text{or} \quad \underline{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} e^{i\omega t} \quad (4.27a)$$

For each (positive) angular frequency, there will be a solution for the column vector consisting of B_1 and B_2 . In general, each column vector will be represented by

$$\underline{a}^{(i)} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \quad (4.27b)$$

Substitute and collect terms to write the matrix equation

$$\begin{bmatrix} \beta + \beta_{12} - m\omega^2 & -\beta_{12} \\ -\beta_{12} & \beta + \beta_{12} - m\omega^2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \underline{0} \quad (4.28)$$

If the matrix has an inverse then we would find that $B_1 = 0 = B_2$ and the atoms would not move from equilibrium. Such a solution does not describe wave motion. Therefore, we must require the matrix to be noninvertible by requiring its determinant to be zero. Solving for the frequency provides four roots. Define the positive angular frequencies

$$\omega_1 = \sqrt{\frac{\beta}{m}} \quad \text{and} \quad \omega_2 = \sqrt{\frac{\beta + 2\beta_{12}}{m}} \quad (4.29)$$

so that all four angular frequencies will be $\pm\omega_1, \pm\omega_2$ (the solutions must be real and consists of sinusoids).

Before continuing, consider the following observation. If one mass were held in place, and the equations solved for the other mass, the oscillation frequency would be $\omega_0 = \sqrt{(\beta + \beta_{12})/m}$. Therefore using Equation 4.29, the coupling for the two masses “splits” the oscillation frequency according to $\omega_1 < \omega_0 < \omega_2$. Suppose N particles of mass m appear in the linear chain and that the N atoms are free to move. For N an even number, there will be $N/2$ frequencies above and $N/2$ frequencies below ω_0 , while for N an odd number, there will be $(N-1)/2$ above, $(N-1)/2$ below, and 1 equal to ω_0 . Also notice the number of degrees of freedom for the atoms matches the number of allowed frequencies.

Each positive frequency provides a different set of $\{B_1, B_2\}$ (i.e., a different eigenmode). Represent each different set of B 's by a column vector

$$\underline{a} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \quad (4.30)$$

Add a superscript to form $\underline{a}^{(i)}$ in order to indicate the particular set of B 's which correspond to the positive eigenfrequency ω_i . In the present case, the two column vectors $\underline{a}^{(i)}$ consisting of B_1 and B_2 (each column vector has different B_1 and B_2) can be found by substituting Equation 4.29 into the matrix Equation 4.28 which can be seen to produce the two solutions $B_1 = B_2$ and $B_1 = -B_2$ for the two different positive frequencies ω_1 and ω_2 , respectively. Therefore the column vector solutions become

$$\underline{a}^{(1)} = \begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \end{pmatrix} = \begin{pmatrix} B_1 \\ B_1 \end{pmatrix} \sim \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \underline{a}^{(2)} = \begin{pmatrix} a_1^{(2)} \\ a_2^{(2)} \end{pmatrix} = \begin{pmatrix} B_1 \\ -B_1 \end{pmatrix} \sim \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad (4.31)$$

A normalization factor of $1/\sqrt{2}$ can be included to normalize the eigenvectors $\underline{a}^{(i)}$ to 1. Equations 4.31 show that the angular frequencies define the modes for the masses to either move 180° out of phase or to move completely in phase (i.e., the displacement between them does not change and they oscillate together). The solutions must be a summation of four terms for the complex exponentials.

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = b_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{i\omega_1 t} + b_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-i\omega_1 t} + b_3 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{i\omega_2 t} + b_4 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-i\omega_2 t} \quad (4.32a)$$

The solutions u_1 and u_2 consist of a linear combination of complex exponentials in time having the four possible frequencies listed by Equations 4.29 and their negative counterparts. By looking at the solution for each u_i , this last summation can be seen to be identical with that obtained from the usual method of solving differential equations since each exponential has different amplitude. However, this last equation accounts for the relation between B_1 and B_2 and therefore reduces eight parameters (for u_1 and u_2) to the four shown. Given that the solutions must be real, the complex exponentials in Equation 4.32a can be combined to write

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sin(\omega_1 t + \phi_1) + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} \sin(\omega_2 t + \phi_2) \quad (4.32b)$$

where c_j represent real numbers. Each individual sinusoidal term is related to a normal mode.

As an important point, the motion of either atom (focus on u_1 for example) has quite complicated time dependence as it consists of a mixture of two different Fourier components. The complexity arises because we focus on the individual atoms (i.e., u_n represents the coordinate of atom # n) rather than a simpler wave motion as described by the “normal coordinates” for which one focuses on specific collective motions of all the atoms as described next. The normal modes appear as sinusoidal waves in space (c.f., the discussion associated with the transverse motion in Figure 4.10). These fundamental modes can be Fourier superposed to describe the more complicated motions of each atom.

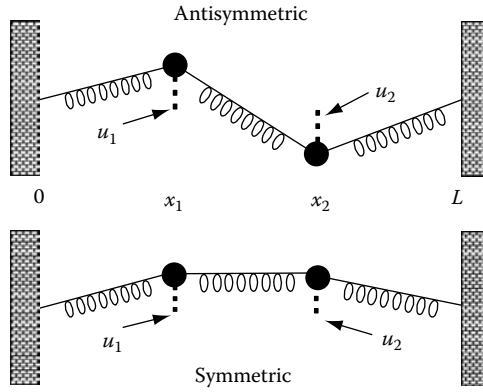


FIGURE 4.10 The two normal modes for transverse oscillations on a spring system with two masses confined to the single transverse motion.

As mentioned, normal modes represent a simpler (and perhaps more intuitive) motion of the atoms (c.f., the discussion associated with the transverse motion in Figure 4.10 above). One looks for a linear combination of normal modes v_j to produce the original motion of each atom u_i . In general, one looks for a transformation matrix A_{ij} which has elements $a_{ij} = A_{ij}$ (for notational convenience) such that

$$u_i = \sum_j a_{ij} v_j \quad \text{or equivalently} \quad \underline{u} = \underline{A} \underline{v} \quad (4.33a)$$

where the a_{ij} are related to the eigenvectors found in Equation 4.31 (for example). In particular, as shown in Section 4.5.3, and writing $\underline{A}$ in column vectors consisting of the columns formed by a_{ij}

$$\underline{A} = [\underline{a}^{(1)}, \underline{a}^{(2)}, \dots] = \left[\begin{pmatrix} a_{i=1}^{(j=1)} \\ a_2^{(1)} \\ \vdots \end{pmatrix} \begin{pmatrix} a_{i=1}^{(j=2)} \\ a_2^{(2)} \\ \vdots \end{pmatrix} \dots \right] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & \vdots \\ a_{31} & \dots & \end{bmatrix} \quad (4.33b)$$

The matrix $\underline{A}$ consists of eigenvectors and should remind the reader of the matrix used to make a matrix diagonal from Chapter 3; however, we will choose a normalization as necessary for convenience. For our case of two particles, Equation 4.31 provides the matrix

$$\underline{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad (4.33c)$$

The coordinates v_j for the normal modes obtain from a linear combination of the atomic coordinates u_i and resemble the coordinates for the motion of the center of mass and the coordinates of the group of atoms with respect to the center of mass. We find

$$\begin{aligned} u_1 &= v_1 + v_2 & \text{or equivalently} & & v_1 &= (u_1 + u_2)/2 \\ u_2 &= v_1 - v_2 & & & v_2 &= (u_1 - u_2)/2 \end{aligned} \quad (4.34)$$

Substitute into Equation 4.26 and separate variables to find

$$\begin{aligned} m\ddot{v}_1 + (\beta + 2\beta_{12})v_1 &= 0 \\ m\ddot{v}_2 + \beta v_2 &= 0 \end{aligned} \quad (4.35)$$

TABLE 4.1
Specific Examples for the Normal Modes

Initial Conditions	Solutions
$u_1(0) = -u_2(0)$	$v_1(t) = 0 \rightarrow u_1(t) = -u_2(t)$
$\dot{u}_1(0) = -\dot{u}_2(0)$	$\omega_1 = \sqrt{\frac{\beta + 2\beta_{12}}{m}}$
$u_1(0) = u_2(0)$	$v_2(t) = 0 \rightarrow u_1(t) = u_2(t)$
$\dot{u}_1(0) = \dot{u}_2(0)$	$\omega_2 = \sqrt{\frac{\beta}{m}}$

The uncoupled solutions can be written as

$$\begin{aligned} v_1(t) &= d_1 e^{i\omega_1 t} + d_2 e^{-i\omega_1 t} \\ v_2(t) &= d_3 e^{i\omega_2 t} + d_4 e^{-i\omega_2 t} \end{aligned} \quad (4.36)$$

where d_i are constants. The motion can be easily visualized for the specific initial conditions given in Table 4.1.

The first set of initial conditions corresponding to ω_1 provide a stationary center of mass and the two atoms oscillate 180° out of phase. The second set corresponding to ω_2 shows both atoms oscillate in phase which gives the center-of-mass a sinusoidal time dependence.

Instead of the longitudinal waves shown in Figure 4.8, consider the transverse waves shown in Figure 4.10 where it is easy to see the antisymmetric character for ω_1 and the symmetric character for ω_2 . Notice the shape of the normal modes along the x -axis approximates a sine wave with wavelength either $\lambda = L$ or $\lambda = 2L$ which provides a wave vector of either $k = 2\pi/L$ or $k = \pi/L$. Notice further, the number of normal modes, frequencies, and wave numbers k coincide with the number of degrees of freedom of 2 for the system. The number of degrees of freedom equals the number of dimensions that the particles can independently move. Each atom can move in one direction in this case but including the two atoms provides the 2 degrees of freedom. The modes of a system can refer to the frequencies, wave numbers, polarization, or shapes depending on how the term appears in context. For shape, one refers to the time-independent shape as the mode (a time-independent sinusoid in this case) but more exactly refers to the time-independent eigenfunctions of the wave equation.

The normal modes could have been found from Equation 4.26 and Figure 4.10 (longitudinal motion) by assuming a solution of the form

$$u_n = u(x_n, t) = A_k e^{ikx_n - i\omega_k t}$$

where

A represents the amplitude

x_n the equilibrium position has the value $x_n = na$, where a provides the atomic spacing at equilibrium

The boundary conditions determine the values of k .

Section 4.5.3 discusses the theoretical basis for normal modes of coupled oscillators with attention to wave motion of a linear array of N masses coupled by quadratic potentials (i.e., springs). Section 4.5.3 first focuses on the motion of each individual mass with coordinate u_n . The section shows there results an $N \times N$ determinant equation that must be solved for the fundamental frequencies (i.e., the frequencies of the normal modes).

4.5.3 LAGRANGIAN AND THE NORMAL MODES

The objective consists of showing there exists a transformation from the original coordinates u_i to new coordinates v_i given by matrix notation as

$$u_i = \sum_j a_{ij} v_j \quad \text{or equivalently} \quad \underline{u} = \underline{A} \underline{v} \quad (4.37a)$$

such that the Lagrangian becomes a sum of independent modes according to

$$L = \frac{1}{2} \sum_i [\dot{v}_i^2 - \lambda_i v_i^2] \quad \text{or equivalently} \quad L = \frac{1}{2} \{ \dot{\underline{v}}^T \dot{\underline{v}} - \underline{v}^T \underline{\lambda} \underline{v} \} \quad (4.37b)$$

where the original Lagrangian for a quadratic potential has the form

$$L = T - V = \frac{1}{2} \sum_{i,j} [T_{ij} \dot{u}_i \dot{u}_j - V_{ij} u_i u_j] \quad (4.38)$$

The matrix $\underline{\lambda}$ has all zero elements except those along the diagonal that have the value λ_i ; that is, the matrix has the elements $\lambda_{ij} = \lambda_i \delta_{ij}$. Both T_{ij} and V_{ij} are symmetric ($T^T = T$ and $V^T = V$). The original Lagrangian in Equation 4.38 produces the equation of motion

$$\sum_j (T_{ij} \ddot{u}_j + V_{ij} u_j) = 0 \quad \text{or} \quad \underline{T} \ddot{\underline{u}} + \underline{V} \underline{u} = 0 \quad (4.39a)$$

Equation 4.39a shows that the motion of each individual particle couples with every other to produce complicated motions. If both T_{ij} and V_{ij} are diagonal, then the equations of motion become

$$T_{ii} \ddot{u}_i + V_{ii} u_i = 0 \quad (4.39b)$$

and the motions decouple since the summation is eliminated.

That the potential V in Equation 4.38 depends on the product $u_i u_j$ can easily be seen by Taylor expanding the potential V with the previous notation of $q_i = x_i + u_i$ where x_i represents the equilibrium position of atom $\#i$ and u_i represents the displacement of atom $\#i$ from equilibrium (also see Equation 4.22).

$$V(q_1, q_2, \dots) = V(x_1, x_2, \dots) + \sum_i \left. \frac{\partial V}{\partial q_i} \right|_0 u_i + \frac{1}{2} \sum_{i,j} \left. \frac{\partial^2 V}{\partial q_i \partial q_j} \right|_0 u_i u_j + \dots \quad (4.40)$$

Here the subscript 0 signifies that the functions and derivatives must be evaluated at the equilibrium positions x_i . A similar result obtains for the kinetic energy T . Notice that $V_{ij} = \partial_{q_i} \partial_{q_j} V$ (evaluated at the equilibrium position) represent a collection of expansion coefficients from Equation 4.40 that can be written as a matrix $\underline{V}$. The first term in Equation 4.40 can be taken as zero by shifting the zero of energy, while the second term linear in u_i must be zero by consequence of evaluating the derivative at the equilibrium position.

The Lagrangian in Equation 4.38 can be written in matrix notation as

$$L = \frac{1}{2} \{ \dot{\underline{u}}^T \underline{T} \dot{\underline{u}} - \underline{u}^T \underline{V} \underline{u} \} \quad (4.41)$$

where the superscript T represents the transpose. The matrix notation will help simplify the mathematical manipulations. Both matrices $\underline{T}$ and $\underline{V}$ are real and symmetric (i.e., the matrix and its transpose are identical). The demonstration for the normal modes starts by substituting the coordinate transformation from Equation 4.37a into Equation 4.41.

$$L = \frac{1}{2} \{ \dot{\underline{v}}^T (\underline{A}^T \underline{T} \underline{A}) \dot{\underline{v}} - \underline{v}^T (\underline{A}^T \underline{V} \underline{A}) \underline{v} \} \quad (4.42)$$

It is only necessary to show that $\underline{A}^T \underline{T} \underline{A} = \underline{1}$ and $\underline{A}^T \underline{V} \underline{A} = \underline{\lambda}$ where $\underline{1}$ represents the unit matrix and both $\underline{1}$ and $\underline{\lambda}$ are diagonal.

Similar to Section 4.5.2, the fundamental modes can be found by substituting

$$\underline{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \end{pmatrix} \sim \begin{pmatrix} B_1 \\ B_2 \\ \vdots \end{pmatrix} e^{i\omega t} = \underline{B} e^{i\omega t} \quad (4.43)$$

into Equation 4.39a to find

$$(\underline{V} - \omega^2 \underline{T}) \underline{B} = 0 \quad (4.44)$$

As before, one must require $\det(\underline{V} - \omega^2 \underline{T}) = 0$ in order that the amplitudes B_i can be nonzero. For the number N atoms capable of moving, the matrix will be $N \times N$ and there will be N positive frequencies ω_j (and N negative frequencies $-\omega_j$) and N eigenvectors $\underline{a}^{(j)} = \underline{B}$ (one for each pair of frequencies $\omega_j, -\omega_j$). The positive and negative frequency pair will combine to produce a real sinusoidal oscillation. Identify the $N \times N$ diagonal matrix $\underline{\lambda}$ as $\lambda_{ij} = \omega_i^2 \delta_{ij}$ where δ_{ij} represents the Kronecker delta function. Each column vector has the form

$$\underline{a}^{(j)} = \begin{pmatrix} a_{i=1}^{(j)} \\ a_{i=2}^{(j)} \\ \vdots \end{pmatrix} = \begin{pmatrix} a_{i=1,j} \\ a_{i=2,j} \\ \vdots \end{pmatrix} \quad (4.45a)$$

where j designates the particular column vector. The second column vector in Equation 4.45a, develops the subscripts for the column vectors to be arranged as a square matrix

$$\underline{A} = [\underline{a}^{(1)}, \underline{a}^{(2)}, \dots] = \left[\begin{pmatrix} a_{i=1}^{(j=1)} \\ a_2^{(1)} \\ \vdots \end{pmatrix} \begin{pmatrix} a_{i=1}^{(j=2)} \\ a_2^{(2)} \\ \vdots \end{pmatrix} \dots \right] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & \vdots \\ a_{31} & \dots & \end{bmatrix} \quad (4.45b)$$

where the index i indicates the row. Applying these definitions changes Equation 4.44 into

$$\underline{V} \underline{a}^{(i)} = \underline{\lambda} \underline{T} \underline{a}^{(i)} \quad \text{or} \quad \underline{V} \underline{A} = \underline{\lambda} \underline{T} \underline{A} \quad (4.46)$$

which has the form of an eigenvalue equation (with a “weight” T , see Section 2.4.5).

For each different positive frequency ω_j , there will be a different column vector $\underline{a}^{(j)} = a_{ij}$. Then Equation 4.46 can be viewed as the sequence

$$[\underline{V} - \omega_1^2 \underline{T}] \underline{a}^{(1)} = 0 \quad \dots \quad [\underline{V} - \omega_j^2 \underline{T}] \underline{a}^{(j)} = 0 \dots \quad (4.47a)$$

or letting $\lambda_{ij} = \lambda_j \delta_{ij} = \omega_j^2 \delta_{ij}$ provides

$$[\underline{V} - \lambda_j \underline{T}] \underline{a}^{(j)} = 0 \quad \text{or} \quad \underline{V} \underline{a}^{(j)} = \lambda_j \underline{T} \underline{a}^{(j)} \quad (4.47b)$$

which form column # j in the square matrix $\underline{A}$.

Next we show that $\underline{A}^T \underline{TA} = \underline{1}$. Similar to showing Hermitian operators have orthonormal eigenvectors, consider two different eigenvalues $\lambda_\xi \neq \lambda_\eta$ and use the second of Equation 4.47b to write

$$\underline{V} \underline{a}^{(\xi)} = \lambda_\xi \underline{T} \underline{a}^{(\xi)} \quad \text{and} \quad \underline{V} \underline{a}^{(\eta)} = \lambda_\eta \underline{T} \underline{a}^{(\eta)} \quad (4.48)$$

Multiply the first by the transpose $\underline{a}^{(\eta)T}$ and the second by $\underline{a}^{(\xi)T}$ to find

$$\underline{a}^{(\eta)T} \underline{V} \underline{a}^{(\xi)} = \lambda_\xi \underline{a}^{(\eta)T} \underline{T} \underline{a}^{(\xi)} \quad \text{and} \quad \underline{a}^{(\xi)T} \underline{V} \underline{a}^{(\eta)} = \lambda_\eta \underline{a}^{(\xi)T} \underline{T} \underline{a}^{(\eta)} \quad (4.49)$$

Take the transpose of the second Equation 4.49 and subtract the two to obtain

$$0 = (\lambda_\xi - \lambda_\eta) \underline{a}^{(\eta)T} \underline{T} \underline{a}^{(\xi)} \quad (4.50)$$

Since the eigenvalues are not equal, the last equation requires

$$\underline{a}^{(\eta)T} \underline{T} \underline{a}^{(\xi)} = 0 \quad (4.51a)$$

The eigenvectors $\underline{a}^{(i)}$ can be normalized so that for the same eigenvectors, one can produce (refer to Goldstein's book on *Classical Mechanics* for more details)

$$\underline{a}^{(\eta)T} \underline{T} \underline{a}^{(\xi)} = \delta_{ij} \quad \text{or} \quad \underline{A}^T \underline{TA} = \underline{1} \quad (4.51b)$$

Consequently Equation 4.46 can be rewritten by multiplying on the left by $\underline{A}^T$ and using Equation 4.51b to find

$$\underline{A}^T \underline{VA} = \underline{\lambda} \quad (4.52)$$

As a result Equations 4.51b and 4.52 show that both $\underline{A}^T \underline{TA} = \underline{1}$ and $\underline{A}^T \underline{VA} = \underline{\lambda}$ represent diagonal matrices which then justifies Equations 4.39a and 4.39b. The modes represented by the v_i are the normal modes.

4.6 CLASSICAL FIELD THEORY

Up to now, the discussion has centered on the classical Lagrangian and Hamiltonian for discrete sets of generalized coordinates and their conjugate momentum. Now we turn our attention to systems with an uncountably infinite number of coordinates. The present section first discusses the relation between discrete and continuous system, and then shows how the Lagrangian for sets of discrete coordinates leads to the Lagrangian for the continuous set of coordinates. This latter Lagrangian begins the study of classical field theories since it can produce the Maxwell equations, the Schrödinger equation, and it begins the quantum field theory for particles and the quantum electrodynamics. The present section demonstrates the Lagrangian for the wave motion in a continuous media that has applications to phonon fields and provides an example for the later field theory of electromagnetic fields.

4.6.1 LAGRANGIAN AND HAMILTONIAN DENSITY

For systems with a continuous set of generalized coordinates, Lagrange's and Hamilton's formulation of dynamics must be generalized. First, we discuss the generalized coordinates and velocities. There are an uncountable number of these coordinates. Second, we show how a continuous system can be viewed as a discrete one with a countable number of generalized coordinates. Third, we derive the generalized momentum for the Hamiltonian density. We end the discussion with a summary. The following sections apply the procedure to wave motion in a continuous medium.

For the continuous coordinate case, consider the following imagery. Suppose the indices x, y, z in $\vec{r} = x\hat{x} + y\hat{y} + z\hat{z}$ label each point in space. The value of a function $\eta(\vec{r}, t) = \eta(x, y, z, t)$ serves as a generalized coordinate indexed by the point $\vec{r}$. Figure 4.11 shows some of the generalized coordinates along the z -direction. The lower left side shows a small volume of space with a field (electromagnetic in origin). The field has a different value for each point. The field at a particular point is the generalized coordinate at that point. The lower right side shows another example for the generalized coordinates. Here η represents the displacement of small masses. The generalized velocities are given by $\dot{\eta}$.

Now let us discuss how the continuous coordinates $\eta(\vec{r}, t)$ compare with the discrete ones q_i . The top panel of Figure 4.11 shows all of space divided into many cells of volume ΔV_i . In "each" cell, the field $\eta(z, t)$ takes on many similar values. We can define the "discrete" generalized coordinates by the average

$$q_i(t) = \frac{1}{\Delta V_i} \int_{\Delta V_i} dV \eta(\vec{r}, t) \quad (4.53)$$

The q_i represent the average value of the continuous coordinate in the given cell. Making ΔV_i small enough means that the η under the integral is approximately constant so that

$$q_i(t) = \frac{1}{\Delta V_i} \int_{\Delta V_i} dV \eta(\vec{r}, t) \rightarrow \eta(\vec{r}, t) \quad (4.54)$$

Notice that the small volume ΔV_i is associated with the points x, y, z in space and not with the "tops" of $\eta(\vec{r}, t)$. In Section 4.6.2, we will show displaced small boxes but those will be different boxes. Those boxes will refer to actual chunks of mass displaced from equilibrium. The procedure given in the present section uses the small cells in Figure 4.11 to show how the continuous and discrete Lagrangians can be interrelated.

Next we compare the Lagrangians for the two systems. For continuous sets of coordinates, people usually work with the Lagrange density $\mathcal{L}$ defined through

$$L = \int_V dV \mathcal{L} \quad (4.55)$$

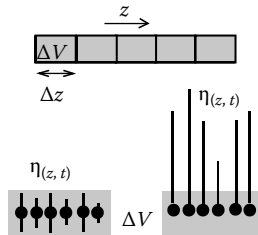


FIGURE 4.11 Top portion shows space divided into cells. Bottom portion shows two types of continuous coordinates. Left side shows a field and the right side shows displacement of small masses.

where the Lagrange density has units of “energy per volume.” The Lagrange density has the form

$$\mathcal{L} = \mathcal{L}(\eta, \dot{\eta}, \partial_i \eta) \quad (4.56)$$

where $i = 1, 2, 3$ refers to derivatives with respect to x, y, z , respectively. The Lagrange density refers to a single point in space (or possibly two arbitrarily close points due to the derivatives).

On the other hand, suppose we divide all space into cells of volume ΔV_i with $q_i, \dot{q}_i$ being the generalized coordinate and velocity in cell $\#i$. The full Lagrangian must have the form

$$L = L(q_i, \dot{q}_i, q_{i\pm 1}) \quad (4.57)$$

where the $q_{i\pm 1}$ allows for derivatives. Especially note that all coordinates $i = 1, 2, 3, 4, \dots$ occur in the full Lagrangian. Now to make the connection with the Lagrange density, apply the cellular space to the full Lagrangian in Equation 4.57. Dividing up the volume V into cells so that $V = \sum_i \Delta V_i$ we can write

$$L(q_i, \dot{q}_i, q_{i\pm 1}) = \int_V dV \mathcal{L}(q_i, \dot{q}_i, q_{i\pm 1}) = \sum_i \int_{\Delta V_i} dV \mathcal{L}(q_i, \dot{q}_i, q_{i\pm 1}) \quad (4.58)$$

The definition of an average from calculus provides

$$\bar{\mathcal{L}}_i = \frac{1}{\Delta V_i} \int_{\Delta V_i} dV \mathcal{L} \quad \text{so that} \quad L = \sum_i \int_{\Delta V_i} dV \mathcal{L} = \sum_i \Delta V_i \bar{\mathcal{L}}_i(q_i, \dot{q}_i, q_{i\pm 1}) \quad (4.59)$$

where now each ΔV_i has one q_i and one $\dot{q}_i$ associated with it on account of Equation 4.53. Making ΔV_i small enough produces the result in Equation 4.54, namely $q_i \rightarrow \eta$. Similarly small enough ΔV_i allows one to replace the average Lagrangian density with the value of Lagrangian at a “point” as $\bar{\mathcal{L}} \rightarrow \mathcal{L}$. As a result, in the limit $\Delta V_i \rightarrow 0$, the full Lagrangian in Equation 4.59 becomes

$$L = \sum_i \Delta V_i \bar{\mathcal{L}}_i(q_i, \dot{q}_i, q_{i\pm 1}) \rightarrow \int dV \mathcal{L}(\eta, \dot{\eta}, \partial_i \eta) \quad (4.60)$$

This last equation shows how discrete coordinates and the corresponding Lagrangian produce the continuous coordinates and the Lagrangian density.

Finally, we compare the full Hamiltonian with the Hamiltonian density. The full Hamiltonian can be written as

$$H = H(q_i, p_i) = \sum_i p_i \dot{q}_i - L = \sum_i p_i \dot{q}_i - \sum_i \Delta V_i \bar{\mathcal{L}}_i \quad (4.61)$$

We can calculate p_j by the usual method

$$p_j = \frac{\partial L}{\partial \dot{q}_j} = \frac{\partial}{\partial \dot{q}_j} \sum_i \Delta V_i \bar{\mathcal{L}}_i = \sum_i \Delta V_i \frac{\partial \bar{\mathcal{L}}_i}{\partial \dot{q}_j} = \Delta V_j \frac{\partial \bar{\mathcal{L}}_j}{\partial \dot{q}_j} \quad (4.62)$$

where the summation in last term disappears because we assume $\bar{\mathcal{L}}_i$ depends only on $\dot{q}_j$ (along with q_j) and the relation $\partial \dot{q}_i / \partial \dot{q}_j = \delta_{ij}$ holds. Notice how the momentum (Equation 4.62) depends on the volume of the small box whereas the relation $q_j \rightarrow \eta$ does not. For continuous material systems (as opposed to electromagnetic systems), one often writes the momentum in terms

of a small mass which means the momentum is indeed proportional to the small volume $p_j \sim (\Delta m)\dot{q}_j \sim (\Delta V)\rho\dot{q}_j \sim (\Delta V)\pi_j$ where ρ represents the mass density. Similar considerations apply to other continuous systems as well. Therefore, the momentum density can be defined as

$$\Delta V_j \pi_j = p_j = \frac{\partial L}{\partial \dot{q}_j} = \Delta V_j \frac{\partial \bar{L}_j}{\partial \dot{q}_j} \rightarrow \pi_j = \frac{\partial \bar{L}_j}{\partial \dot{q}_j} \xrightarrow{\Delta V_i \rightarrow 0} \pi(\vec{r}, t) = \frac{\partial \mathcal{L}(\eta, \dots)}{\partial \dot{\eta}} \quad (4.63)$$

The full Hamiltonian can be written as a Hamiltonian density

$$H = \int dV \mathcal{H} \quad (4.64)$$

We can write

$$\int d^3x \mathcal{H} = H = \sum_i p_i \dot{q}_i - L = \sum_i \Delta V_i \pi_i \dot{q}_i - \sum_i \Delta V_i \bar{L}_i \rightarrow \int d^3x [\pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \mathcal{L}]$$

and identify the Hamiltonian density as

$$\mathcal{H} = \pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \mathcal{L} \quad (4.65)$$

TABLE 4.2
Summary of Results

Lagrange density	$\mathcal{L} = \mathcal{L}(\eta, \dot{\eta}, \partial_i \eta)$
Lagrangian	$L = \int_V dV \mathcal{L}$
Hamiltonian density	$\mathcal{H} = \pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \mathcal{L}$
Hamiltonian	$H = \int dV \mathcal{H}$
Momentum density	$\pi(\vec{r}, t) = \frac{\partial \mathcal{L}(\eta, \dots)}{\partial \dot{\eta}}$
Hamilton's canonical equations	$\dot{\eta} = \frac{\partial \mathcal{H}}{\partial \pi} \quad \dot{\pi} = -\frac{\partial \mathcal{H}}{\partial \eta}$

4.6.2 LAGRANGE DENSITY FOR 1-D WAVE MOTION

Now we develop the Lagrangian for 1-D wave motion in a continuous medium. As discussed in the previous section, we imagine each point in space to be labeled by indices x, y, z according to $\vec{r} = x\hat{x} + y\hat{y} + z\hat{z}$. The value of a function $\eta(\vec{r}, t) = \eta(x, y, z, t)$ serves as a generalized coordinate indexed by the point $\vec{r}$. Figure 4.12 shows transverse wave motion along the z -axis with η gives the

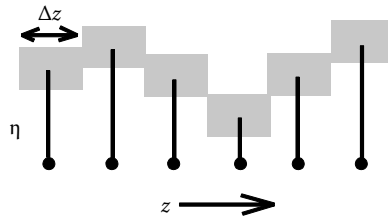


FIGURE 4.12 Displacement of masses at various points along the z -axis.

displacement. The generalized velocity at the point x, y, z can be written as $\dot{\eta}$. Two important notes are in order. First, note that x, y, z do not depend on time since they are treated as indices. Second, the small boxes appearing in Figure 4.12 represent small chunks of matter that the wave displaces from equilibrium. The coordinate q_i denotes the average displacement of the scalar field η for the small chunk.

The description of wave motion requires a partial differential equation involving partial derivatives. We require the partial derivatives to appear in the argument of the Lagrangian. These spatial derivatives take the form $\partial_i \eta$ where i refers to one of the indices x, y, z . For example, $i = 3$ gives $\partial_3 \eta = \partial \eta / \partial z$. For the purpose of the Lagrangian, the spatial derivatives must be independent of each other and of the coordinates.

$$\frac{\partial(\partial_i \eta)}{\partial(\partial_j \eta)} = \delta_{ij} \quad \frac{\partial(\partial_i \eta)}{\partial \eta} = 0 \quad \frac{\partial \eta}{\partial(\partial_i \eta)} = 0$$

The Lagrangian can be written as

$$L = L(\eta, \dot{\eta}, \partial_i \eta) = L(\eta, \dot{\eta}, \partial_1 \eta, \partial_2 \eta, \partial_3 \eta) \quad (4.66)$$

For the transverse wave motion, the partial derivatives actually enter the Lagrangian as a result of the generalized forces acting on each element of volume.

We need to minimize the action

$$I = \int_{t_1}^{t_2} dt L \quad (4.67)$$

However, for continuous systems (i.e., systems with continuous sets of generalized coordinates), it is customary to work with the “Lagrange density” defined by

$$I = \int_{t_1}^{t_2} dt L = \int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt d^3x \mathcal{L}(\eta, \dot{\eta}, \partial_i \eta) \quad (4.68)$$

The Lagrange density $\mathcal{L}$ has units of energy per volume. To find the minimum action, we must vary the integral I so that $\delta I = 0$ where δ represents a small variation due to variations in the path between endpoints. In the process, a partial integration produces a “surface term.” We assume two boundary conditions: one for the time integral and one for the spatial integral. For the time integral, the set of displacements η must be fixed at times t_1, t_2 so that $\delta \eta(t_1) = 0 = \delta \eta(t_2)$. For the spatial integrals, we assume either periodic boundary conditions or fixed-endpoint conditions so that the surface term vanishes.

Now let us find the extremum of the action in Equation 4.68

$$0 = \delta I = \int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt d^3x \delta \mathcal{L}(\eta, \dot{\eta}, \partial_i \eta) = \int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt d^3x \left[\frac{\partial \mathcal{L}}{\partial \eta} \delta \eta + \frac{\partial \mathcal{L}}{\partial \dot{\eta}} \delta \dot{\eta} + \frac{\partial \mathcal{L}}{\partial(\partial_i \eta)} \delta(\partial_i \eta) \right]$$

where we use the Einstein convention for repeated indices in a product, namely $A_i B_i = \sum_i A_i B_i$. Interchanging the differentiation with the variation produces

$$0 = \delta I = \int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt d^3x \left[\frac{\partial \mathcal{L}}{\partial \eta} \delta \eta + \frac{\partial \mathcal{L}}{\partial \dot{\eta}} \frac{\partial}{\partial t} \delta \eta + \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} \partial_i \delta \eta \right]$$

Integrating by parts and using the fact that both the temporal- and spatial-surface terms do not contribute, we find

$$\int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt d^3x \left[\frac{\partial \mathcal{L}}{\partial \eta} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} - \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} \right] \delta \eta = 0$$

Given that the variation at each point is independent of every other, we find Lagrange's equations for the continuous media

$$\frac{\partial \mathcal{L}}{\partial \eta} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} - \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} = 0 \quad (4.69)$$

where the repeated index convention must be enforced on the last term. Notice that the first two terms look very similar to the usual Lagrange equation for the discrete set of generalized coordinates. If desired, we can also include generalized forces in the formalism so that the motion of the waves can be “driven” by an outside force.

Example 4.12

Suppose the Lagrange density has the form $\mathcal{L} = \frac{\rho}{2} \dot{\eta}^2 - \frac{\beta}{2} (\partial_z \eta)^2$ for 1-D motion propagating along the z-direction, where ρ, β resemble the mass density and spring constant (Young's modulus) for the material, and $\eta = \eta(z, t)$. Find the applicable wave equation.

SOLUTION

Lagrange's equation has the following terms

$$\frac{\partial \mathcal{L}}{\partial \eta} = 0 \quad \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} = \rho \ddot{\eta} \quad \frac{\partial}{\partial z} \frac{\partial \mathcal{L}}{\partial (\partial_z \eta)} = -\beta \frac{\partial^2 \eta}{\partial z^2}$$

Equation 4.69 then gives

$$\frac{\partial^2 \eta}{\partial z^2} - \frac{\rho}{\beta} \ddot{\eta} = 0 \quad \text{with speed } v = \sqrt{\beta/\rho}$$

The reader can refer to Section 6.14 for applications to phonons.

Example 4.13

Find π_η for the previous example

SOLUTION

$\pi_\eta = \frac{\partial \mathcal{L}}{\partial \dot{\eta}} = \rho \dot{\eta}$. Notice this last result agrees with the idea of momentum $p = mv$ by setting $m = \rho \Delta V$ and $p = \pi \Delta V$ where ρ represents mass per volume, and using $v = \dot{\eta}$.

4.7 LAGRANGIAN AND THE SCHRÖDINGER EQUATION

The quantum theory relies primarily on the Schrödinger wave equation to describe the dynamics of quantum particles. The present section shows how the Lagrangian formulation leads to the Schrödinger wave equation that treats particles as waves. The quantum theory will explore these concepts in more detail.

4.7.1 SCHRÖDINGER WAVE EQUATION

As a mathematical exercise, we start with a Lagrangian density

$$\mathcal{L} = i\hbar\psi^*\dot{\psi} - \frac{\hbar^2}{2m}\nabla\psi^* \cdot \nabla\psi - V(r)\psi^*\psi \quad (4.70a)$$

or equivalently

$$\mathcal{L} = i\hbar\psi^*\dot{\psi} - \frac{\hbar^2}{2m}\sum_j \partial_j\psi^* \partial_j\psi - V(r)\psi^*\psi \quad (4.70b)$$

where $j = x, y, z$ the Lagrangian is

$$L = \int d^3x \mathcal{L} \quad (4.70c)$$

The Lagrange density is a functional of the independent coordinates ψ, ψ^* and their derivatives $\partial_j \psi, \partial_j \psi^*$ where $j = x, y, z$.

The variation of L leads to the Euler–Lagrange equations of the form

$$\frac{\partial \mathcal{L}}{\partial \phi} - \sum_{\mu} \partial_{\mu} \frac{\partial \mathcal{L}}{\partial (\partial_{\mu} \phi)} = 0 \quad (4.71a)$$

where $\mu = x, y, z, t$ and $\phi = \psi, \psi^*$. Setting $\phi = \psi^*$ provides

$$\frac{\partial \mathcal{L}}{\partial \psi^*} - \sum_{\mu} \partial_{\mu} \frac{\partial \mathcal{L}}{\partial (\partial_{\mu} \psi^*)} = 0 \quad (4.71b)$$

Evaluating the first term in Equation 4.71a produces

$$\frac{\partial \mathcal{L}}{\partial \psi^*} = \frac{\partial}{\partial \psi^*} \left[i\hbar\psi^*\dot{\psi} - \frac{\hbar^2}{2m}\sum_j \partial_j\psi^* \partial_j\psi - V(r)\psi^*\psi \right] = i\hbar\dot{\psi} - V(r)\psi$$

The argument of the second term in Equation 4.71b produces

$$\frac{\partial \mathcal{L}}{\partial (\partial_{\mu} \psi^*)} = \frac{\partial}{\partial (\partial_{\mu} \psi^*)} \left\{ i\hbar\psi^* \partial_t \psi - \frac{\hbar^2}{2m} \sum_j \partial_j \psi^* \partial_j \psi - V(r)\psi^* \psi \right\} = \begin{cases} 0 & \mu = t \\ -\frac{\hbar^2}{2m} \partial_j \psi & \mu = j \end{cases}$$

The summation in Equation 4.71b becomes

$$i\hbar\dot{\psi} - V(r)\psi + \frac{\hbar^2}{2m} \sum_j \partial_j \partial_j \psi = 0$$

Therefore, we find the Schrödinger wave equation

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V(r)\psi = i\hbar\dot{\psi} \quad (4.72)$$

4.7.2 HAMILTONIAN DENSITY

We can find the classical Hamiltonian density (energy per unit volume)

$$\mathcal{H} = \pi\dot{\psi} - \mathcal{L} \quad (4.73a)$$

where π is the momentum conjugate to ψ and the total energy is

$$H = \int d^3x \mathcal{H} \quad (4.73b)$$

The conjugate momentum is defined by

$$\pi = \frac{\partial \mathcal{L}}{\partial \dot{\psi}} \quad (4.74)$$

For the Lagrange density in Equation 4.70, we find

$$\pi = \frac{\partial \mathcal{L}}{\partial \dot{\psi}} = \frac{\partial}{\partial \dot{\psi}} \left\{ i\hbar\psi^* \dot{\psi} - \frac{\hbar^2}{2m} \sum_j \partial_j \psi^* \partial_j \psi - V(r)\psi^* \psi \right\} = i\hbar\psi^*$$

The classical Hamiltonian density becomes

$$\mathcal{H} = \pi\dot{\psi} - \mathcal{L} = i\hbar\psi^* \dot{\psi} - \left\{ i\hbar\psi^* \dot{\psi} - \frac{\hbar^2}{2m} \nabla\psi^* \cdot \nabla\psi - V(r)\psi^* \psi \right\} = \frac{\hbar^2}{2m} \nabla\psi^* \cdot \nabla\psi + V(r)\psi^* \psi$$

Often times the Lagrange density is stated as

$$\mathcal{L} = i\hbar\psi^* \dot{\psi} + \frac{\hbar^2}{2m} \psi^* \nabla^2 \psi - V(r)\psi^* \psi = \psi^* \left(i\hbar\partial_t + \frac{\hbar^2}{2m} \nabla^2 - V \right) \psi \quad (4.75)$$

This last equation comes from Equation 4.70 by partially integrating and assuming the surface terms are zero. The Hamiltonian density then has the form

$$\mathcal{H} = \pi\dot{\psi} - \mathcal{L} = \psi^* \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \psi \quad (4.76)$$

The same results could have equally well been found by partially integrating Equation 4.73b using Equation 4.76 and taking the surface terms to be zero. In terms of the quantum theory, the classical Hamiltonian is most related to the average energy

$$H = \int d^3x \psi^* \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \psi = \langle \psi | H_{\text{sch}} | \psi \rangle \quad (4.77a)$$

where

$$H_{\text{sch}} = -\frac{\hbar^2}{2m} \nabla^2 + V \quad (4.77b)$$

4.8 BRIEF SUMMARY OF THE STRUCTURE OF SPACE-TIME

The theory of relativity is becoming increasingly important in a number of areas of engineering such as for the operation of the free electron laser. More importantly it sets limits for modern technology in terms of signal propagation speed. The structure of space-time represents one of the most fundamental notions in calculating the behavior of systems. For the most part, the theory of relativity will not be used in this text. However, significant amounts of the notation will be found throughout. For these reasons, we include a brief section.

The theory of relativity grew from the failure of experiments to detect an ether, which was postulated to be a deformable medium permeating all space for the sole purpose of sustaining light wave propagation. Einstein formulated several postulates. One required the speed of light to be a constant independent of the speed of the observer. The first section shows basic reasoning that allows us to conclude “space must warp” and that the universe must be made of a single entity, space-time. Rotations of 4-vectors mix time and space as well as energy and mass. The well-known relation $E = mc^2$ represents for length of a 4-vector (in the rest frame).

The introduction given here must be kept short. The name “special theory of relativity” suggests the postulates remain in the domain of unverified theory. Nothing could be further from reality. The implications of special relativity have been verified repeatedly for over 100 years. One might think that the “general theory of relativity,” normally associated with gravitation and black holes, would not have any application to solid state. Well, the clock rate on GPS satellites must incorporate corrections factors due to their position in the gravitational field; these corrections come from the general relativity. There exist many excellent texts on the subject of special relativity. Two of my favorites, somewhat older than others, but good introductions (1) *Space-Time Physics* by E.F. Taylor and J.A. Wheeler and (2) *Relativity* by A. Einstein. The first text should be rated “do not miss” for its clarity of basic concepts. R.A. Mould also has a very good book titled *Basic Relativity*.

4.8.1 INTRODUCTION TO SPACE-TIME WARPING

Let us consider two observers in uniform relative motion. Each observer has spatial dimensions x, y, z and time t . Assume observer O moves past observer O' along is $-z'$ -axis and that the two origins overlap at $t = t' = 0$.

The “space-time warp” can be discovered as follows. Observer O' sends out a pulse of light (at $t' = 0$) along the x' -axis as shown in Figure 4.13. At time t' the light is absorbed at point x' . Observer O sees the light pulse absorbed at (x, z) at time t . According to O', the light travels a distance $x' = ct'$. According to O, the light travels the distance $r = \sqrt{x^2 + z^2} = ct$ where $z = vt$. The special theory of relativity postulates that the speed of light must be the same in each case. The two observers can verify that $x = x'$. We can combine the equations to find the time-dilation effect.

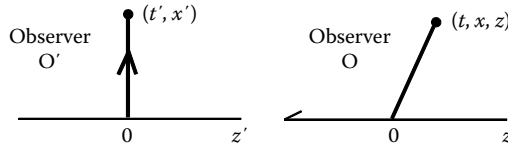


FIGURE 4.13 A light beam and two observers in relative motion.

$$x'^2 = c^2 t'^2 \quad \text{and} \quad x^2 + z^2 = c^2 t^2 \quad \rightarrow \quad x^2 - v^2 t^2 = c^2 t^2$$

Using the fact that $x' = x$ provides

$$t = \frac{t'}{\sqrt{1 - \left(\frac{v}{c}\right)^2}}$$

Clocks run slow when in uniform motion with respect to the observer. We can similarly consider a light beam traveling along the z -direction to find the length-contraction formula

$$x = x' \sqrt{1 - \left(\frac{v}{c}\right)^2}$$

As an important point, the time-dilation and length-contraction formulas actually apply to time and length intervals. Therefore the last two equations should be more correctly written as

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} \quad \Delta x = \Delta x' \sqrt{1 - \left(\frac{v}{c}\right)^2}$$

It should be clear, that in order to keep the speed of light constant, the time and space intervals must behave in a somewhat nonintuitive fashion. We will see from the Lorentz transformation, space and time become intermixed. The Lorentz transformation relates the coordinates (space and time) in one reference frame to the coordinates in another one. The coordinates (x, y, z, t) form a 4-vector. In fact, the Lorentz transformations relate any relativistic 4-vector in one frame to that in another (not just the coordinates). Writing formulas in correct 4-vector notation constitutes covariant notation. We will see that another 4-vector consists of energy and momentum. It gives rise to the famous $E = mc^2$ formula.

The next section discusses the Lorentz transformation. We state results. Notice how the transformation intermixes time and space. Also notice how, rather than using time and space intervals, it focuses on the coordinates themselves. One of the main results centers on the length of the 4-vectors. We cannot use the ordinary sum-of-the-squares type formula. However, once the length is defined, we will see that it is invariant under Lorentz transformations. This is very similar to rotations in Euclidean space that leave the length of a vector unaltered.

4.8.2 MINKOWSKI SPACE

Minkowski space provides the basic mathematical construct for space-time (special theory of relativity) as discussed in Section 2.10. It consists of a set of 4-vectors and a psuedo-inner product. The psuedo-inner product (metric) is somewhat different than the ordinary Euclidean one. However, the inner product directly relates to the manner in which the coordinates transform. Recall from Chapter 2, the definition of the inner product gives rise to the fact that unitary transformations exist which do not alter the length of a vector. The unitary transformations are normally viewed as rotations in Hilbert space. The “inner product” for Minkowski space does not fully satisfy the

properties of an inner product. For example, if the “inner product” of a 4-vector with itself produces zero, we do not necessarily have the vector itself being zero. In addition, the length of a vector is not necessarily positive definite.

In relativity, we are interested in Lorentz transformations because they relate physical quantities in one coordinate system to another in uniform motion with respect to the first. So if we know the position and time of an event in one reference frame, we can find the position and time of that event in any other reference frame. Or if we know energy and momentum in one, we can find the energy and momentum in another. It just so happens, that these Lorentz transformations appear as rotations in Minkowski space.

As just mentioned, the 4-vectors transform according to the Lorentz transformation, which allows us to calculate the 4-vector in any moving reference frame if we know it in at least one other. In many cases, the easiest procedure consists of calculating the interesting quantities in a “rest frame” and then applying the Lorentz transformation to find the corresponding quantities in the moving reference frame. As previously mentioned, the Lorentz transformation originates in the experimental fact that the speed of light must be a constant regardless of the state of motion of the observer. There does not exist a “stationary” coordinate system in the universe and therefore we cannot define a naturally preferred reference frame. The mathematical formulation of valid physical laws must be independent of any particular reference frame; this means that the equations must be applicable to any coordinate system regardless of its state of uniform motion. Mathematical expressions valid for all reference frames are termed “relativistically covariant”; i.e., they retain their form under a Lorentz transformation.

Now we investigate the 4-vectors found in Minkowski space. The first version uses complex numbers for later convenience with the Lorentz transformations. We now list some examples of Minkowski space. We can have a Minkowski space with space-time 4-vectors or another Minkowski space of energy–momentum 4-vectors. All use the same pseudo-inner product. The list of 4-vectors include (where $i = \sqrt{-1}$)

Position–time	$x_\mu = (x_1, x_2, x_3, x_4) = (\vec{x}, ix_0) = (\vec{x}, ict)$
Four-gradient	$\partial_\mu = \frac{\partial}{\partial x_\mu} = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_4} \right) = \left(\nabla, \frac{1}{ic} \frac{\partial}{\partial t} \right)$
Momentum–energy	$p_\mu = (c\vec{p}, iE)$

(where E denotes the total energy and not just the energy in the rest mass)

Vector–scalar potential	$A_\mu = (\vec{A}, iA_0) = (\vec{A}, i\Phi)$
Current–charge density	$J_\mu = (\vec{J}, ic\rho)$

In particular, notice the order of the components and the imaginary number. Later, we will show another convention that eliminates the imaginary number and changes the order.

The psuedo-inner product is defined by appending the imaginary number i to one of the components of the coordinates (see Section 2.10). Strictly speaking, this modifies the coordinates to make it possible to use a Euclidean inner product. Let $A_\mu = (a_1, a_2, a_3, ia_4)$ and $B_\mu = (b_1, b_2, b_3, ib_4)$ be two 4-vectors in a Minkowski space, where the components a_μ, b_μ are all real.

$$\vec{A} \cdot \vec{B} = \sum_{\mu=1}^4 A_\mu B_\mu = a_1 b_1 + a_2 b_2 + a_3 b_3 - a_4 b_4$$

To actually define the pseudo-inner product, one does not append the “ i ” to the coordinates but instead, directly adopts the dot product shown in the previous equation.

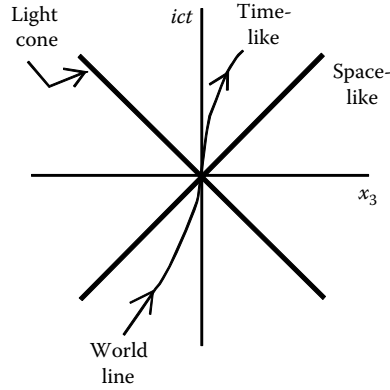


FIGURE 4.14 The light cone divides space into three regions.

Many times, we use the Einstein convention for repeated indices in a product to mean

$$A_{\mu}B_{\mu} \equiv \sum_{\mu=1}^4 A_{\mu}B_{\mu}$$

For example, using $x_{\mu} = (x_1, x_2, x_3, x_4) = (\vec{x}, ix_0) = (\vec{x}, ict)$ we find

$$x_{\mu}x_{\mu} = \vec{x} \cdot \vec{x} - c^2t^2$$

where the calculation of $\vec{x} \cdot \vec{x}$ proceeds as the usual inner product between Euclidean vectors. However, the previous equation is not the same as the typical Euclidean inner product because of the ‘minus’ sign (see Section 2.10).

Basically, the pseudo-inner product divides space-time into three regions (Figure 4.14) bounded by a “light cone.” The three regions determine whether the origin can be connected to other points by a signal not exceeding the speed of light. A 4-vector $x_{\mu} = (\vec{r}, ict)$ is time-like if $r^2 < c^2t^2$, space-like if $r^2 > c^2t^2$ and light-like if $r^2 = c^2t^2$. A world line is created by a particle as it moves through space-time. A differential element of length along the world line can be found from Pythagoras relation

$$(dL)^2 = \sum_{\mu} (dx_{\mu})^2 = (d\vec{x}) \cdot (d\vec{x}) - c^2(dt)^2$$

which is independent of coordinate system. The differential “proper time” $d\tau$ is defined to be the differential length of the position 4-vector as measured in the reference frame at rest with the particle (i.e., the reference frame traveling with the particle). In this case, $dx = 0$ and so $dL = ic d\tau$. The time interval $d\tau$ is measured by a clock at rest with the moving particle. Using the fact that the length of the 4-vector is invariant under a Lorentz transformation (the length is a scalar), the differential interval $dL = ic d\tau$ in any reference frame has the value

$$(d\tau)^2 = -\frac{1}{c^2} \sum_{\mu} (dx'_{\mu})(dx'_{\mu})$$

which leads to the usual time-dilation formula. The four-velocity is defined to be

$$dv_{\mu} = \frac{dx_{\mu}}{d\tau}$$

which is a valid 4-vector since $d\tau$ is a scalar.

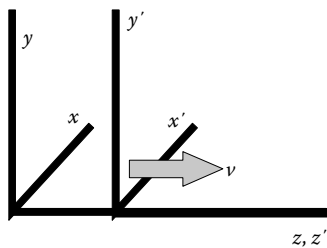


FIGURE 4.15 Prime system moves along the positive z-axis.

4.8.3 LORENTZ TRANSFORMATION

If the components of a 4-vector are known in one reference frame (i.e., space-time coordinate system) then they are known in any other by using the Lorentz transformation. As shown in Figure 4.15, we assume that the motion between two reference frames is along the $z = x_3$ axis. In particular, the primed system moves along the positive z -direction with speed v . In order to calculate physical quantities, we do not especially try to picture the situation using old Galilean intuition, but instead picture mathematical rotations in Minkowski space.

First consider rotations in Euclidean space (basically, the complex i changes Euclidean space into Minkowski space). Figure 4.16 shows a rotation of a 2-D vector $\vec{r}$ by an angle θ which is equivalent to rotating the reference frame by $-\theta$. The rotated vector is related to the original one by the operator $\hat{R}$

$$|r'\rangle = \hat{R}|r\rangle \tag{4.78}$$

which has the matrix

$$\underline{R} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \tag{4.79}$$

The Lorentz transformation rotates the x_3 and x_4 components for motion along the z -direction according to

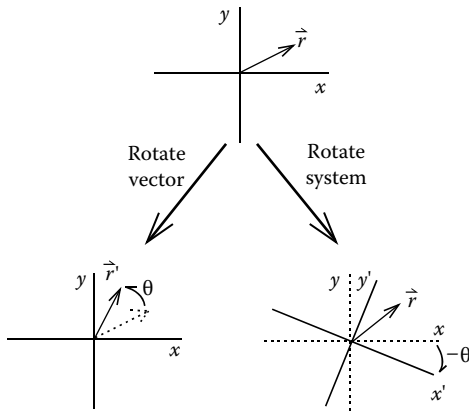


FIGURE 4.16 Rotate either the vector or the coordinate system.

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos \theta & -\sin \theta \\ 0 & 0 & \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad (4.80)$$

where the other components x_1 and x_2 are unaffected by motion along the z -direction. The same transformation holds for all of the different types of 4-vectors. The transformation equation can be written in terms of typical parameters using the following definitions

$$\theta = i\alpha \quad \beta = \frac{v}{c} \quad \gamma = \frac{1}{\sqrt{1-\beta^2}} \quad \beta = \tanh(\alpha) \quad (4.81)$$

where “tanh” is the hyperbolic tangent and the last relation leads to

$$\cosh(\alpha) = \frac{1}{\sqrt{1-\beta^2}} = \gamma \quad \sinh(\alpha) = \frac{\beta}{\sqrt{1-\beta^2}} = \beta\gamma \quad (4.82)$$

We have

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ ict' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos \theta & -\sin \theta \\ 0 & 0 & \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ ict \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cosh(\alpha) & i \sinh(\alpha) \\ 0 & 0 & -i \sinh(\alpha) & \cosh(\alpha) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ ict \end{pmatrix} \quad (4.83)$$

or, more simply

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ ict' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & i\beta\gamma \\ 0 & 0 & -i\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ ict \end{pmatrix} \quad (4.84)$$

The discussion above shows that 4-vectors transform according to

$$x'_\mu = R_{\mu\nu} x_\nu \quad (4.85)$$

so that the components in one reference frame can be related to the components in a second one in uniform motion (along z) with respect to the first. The transformation matrix is

$$R_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & i\beta\gamma \\ 0 & 0 & -i\beta\gamma & \gamma \end{pmatrix} \quad \mu, \nu = 1, 2, 3, 4$$

In reference to Equation 4.80, rotations in Minkowski space are orthogonal in the sense that $\underline{R}^{-1} = \underline{R}^T$ and the length of a 4-vector

$$x_\mu x_\mu = \sum_{\mu=1}^4 x_\mu x_\mu = \vec{x} \cdot \vec{x} - c^2 t^2$$

is left invariant under the transformation. Note the convention of an implied sum over repeated indices. The invariance is easy to see using matrix notation

$$x'_\mu x'_\mu \equiv \underline{x'}^T \underline{x'} = (\underline{Rx})^T (\underline{Rx}) = \underline{x} \underline{R}^T \underline{Rx} = \underline{x} \underline{x} = x_\mu x_\mu$$

The length of a 4-vector is therefore a scalar under the Lorentz transformation.

As a note, tensors $F_{\mu\nu}$ transform according to

$$F_{\alpha\beta} = R_{\alpha\mu} R_{\beta\nu} F_{\mu\nu}$$

where repeated indices are summed. Once the components of the tensor $F_{\mu\nu}$ are known in one reference frame, they are known in all others in uniform motion with respect to the first. One especially nice example concerns the electromagnetic field. We can show that a magnetic field is really an electric field in motion! That is, if we have an electric field due to a stationary point charge in one frame, then in a second frame in uniform motion, an observer will see both electric and magnetic fields! The motion between the two frames has converted a portion of the electric field into a magnetic field.

4.8.4 SOME EXAMPLES

As a first example, let us demonstrate the time-dilation formula using the Lorentz transformation equations. Suppose a clock is situated at the origin of the unprimed reference system. Find the time in the primed system. Using Equation 4.80

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ ict' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & i\beta\gamma \\ 0 & 0 & -i\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ ict \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & i\beta\gamma \\ 0 & 0 & -i\beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ ict \end{pmatrix} \quad (4.86)$$

We find

$$t' = \gamma t = \frac{t}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} \quad (4.87)$$

As a second example, the length of a 4-vector provides the momentum–energy relation. Starting with $p_\mu = (c\vec{p}, iE)$, we find

$$p_\mu p_\mu = c^2 \vec{p}^2 - E^2 \quad (4.88)$$

However, in a reference frame at rest with respect to the particle, we have $\vec{p} = 0$ and only the rest mass contributes to the total energy of the particle using $E = mc^2$. The length of the energy–momentum vector is invariant under Lorentz transformations. Therefore, the length of the energy–momentum 4-vector in any reference frame is given by

$$p'_\mu p'_\mu = p_\mu p_\mu = c^2 \vec{p}^2 - E^2 = -(mc^2)^2 \quad (4.89)$$

where m is the rest mass. Substituting for the 4-momentum p'_μ we find

$$E' = \sqrt{c^2 \vec{p}'^2 + (mc^2)^2} \rightarrow E = \sqrt{c^2 \vec{p}^2 + (mc^2)^2} \quad (4.90)$$

where we drop the prime notation since the formula must be correct in any reference frame regardless of its state of uniform motion. Using the Lorentz transformation, it is possible to show that the momentum in this last equation has the form $\vec{p} = \gamma m \vec{v}$. Equation 4.90 shows that the total energy comes from a momentum-related term ($\sim$ kinetic energy) and a rest mass term (the energy equivalent of the mass of the particle—the famous $E = mc^2$ term). For small momentum, we can make a Taylor expansion of Equation 4.90 to find

$$E = \sqrt{c^2 \vec{p}^2 + (mc^2)^2} = mc^2 \sqrt{1 + \frac{c^2 \vec{p}^2}{(mc^2)^2}} \cong mc^2 + \frac{1}{2} mv^2 + \dots \quad (4.91)$$

Section 2.10 shows the alternate notation using tensors and the metric.

4.9 REVIEW EXERCISES

- 4.1** Consider a solid rigid mass M rotating at angular speed $\dot{\theta}$ (in radians) about an origin fixed in space. Show the kinetic energy can be written as $T = I\dot{\theta}^2/2$ where $I = \int dm r^2$ and r is the distance to the mass dm from the origin. Start with $dT = (dm)v^2/2$.
- 4.2** Consider a system of N noninteracting point particles. Particle $\#i$ has mass m_i and vector $\vec{r}_i$ pointing from the origin to the particle. The center of mass can be written as $\vec{R} = \sum_{i=1}^N \vec{r}_i \frac{m_i}{M}$ where $M = \sum_{i=1}^N m_i$. Show the momentum of the system of N particles can be written as $\vec{P} = M\dot{\vec{R}}$. Further show the total externally applied force $F = \sum_{i=1}^N F_i$ accelerates the center of mass according to $\vec{F} = M\ddot{\vec{R}}$.
- 4.3** The torque is defined by $\vec{\tau} = \sum_{i=1}^N \vec{r}_i \times \vec{F}_i$ where Problem 4.2 defines the symbols. The angular momentum is defined by $\vec{L} = \sum_{i=1}^N \vec{r}_i \times \vec{p}_i$. Show $\vec{\tau} = \frac{d\vec{L}}{dt}$.
- 4.4** Define the center of mass as in Problem 4.2. Suppose the vector $\vec{r}_i$ represents the position of mass m_i in some arbitrary but fixed coordinate system while $\vec{r}'_i$ represents the position with respect to the origin of the center of mass system (i.e., place a coordinate system at point $\vec{R}$ defined in Problem 4.2). The vectors can be related by the relation $\vec{r}_i = \vec{R} + \vec{r}'_i$. Show the total angular momentum can be written as

$$\vec{L} = \vec{R} \times \vec{P} + \sum_{i=1}^N \vec{r}'_i \times \vec{p}'_i$$

where $\vec{p}'_i$ is the momentum with respect to the center of mass coordinates.

- 4.5** Use the definitions in the previous problems to show the kinetic energy T of a solid body can be expressed as the sum of the kinetic energy of the center of mass and the motion about the center of mass.

$$T = \frac{1}{2} M \dot{\vec{R}}^2 + \frac{1}{2} I \dot{\theta}^2$$

- 4.6** Assume the pulley has mass M and radius R and that it supports two masses as in Figure P4.6. Use the results of Problem 4.1.
- Find moment of inertia I for the pulley with uniform mass distribution.
 - Write the total kinetic and potential energy for the system in terms of θ and $\dot{\theta}$.
 - Use the Lagrangian to find the equation of motion and solve it.
 - Find the momentum conjugate to θ .

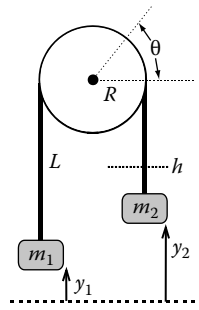


FIGURE P4.6 Pulley system.

- 4.7 Find the Equations of motion for the pulley system in Figure P4.6 for the case of a stretchable string with spring constant k . Assume the equilibrium length of the string is L (without masses attached), the string can be both compressed and stretched (obeys Hook's law), and the pulley is massless. Further assume that (without masses) $y_2 = h$ when $y_1 = 0$. Decouple and solve the equations of motion by using the new coordinates $y_+ = y_1 + y_2$ and $y_- = y_1 - y_2$.
- 4.8 Consider a cylinder of mass M , length L , and radius R constrained to roll down a plane as shown in Figure P4.8. Find the equation of motion and solve it.

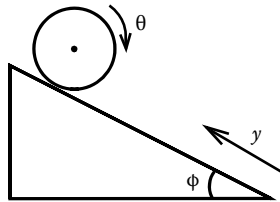


FIGURE P4.8 A cylinder rolling down the plane.

- 4.9 Consider a mass m connected to a spring with spring constant k . Assume the equilibrium position of the mass is at $x = 0$.
- Write the Hamiltonian for the system.
 - Use Hamilton's canonical equations to find an expression for $\dot{x}$ and $\dot{p}$.
 - Use the results of part b to write an equation for position x alone and solve it.
- 4.10 Find the Hamiltonian for Problem 4.7 and then write expressions for $\dot{y}_1$, $\dot{y}_2$, $\dot{p}_1$, $\dot{p}_2$. You can start from the basic definition $H = \sum_i p_i \dot{q}_i - L$.
- 4.11 Find the Hamiltonian for Problem 4.8 and then use Hamilton's canonical relations.
- 4.12 Use the Poisson brackets to demonstrate the following relations

$$[A, A] = 0 \quad [A, B] = -[B, A] \quad [A, c] = 0$$

$$[A + B, C] = [A, C] + [B, C] \quad [A, BC] = [A, B]C + B[A, C]$$

- 4.13 Use the Poisson brackets to show

$$[q_i, q_j] = 0 \quad [p_i, p_j] = 0 \quad [q_i, p_i] = \delta_{ij}$$

where p_j is the momentum conjugate to q_j .

4.14 In the section covering normal coordinates, a Lagrangian was defined by

$$L = T - V = \frac{1}{2} \sum_{i,j} [T_{ij} \dot{u}_i \dot{u}_j - V_{ij} u_i u_j]$$

a. Show the coordinate transformation

$$u_i = \sum_j a_{ij} v_j$$

produces the following two Lagrangians

$$L = \frac{1}{2} \sum_i [\dot{v}_i^2 - \lambda_i v_i^2] \quad \text{and} \quad L = \frac{1}{2} \{ \dot{\mathbf{v}}^T \dot{\mathbf{v}} - \mathbf{v}^T \underline{\lambda} \mathbf{v} \}$$

The matrix $\underline{\lambda}$ has all zero elements except those along the diagonal that have the value λ_i ; that is, the matrix has the elements $\underline{\lambda}_{ij} = \lambda_i \delta_{ij}$.

b. Show that the original Lagrangian in Equation 4.69 produces the equation of motion

$$\sum_j (T_{ij} \ddot{u}_j + V_{ij} u_j) = 0 \quad \text{or} \quad \underline{T} \ddot{\mathbf{u}} + \underline{V} \mathbf{u} = 0$$

which assumes that T_{ij} and V_{ij} are symmetric.

4.14 Suppose an electromagnetic field interacts with charged particle at $\vec{r}_i = \tilde{x}x_i + \tilde{y}y_i + \tilde{z}z_i$ through the vector potential $\vec{A}(\vec{r}_i)$ and electrostatic potential $\phi(\vec{r}_i)$, where $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ represent unit vectors. The Lagrangian has the form

$$L = \sum_i \left\{ \frac{1}{2} m_i r_i^2 - q_i \phi(\vec{r}_i) + \frac{q_i}{c} \vec{A}(\vec{r}_i) \cdot \vec{r}_i \right\}$$

Find the canonical momentum p_{ix} . Explain why two terms appear in the result and what they physically mean.

4.15 Explain why a the following relation must hold for δx_i independent

$$\sum_{i=1}^N f(x_i) \delta x_i = 0 \rightarrow f(x_i) = 0$$

This is similar to a step in the procedure to derive Lagrange's equation. *Hint:* Consider a matrix solution. Keep in mind that δx_1 , for example, can have any number of values such as 0.1, 0.001, etc.

4.16 Assume periodic boundary conditions. Show how

$$0 = \delta I = \int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt \, d^3x \left[\frac{\partial \mathcal{L}}{\partial \eta} \delta \eta + \frac{\partial \mathcal{L}}{\partial \dot{\eta}} \frac{\partial}{\partial t} \delta \eta + \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} \partial_i \delta \eta \right]$$

leads to

$$\int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt \, d^3x \left[\frac{\partial \mathcal{L}}{\partial \eta} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} - \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} \right] \delta \eta = 0$$

Explain and show any necessary conditions of the limits of the spatial integral. Remark, according to the Einstein summation convention, repeated indices must be summed $i = 1, 2, 3$.

- 4.17** Suppose the Lagrange density has the form $\mathcal{L} = \frac{\rho}{2} \dot{\eta}^2 + \frac{\beta}{2} [(\partial_x \eta)^2 + (\partial_y \eta)^2]$ for 1-D motion, where ρ, β resemble the mass density and spring constant (Young's modulus) for the material, and $\eta = \eta(x, y, t)$. Find the equation of motion for η .
- 4.18** If $\mathcal{L} = \frac{\rho}{2} \dot{\eta}^2 + \frac{\beta}{2} (\nabla \eta)^2$ where $(\nabla \eta)^2 = \nabla \eta \cdot \nabla \eta$ and $\eta = \eta(x, y, z)$ then find the equation of motion for η .
- 4.19** Starting with $\mathcal{L} = i\hbar \psi^* \dot{\psi} - \frac{\hbar^2}{2m} \nabla \psi^* \cdot \nabla \psi - V(r) \psi^* \psi$, show the alternate form of the Lagrange density by partial integration.

$$\mathcal{L} = i\hbar \psi^* \dot{\psi} + \frac{\hbar^2}{2m} \psi^* \nabla^2 \psi - V(r) \psi^* \psi = \psi^* \left(i\hbar \partial_t + \frac{\hbar^2}{2m} \nabla^2 - V \right) \psi$$

4.20 Show Hamiltonian

$$\mathcal{H} = \pi \dot{\psi} - \mathcal{L} = \psi^* \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \psi$$

based on the Lagrange density

$$\mathcal{L} = \psi^* \left(i\hbar \partial_t + \frac{\hbar^2}{2m} \nabla^2 - V \right) \psi$$

4.21 In Section 4.6, two equations have the form

$$\begin{aligned} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} &= b_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{i\omega_1 t} + b_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-i\omega_1 t} + b_3 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{i\omega_2 t} + b_4 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-i\omega_2 t} \\ \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} &= c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sin(\omega_1 t + \phi_1) + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} \sin(\omega_2 t + \phi_2) \end{aligned}$$

4.21a Show that for b_i real, we must have the relations $b_1 = b_2^*, b_3 = b_4^*$

4.21b Show that the angles ϕ must be given by

$$\tan \phi_1 = \frac{b_1 + b_2}{i(b_1 - b_2)}$$

and show that the denominator must be real.

4.22 Starting with

$$\begin{aligned} m\ddot{u}_1 + (\beta + \beta_{12})u_1 - \beta_{12}u_2 &= 0 \\ m\ddot{u}_2 + (\beta + \beta_{12})u_2 - \beta_{12}u_1 &= 0 \end{aligned}$$

in Section 4.6, show the results in Table 4.2

$$v_1(t) = 0 \rightarrow u_1(t) = -u_2(t) \quad \text{and that} \quad \omega_1 = \sqrt{\frac{\beta}{m}}$$

and

$$v_2(t) = 0 \rightarrow u_1(t) = u_2(t) \quad \text{and that} \quad \omega_2 = \sqrt{\frac{\beta + 2\beta_{12}}{m}}$$

4.23 Consider three-square matrices (same number of rows and columns) defined by

$$\underline{\lambda} = \lambda_i \delta_{ij} \quad \underline{T} \quad \underline{A} = [\underline{a}^{(1)}, \underline{a}^{(2)}, \dots] = \left[\begin{pmatrix} a_{j=1}^{(i=1)} \\ a_2^{(1)} \\ \vdots \end{pmatrix} \begin{pmatrix} a_{j=1}^{(i=2)} \\ a_2^{(2)} \\ \vdots \end{pmatrix} \dots \right] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & \vdots \\ a_{31} & \dots & \end{bmatrix}$$

Show that a matrix formed from the collection of columns $\{\lambda_1 \underline{T} \underline{a}^{(1)}, \lambda_2 \underline{T} \underline{a}^{(2)}, \dots\}$ must be one in the same matrix as given by $\underline{\lambda} \underline{T} \underline{A}$. It might be easiest to write the product of matrices as sums of the product of the matrix elements.

REFERENCES AND FURTHER READINGS

Mechanics, normal modes, Lagrangians for continuous media

1. Marion J.B., *Classical Dynamics*, Academic Press, New York (1970).
2. Goldstein R., *Classical Mechanics*, Addison-Wesley, Reading, MA (1950).

Feynman path integrals, quantum mechanical Lagrangians

3. Feynman R.P., *QED: The Strange Theory of Light and Matter*, Princeton University Press, Princeton, NJ (1985). An excellent and highly recommended easy-to-read book.
4. Brown L.S., *Quantum Field Theory*, Cambridge University Press, Cambridge, U.K. (1996).

Relativity

5. Taylor E.F. and Wheeler J.A., *Spacetime Physics: Introduction to Special Relativity*, 2nd ed., W.F. Freeman and Company, New York (1992). Easy to read.
6. Mould R.A., *Basic Relativity*, Springer-Verlag, New York (1994).
7. Einstein A., *Relativity: The Special and the General Theory, A Popular Exposition*, Crown Publishers, Inc., New York (1961).
8. Das A., *The Special Theory of Relativity: A Mathematical Exposition*, Springer-Verlag, New York (1993).
9. Wald R.M., *General Relativity*, The University of Chicago Press, Chicago, IL (1984).
10. Misner C.W., Thorne, K.S., and Wheeler J.A., *Gravitation*, W.H. Freeman & Company, San Francisco, CA (1973). One giant book.

Other

11. Sakurai J.J., *Advanced Quantum Mechanics*, Addison-Wesley Publishing Co., Reading, MA (1980).
12. Bjorken J.D. and Drell S.D., *Relativistic Quantum Mechanics*, McGraw-Hill Book Company, New York (1964).
13. Gelfand I.M. and Fomin S.V., *Calculus of Variations*, Dover Publications, Mineola, NY (1991).
14. Kane G., *Modern Elementary Particle Physics*, Addison-Wesley Publishing Co., New York (1994).
15. Naber G.L., *The Geometry of Minkowski Spacetime: An Introduction to the Mathematics of the Special Theory of Relativity*, Dover Publications, Mineola, NY (1992, 2003).

5 Quantum Mechanics

Quantum theory has formed a cornerstone for modern physics, engineering, and chemistry since the early 1900s. It has found significant modern applications in engineering since the development of the semiconductor diode, transistor, and especially the laser in the 1960s. Not until the 1980s did the fabrication and materials growth technology become sufficiently developed to produce quantum well devices (such as quantum well lasers) and to engineer the optical and electrical properties of materials (band gap engineering). One of the major purposes of this chapter is to introduce modern quantum theory in order to engineer new, superior components.

This chapter begins by developing the relation between the quantum theory and the linear algebra. It discusses the most fundamental postulates of the theory and provides a phenomenological development of the Schrödinger wave equation (SWE) although the results might be gleaned from the classical Lagrangian and Hamiltonian with the Poisson bracket relations. Afterward some simple examples for the infinitely and finitely deep wells help clarify the basic postulates. The simple harmonic oscillator has perhaps the most sweeping implications and applications for the quantum theory. For this reason, we present the most applicable formalism for the Harmonic oscillator that uses the operator approach rather than the classical method of partial differential equations. The angular momentum and spin are presented for the study of the atom and for applications for nanodevices, and quantum computing. Next the representation formalism clarifies the distinction between the dynamical operators and the wave functions and their interrelations. The chapter discusses both time-independent and time-dependent perturbation theories. The density operator combines the quantum theory with the classical theory for finding the behavior of a system. The density operator represents one of the basic concepts in the quantum theory. The remainder of the chapter introduces the more advanced material meant to introduce the quantum field theory starting with the second quantization and the propagator.

5.1 RELATION BETWEEN QUANTUM MECHANICS AND LINEAR ALGEBRA

Mathematical abstractions inherent to the linear algebra must be properly interpreted to accurately model the physical world. The theory must represent properties of particles and systems, predict the evolution of the system, and provide the ability to make and interpret observations. Quantum theory began in an effort to describe microscopic (atomic) systems when classical theory gave erred predictions. However, classical and quantum mechanical descriptions must agree for macroscopic systems which comprise the correspondence principle.

Vectors in a Hilbert space represent specific properties of a particle or system. Every physically possible state of the system must be represented by one of the vectors. A single particle corresponds to a single vector (possibly a time-dependent vector in a tensor product space). Hermitian operators represent physically observable quantities such as energy, momentum, and electric field. These operators provide values for the quantities when they act upon a vector in a Hilbert space. The discussion will show how the theory distinguishes measurement operators from Hermitian operators.

The Feynman path integral and principle of least action (through the Lagrangian) lead to the Schrödinger equation, which describes the system dynamics. The method essentially reduces to using a classical Hamiltonian and replacing the dynamical variables with operators. The operators must satisfy commutation relations somewhat similar to the Poisson brackets for classical mechanics.

TABLE 5.1
Physical World, Linear Algebra, and Quantum Theory

Physical World	Mathematics
Observables: Properties that can be measured in a laboratory	Hermitian operators $\hat{\mathcal{H}}$
Specific particle/system properties	Wave functions $ \psi\rangle$
Fundamental motions/states of existence	Basis/eigenvectors $ h\rangle$ of $\hat{\mathcal{H}}$
Value of observable in fundamental motion	$\hat{\mathcal{H}} h\rangle = h h\rangle$
Laboratory measured values, states	Sets $\{h\}$ and $\{ h\rangle\}$
Particle/system has characteristics of all fundamental motions	Superposed wave function $ \psi\rangle = \sum_h \beta_h h\rangle$
Average behavior of a particle	$\langle\psi \hat{\mathcal{H}} \psi\rangle$
Probability of finding value or fundamental motion	Probability amplitude of finding h or $ h\rangle$ is $\langle h \psi\rangle = \beta_h$. Probability = $ \beta_h ^2$
Dynamics of system	Time dependence of operators or vectors—Schrödinger’s equation
Measure state of particle/system	Collapse of $ \psi\rangle$ to basis vector $ h\rangle$. Random collapse does not have equation of motion
Simultaneous measurements of two or more observables	Commuting operators: repeated measurements produce identical values Noncommuting operators: repeated measurements produce a range of values
Complete description of a particle/system	Largest possible set of commuting Hermitian operators

We also need to address the issue of how the particle dynamics (equations of motion) arise. In the classical situation, dynamical variables such as position and momentum can depend on time. The Heisenberg representation in quantum theory gives the time dependence to the Hermitian operators which represent the dynamical variables. In this description, the operators “carry the dynamics of the system” while the wave functions remain independent of time. In this case, the vectors (i.e., wave functions) in Hilbert space appear as a type of “lattice” (or stage) for observation. The result of an observation depends on the time of making the observation through the operators. The Schrödinger representation of the quantum theory provides an interpretation most closely related to classical optics and electromagnetic theory. The wave functions depend on time but the operators do not. This is very similar to saying that the electric field (as the wave function) depends on time because the traveling wave, for example, has the form $e^{ikx - i\omega t}$. We will encounter an intermediate case, the interaction representation, where the operators carry a trivial time-dependence and the wave functions retain the time response to a “forcing function.” All three representations contain identical information.

In this section, we address the following issues listed in Table 5.1: (1) how basis vectors differ from other vectors; (2) the meaning of superposition; (3) the physical meaning of the expansion coefficients of a general vector in a Hilbert space; (4) a picture of the time-dependent wave function; (5) the collapse of the wave function; and (6) observables that cannot be “simultaneously observed” with unlimited precision.

5.1.1 OBSERVABLES AND HERMITIAN OPERATORS

Every physical system must be capable of interacting with the physical world. In the laboratory, the systems come under the scrutiny of other probing systems such as our own physical senses or the equipment in the laboratory. The results of these measurements must be real numbers and not the complex numbers often used for convenience. “Observables,” such as energy or momentum, are

quantities that can be observed and measured in the laboratory and take on only real values. These values can be samples from inherently discrete or continuous ranges. For example, confined electrons have discrete energy values whereas the position of an electron can be in a continuous range.

Suppose measurements of a particular property such as energy H of a system always produce the set of real values $\{E_1, E_2, \dots\}$ and the particle is always found in one of the corresponding states $\{|E_1\rangle, |E_2\rangle, \dots\}$. Based on these values and vectors, we define an energy operator (Hamiltonian $\hat{H}$)

$$\hat{H} = \sum_n E_n |E_n\rangle \langle E_n| \quad (5.1)$$

Applying the Hamiltonian to one of the states produces

$$\hat{H}|E_n\rangle = E_n|E_n\rangle \quad (5.2)$$

We naturally interpret the operation as measuring the value of $\hat{H}$ for a system in the state $|E_n\rangle$. Notice that the operator in Equation 5.1 must be Hermitian since $\hat{H}^+ = \hat{H}$. By assumption, the eigenvalues are real. The number of eigenvectors equals the number of possible states for the system so that each possible state can be represented by a mathematical object; the eigenvectors form a complete set. For these reasons, quantum theory represents observables by Hermitian operators.

The process of “making a measurement” cannot be fully modeled by the eigenvalue equation (Equation 5.2). The operators in the theory operate on vectors in a Hilbert space. A general vector can be written as a superposition of the eigenvectors of $\hat{H}$ and therefore do not have just a single value for the measurement of $\hat{H}$. A physical measurement of $\hat{H}$ causes the wave function to collapse to a random basis vector, which does not follow from the dynamics and does not appear in the effect of the Hermitian operator—more on this later.

5.1.2 EIGENSTATES

The eigenvectors of a Hermitian operator, which corresponds to an observable, are the most fundamental states for a particle or system. Every possible fundamental motion of a particle must be observable (i.e., measurable). This requires that each fundamental physical state of a system or particle must be represented as a basis vector. For example, the various orbitals in an atom correspond to energy eigenvectors since each orbital has a well-defined value for the energy. The basis set must be complete so that all fundamental motions can be detected and represented in the theory. As mentioned in Section 5.1.1, if measurements of particle energy $\hat{H}$ produce the values $\{E_1, E_2, \dots, E_n, \dots\}$ then we can represent the “observed” states by the eigenvectors $\{|E_1\rangle, |E_2\rangle, \dots, |E_n\rangle, \dots\}$ where $\hat{H}|E_n\rangle = E_n|E_n\rangle$. These states must be the most basic states; they form the basis states. Any other state of the system must be a linear combination of these basis states. A linear combination of the basis functions $\{|E_1\rangle, |E_2\rangle, |E_3\rangle, \dots\}$ produces an *average* energy that can differ from the energies $\{E_1, E_2, \dots, E_n, \dots\}$. The distinction between the basis states and the superposed states is quite fundamental to the theory. The particles can only be found in one of the basis states; however, prior to the measurement, they can exist in a superposition state. According to the Copenhagen interpretation, the measurement causes the system to transition from the superposed state to the basis state (sometimes called the “collapse of the wave function”).

The idea of “state” occurs in many branches of science and engineering. A particle or system can usually be described by a collection of parameters. We define a state of the particle or system to be a specific set of values for the parameters. For example, pressure, volume, and temperature specify the state of a gas. In the following, we describe the states found in other areas of study.

What are the states for classical mechanics? The position and momentum describe the motion of a point particle. Therefore, the three position and three momentum components completely specify the state of motion for a single point particle. There are three degrees of freedom.

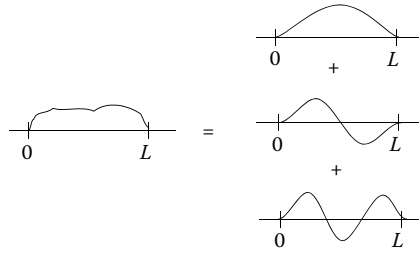


FIGURE 5.1 A classical wave on a string is decomposed into the basic modes (i.e., the basis vectors).

What are the states for classical wave motion on a string? Assume both ends of the string are securely fastened (Figure 5.1). The basis set consists of sine waves normalized to 1

$$\left\{ \langle x | \phi_n \rangle = \phi_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \text{ where } n = 1, 2, 3, \dots \right\}$$

These states (i.e., modes) can be indexed by the allowed wavelengths $\lambda = 2L/n$. The overall shape of the wave specifies the “mode” (and not the amplitude since that corresponds to adding energy to a given mode). A general state of the system consists of a sum over all of the allowed modes (Fourier analysis). A linear combination of the basis vectors defines a general state for the string; the classical wave can have arbitrary magnitude. The linear combination of basis vectors

$$|\psi(t)\rangle = \sum_n \beta_n(t) |\phi_n\rangle$$

gives a general wave function for the vibrating string. Notice that the basic modes (i.e., $|\phi_n\rangle$ or $\phi_n(x)$) do not depend on time. The time dependence of the vibrational motion appears in the expansion coefficients $\beta_n(t)$. The basis set consists of the eigenvectors for the time-independent wave equation. A given coefficient $\beta_n(t)$ provides a “weight” that describes how much of the wave function $|\psi(t)\rangle$ can be attributed to the basis function $|\phi_n\rangle$.

What are the fundamental “modes” in classical optics? The polarization, wavelength, and the propagation vector specify the basic modes. Notice that we do not include the amplitude in the list because we can add any number of photons to the mode (i.e., produce any amplitude we choose) without changing the basic shape. However, in quantum optics, the fundamental states include the photon number as part of the description of the basis states. That is, two basis states characterized by two different numbers of photons in the same mode (same wave vector and polarization) will be orthogonal in the Hilbert space. The optical modes are eigenvectors of the time-independent Maxwell wave equation. We expect that these basic modes will be sinusoidal for a Fabry–Perot cavity. They produce traveling plane waves for free space.

Example 5.1: Polarization in Optics

A single photon travels along the z-axis as shown in Figure 5.2. The photon has components of polarization along the x-axis and along the y-axis, for example, according to

$$\vec{s} = \frac{1}{\sqrt{2}} \vec{x} + \frac{1}{\sqrt{2}} \vec{y}$$

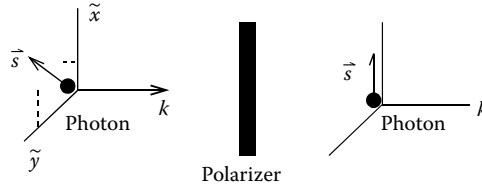


FIGURE 5.2 Polarization.

The electric field is parallel to the polarization $\vec{s}$. We view the single photon as simultaneously polarized along $\tilde{x}$ and along $\tilde{y}$. Suppose we place a polarizer in the path of the photon with its axis along the x -axis. There exists a 50% chance that the photon will be found polarized along the x -axis. The “polarization” state of the incident photon must be the superposition of two basis states $\tilde{x}, \tilde{y}$. We view the single incident photon as being “simultaneously in both polarization states.” The act of observing the photon causes the wave function to collapse to either the $\tilde{x}$ state or to the $\tilde{y}$ state. The polarizer absorbs/reflects the photon if the photon wave function collapses to the $\tilde{y}$ -polarization. The polarizer allows the photon to pass if the photon wave function collapses to the $\tilde{x}$ -polarization. For a single photon, either the photon will be transmitted or it will not; there cannot be any intermediate case.

5.1.3 MEANING OF SUPERPOSITION OF BASIS STATES AND THE PROBABILITY INTERPRETATION

A quantum particle can “occupy” a state

$$|v\rangle = \sum_n \beta_n(t) |\phi_n\rangle \quad (5.3)$$

where basis set $\{|\phi_n\rangle\}$ represents the collection of fundamental physical states. The most convenient basis set consists of the eigenvectors of an operator of special interest to us. For our discussion here, assume that we have most interest in the energy of the particle. We therefore choose the basis set to be the eigenvectors of the energy operator (i.e., the Hamiltonian $\hat{H}$). This means that we make measurements of the energy and therefore find a specific set of states $|\phi_n\rangle$ (such as might represent the atomic orbitals or energy levels in laser material) and the corresponding energy values E_n . The states and energy values satisfy the eigenvector equation

$$\hat{H}|\phi_n\rangle = E_n|\phi_n\rangle$$

The superposed wave function $|v\rangle$ refers to a particle (or system) having attributes from all of the states in the superposition. The particle simultaneously exists in all of the basic states making up the superposition. In Figure 5.3, for example, an observation of the energy of the particle in the state $|v\rangle$ with the energy basis set will find it with energy E_1 or E_2 or E_3 . Before the measurement,

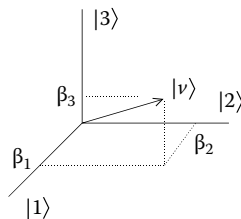


FIGURE 5.3 The vector is a linear combination of basis vectors.

one might view the particle as having some mixture of all three energies in a type of average. The measurement forces the electron to decide on the actual energy. One can easily calculate the average energy of the superposed state for Figure 5.3 (assuming $|v\rangle$ normalized to 1—more on this will be discussed later in this chapter)

$$\langle v|\hat{H}|v\rangle = \sum_n E_n |\beta_n|^2$$

which does not necessarily have the same value as found for the observed state of the particle such as E_1 . It would appear that energy conservation has been violated. However, the product $\langle v|\hat{H}|v\rangle$ corresponds to the classical value of energy and obtains for the single particle only after repeated measurements for the particle in the same state $|v\rangle$ or for many particles in the state $|v\rangle$. As a side comment, it is interesting that Newton's laws assume that a physical observable has an "actual value" and measurements produce an average value that can differ from the actual value only through errors in the measurement process. That is, by refining the measurement technique, one can make the measured average value come closer to the "actual value." Quantum mechanics essentially denies the existence of this type of "actual value." With this paradigm in mind, one realizes that all of the classical laws apply to average values while ignoring the physical reality of the standard deviation.

Not just any superposition wave function can be used for the quantum theory. All quantum mechanical wave functions must be normalized to have unit length including those constructed of a superposition of basis functions $\langle v|v\rangle = 1$ and not just the eigenvectors of a Hermitian operator that satisfy $\langle \phi_m|\phi_n\rangle = \delta_{mn}$. All of the vectors are normalized to one in order to interpret the components as a probability (next section). Therefore, the functions appropriate for the quantum theory define a surface for which all of its points are exactly 1 unit away from the origin. For the three-dimensional (3-D) case, the surface makes a unit sphere. The set of wave functions does not form a vector space since the zero vector cannot be in the set. The valid wave functions differ by their direction in Hilbert space. Once in a while, people do not normalize the wave functions, but then state that only the direction defines the state of the system; however, we will normalize in this book. The direction defines the properties of the system (or particle) through the expansion coefficients β_n in Equation 5.3.

5.1.4 PROBABILITY INTERPRETATION

Perhaps most important, the quantum theory interprets the expansion coefficients β_n in the superposition $|v\rangle = \sum_n \beta_n |n\rangle = \sum_n |n\rangle \langle n|v\rangle$ as a probability amplitude.

$$\text{Probability amplitude} = \beta_n = \langle n|v\rangle \quad (5.4)$$

To be more specific, assume we make a measurement of the energy of the particle. The quantized system allows the particle to occupy a discrete number of "fundamental" states $|\phi_1\rangle, |\phi_2\rangle, |\phi_3\rangle \dots$ with respective energies $E_1, E_2, \dots$. A measurement of the energy can only yield one of the numbers E_n and the particle must be found in one of the fundamental states $|\phi_n\rangle$. The probability that the particle is found in state $|n\rangle = |\phi_n\rangle$ is given by (also see Section 2.11)

$$P(n) = |\beta_n|^2 = |\langle n|v\rangle|^2 \quad (5.5)$$

Keep in mind that a probability function must satisfy certain conditions including

$$P(n) \geq 0 \quad \text{and} \quad \sum_n P(n) = 1 \quad (5.6)$$

Let us check that Equation 5.5 satisfies these two properties. It satisfies the first property since $|\beta_n|^2$ is nonnegative. The second property in Equation 5.6 holds since the vector $|\nu\rangle$ is normalized to one as seen as follows:

$$1 = \langle \nu | \nu \rangle = \sum_m \sum_n \beta_m^* \beta_n \langle \phi_m | \phi_n \rangle = \sum_n |\beta_n|^2 = \sum_n P(n) \quad (5.7)$$

So the normalization condition for the wave function requires the summation of all probabilities to equal unity. The usual theory of “Fourier series” interprets the expansion coefficients β_n in Equation 5.3 as weights which say how much of a certain basis vector (sine or cosine for example) makes up the overall wave function. Now, for quantum theory, the normalization of the wave functions suggests that we interpret the “weight” as a probability. Also notice that the sum always gives “one” in Equation 5.7 even though each individual β_n might change with time.

We can handle continuous coordinates in a similar fashion except use integrals and Dirac delta functions rather than the discrete summations and Kronecker delta functions. Projecting the wave function onto the spatial-coordinate basis set $\{|x\rangle\}$ also provides a probability amplitude. It refers to a probability that depends on position. Suppose a quantum particle occupies state $|\psi\rangle$ that can be expanded as

$$|\psi\rangle = \int dx |x\rangle \langle x | \psi \rangle = \int dx |x\rangle \psi(x)$$

The component of the vector gives the probability amplitude $\psi(x)$. These wave functions $\psi(x)$ usually come from the Schrödinger equation. The square of this probability amplitude $\langle x | \psi \rangle = \psi(x)$ gives the probability density $\rho(x) = \psi^*(x)\psi(x)$ (probability per unit length); it describes the probability of finding the particle at “point x ” (refer to Appendix D for a review of probability theory). We require that all quantum mechanically acceptable wave functions have unit length. For the continuous case, this normalization requirement leads to integrals over the probability density.

$$1 = \langle \psi | \psi \rangle = \langle \psi | \hat{1} | \psi \rangle = \int_{\text{all } x} dx \langle \psi | x \rangle \langle x | \psi \rangle = \int_{\text{all } x} dx \psi^*(x) \psi(x)$$

Therefore the density can be interpreted as a probability density. For three spatial dimensions, $\rho(\vec{r})dV = \psi^*(\vec{r})\psi(\vec{r})dV$ represents the probability of finding a particle in the infinitesimal volume dV centered at the position $\vec{r}$

$$\text{PROB}(a \leq x \leq b, c \leq y \leq d, e \leq z \leq f) = \int_a^b \int_c^d \int_e^f dx dy dz \rho(x, y, z) = \int_V dV \rho$$

Several types of reasoning on probability are quite common for the quantum theory. Unlike classical probability theory, we cannot simply add and multiply probabilities. In quantum theory, the probability amplitudes “add” and “multiply.” Consider a succession of events occurring at the space-time points $\{(x_0, t_0), (x_1, t_1), (x_2, t_2) \dots\}$ on the history path in Figure 5.4. The probability amplitude $\psi(x, t)$ of the succession of events all on the same history path consists of the product $\psi(x, t) = \prod_i \psi_i(x_i, t_i)$. Without superposition, the probability for successive events (the square of the amplitude) reduces to the product of the probabilities as found in classical probability theory. Superposition requires the phase of the amplitude to be taken into account similar to that for the electromagnetic field before calculating the total power.

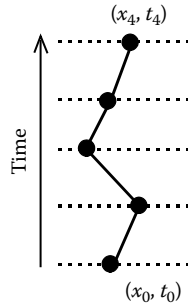


FIGURE 5.4 A succession of events on a single history path.

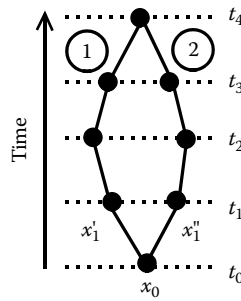


FIGURE 5.5 Parallel history paths.

For the case of two independent events such as two occurring at the same time, the probability amplitudes add (Figure 5.5) $\psi(x, t) = \psi_1(x'_1, t_1) + \psi_2(x''_2, t_1)$ where all wave functions depend on (x, t) at the destination point (really need a propagator).

A measurement of an observable $\hat{A}$ for $|\psi\rangle = \sum_n \beta_n |a_n\rangle$ produces exactly one of the eigenvalues $\{a_1, a_2, \dots\}$ and shows that the particle must be in one of the corresponding eigenstates $\{|a_1\rangle, |a_2\rangle, \dots\}$. The classical probability of finding the particle in state a_i or a_j can be written as

$$P(a_i \text{ or } a_j) = P(a_i) + P(a_j) - P(a_i \text{ and } a_j)$$

The two events are mutually exclusive in this case so that $P(a_i \text{ and } a_j) = 0$ and

$$P(a_i \text{ or } a_j) = |\beta_i|^2 + |\beta_j|^2$$

When people look for the results of measurements on a quantum system, even though there exists an infinite number of wave functions $|\psi\rangle$, they often consider only the basis states and eigenvalues.

5.1.5 AVERAGES

We use the quantum mechanical probability density in a slightly different manner than the classical ones. Consider a particle (or system) in state

$$|\psi\rangle = \sum_n \beta_n |a_n\rangle \quad (5.8)$$

where $\{a_1, a_2, \dots\}$ and $\{|a_1\rangle, |a_2\rangle, \dots\}$ are the eigenvalues and eigenvectors for the observable $\hat{A}$. The quantum mechanical average value of $\hat{A}$ can be written as $\langle\psi|\hat{A}|\psi\rangle$.

An average can be computed by projecting the wave function onto either the eigenvector basis set or the coordinate basis set. Consider the eigenvectors first. Using Equation 5.8 we find

$$\langle\psi|\hat{A}|\psi\rangle = \sum_n a_n |\beta_n|^2 \quad (5.9)$$

This expression agrees with the classical probability expression for averages $E(A) = \sum_n a_n P_n$ where $E(A) = \langle A \rangle = \bar{A}$ represents the expectation value of a random variable A , which is not an operator in the classical probability theory. For the quantum operator, the range of $\hat{A}$ can be viewed as the outcome space $\{a_1, a_2, \dots\}$. Next, projecting into coordinate space, the average can be written as

$$\langle\psi|\hat{A}|\psi\rangle = \langle\psi|\left\{\int dx |x\rangle\langle x|\right\}\hat{A}|\psi\rangle = \int dx \psi^*(x) \hat{A} \psi(x) \quad (5.10a)$$

Notice that we must maintain the order of operators and vectors. (also see Section 2.11 and Appendix D). As discussed later, the use of the coordinate projector means that the operator A is now written as a functional of x (such as derivative) rather than as an abstract vector operator.

We define the variance of a Hermitian operator by

$$\sigma^2 = E(\hat{O} - \langle\hat{O}\rangle)^2 = E(\hat{O}^2 - 2\hat{O}\langle\hat{O}\rangle + \langle\hat{O}\rangle^2) = E(\hat{O}^2) - \langle\hat{O}\rangle^2 = \langle\hat{O}^2\rangle - \langle\hat{O}\rangle^2 \quad (5.10b)$$

The standard deviation becomes

$$\sigma = \sqrt{\langle\hat{O}^2\rangle - \langle\hat{O}\rangle^2} \quad (5.10c)$$

Three comments need to be made. First, to compute the expectation value or the variance, the wave function must be known. The components of the wave function give the probability amplitude. This is equivalent to knowing the probability function in classical probability theory. Second, from an ensemble point of view, the expectation of an operator really gives the average of an observable when making multiple observations on the same state. The quantity $\langle\hat{O}\rangle \equiv \langle\psi|\hat{O}|\psi\rangle$ gives the average of the observable $\hat{O}$ in the single state $|\psi\rangle$. For example, consider $|a\rangle$ to be an eigenstate of $\hat{A}$. Repeated measurements of the operator $\hat{A}$ produce the average

$$\langle a|\hat{A}|a\rangle = \langle a|a|a\rangle = a\langle a|a\rangle = a$$

The variance is obviously zero.

Non-Hermitian operators do not necessarily have a unique definition for the variance. Consider a variance defined similar to a classical variance $\text{Var}(O) = \langle(O - \bar{O})^*(O - \bar{O})\rangle$. For simplicity, set $\bar{O} = 0$ so that $\text{Var}(O) = \langle O^* O \rangle$. Replacing O with $\hat{O}$ and O^* with $\hat{O}^+$ produces the three possibilities of $\langle\hat{O}^+ \hat{O}\rangle$, $\langle\hat{O} \hat{O}^+\rangle$, and $\langle\frac{1}{2}\hat{O}^+ \hat{O} + \frac{1}{2}\hat{O} \hat{O}^+\rangle$ out of an infinite number. The adjoint can be dropped for Hermitian operators and all possibilities reduce to the one listed Equation 5.10c.

Example 5.2: Find the Standard Deviation for the Operator $\hat{A}$ in an Eigenstate $|a\rangle$

We need $\langle\hat{A}^2\rangle$. We can calculate it as follows:

$$\langle\hat{A}^2\rangle = \langle a|\hat{A}^2|a\rangle = \langle a|\hat{A}\hat{A}|a\rangle = \langle a|a^2|a\rangle = a^2$$

The average can also be found

$$\langle a|\hat{A}|a\rangle = \langle a|a|a\rangle = a\langle a|a\rangle = a$$

Therefore the standard deviation must be

$$\sigma = \sqrt{\langle \hat{A}^2 \rangle - \langle \hat{A} \rangle^2} = \sqrt{a^2 - a^2} = 0$$

Example 5.3: The Infinitely Deep Square Well

Find the expectation value of the position x for an electron in state n where the basis functions are

$$\left\{ \phi_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \right\}$$

SOLUTION

$$\langle x \rangle = \langle n|x|n \rangle = \int_0^L dx \phi_n^* x \phi_n = \frac{2}{L} \int_0^L dx x \sin^2\left(\frac{n\pi x}{L}\right) = \frac{L}{2}$$

5.1.6 MOTION OF THE WAVE FUNCTION

The SWE, a partial differential equation, provides the dynamics of particle through the wave function. Section 5.2 shows that the Schrödinger equation has the form

$$\hat{\mathcal{H}}|\Psi\rangle = i\hbar \frac{\partial}{\partial t} |\Psi\rangle \quad (5.11)$$

Solving the Schrödinger equation by the method of orthonormal expansions provides the energy basis functions $\{|1\rangle = |\phi_1\rangle, |2\rangle = |\phi_2\rangle, \dots\}$. It also gives the time dependence of $|\Psi\rangle$ which appears in the coefficients β in the basis vector expansion

$$|\Psi(t)\rangle = \sum_n \beta_n(t) |n\rangle$$

The wave function $|\Psi\rangle$ moves in Hilbert space since the coefficients β_n depend on time. Notice that the wave function stays within the given Hilbert space and never moves out of it! This is a result of the fact that the eigenvectors form a complete set.

A formal solution to Equation 5.11 can be found when the Hamiltonian does not depend on time. We will see later that

$$|\Psi(t)\rangle = e^{\frac{\hat{\mathcal{H}}(t-t_0)}{i\hbar}} |\Psi(t_0)\rangle \quad (5.12)$$

where $|\Psi(t_0)\rangle$ is the initial wave function. The operator

$$\hat{u}(t, t_0) = e^{\frac{\hat{\mathcal{H}}(t-t_0)}{i\hbar}} \quad (5.13)$$

moves the wave function $|\psi\rangle = |\psi(t)\rangle$ in time according to

$$|\psi(t)\rangle = \hat{u}(t - t_0) |\psi(t_0)\rangle$$

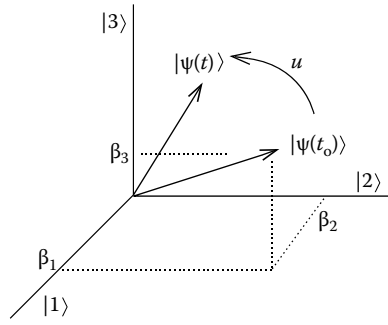


FIGURE 5.6 The evolution operator causes the wave function to move in Hilbert space. The unitary operator depends on the Hamiltonian. Therefore, it is really the Hamiltonian that causes the wave function to move.

as shown in Figure 5.6. Also, because all quantum mechanical wave functions have unit length and never anything else, the operator $\hat{u}$ must be unitary! The coefficients depend on time and so do the probabilities $P(n) = |\langle n | \psi(t) \rangle|^2 = |\beta_n(t)|^2$. We will see some simple examples in the next section where the total Hamiltonian does not depend on time and therefore, β 's depend on time only through a trivial phase factor of the form $e^{i\omega t}$, and therefore probabilities $P(n) = |\beta_n|^2$ do not depend on time.

5.1.7 COLLAPSE OF THE WAVE FUNCTION

The collapse of the wave function is one of the most interesting aspects of quantum theory (certainly one of the most imaginative). Comparing the quantum wave function with a classical wave (on a string for example) helps to highlight some of the differences. We already know one distinction in that the quantum wave functions must always be normalized to unity. A second distinction concerns the process of making a measurement on the Fourier superimposed wave. The collapse deals with how a superposed wave function behaves when a measurement is made of an observable. The collapse is random and outside the normal evolution of the wave function; a dynamical equation does not govern the collapse.

First, we introduce the collapse of the wave function. Suppose we are most interested in the energy of the system (although any Hermitian operator will work) and that the energy has quantized values $\{E_1, E_2, \dots\}$ where $\mathcal{H} |\phi_n\rangle = E_n |\phi_n\rangle$. Further assume that an electron resides in a superposed state

$$|\psi\rangle = \sum_n \beta_n |\phi_n\rangle \quad (5.14)$$

Making a measurement of the energy produces a single energy value E_n (for example). To obtain the single value E_n , the particle must be in the single state $|\phi_n\rangle$. We therefore realize that making a measurement of the energy somehow changes the wave function from $|\psi\rangle$ to $|\phi_n\rangle$. How does the wave function $|\psi\rangle$ collapse to $|\phi_n\rangle$?

Let us now be a little more specific about the meaning of the collapse of the wave function by using an example of an electron in an infinitely deep well. The upward pointing arrows at the sides of the well in Figure 5.7 show that the potential energy V becomes infinite there. We find only certain allowed wavelengths for the electron wave. Those sine waves fitting exactly in the well provide the most fundamental states $|\phi_n\rangle$

$$\left\{ \phi_n(x) = \langle x | \phi_n \rangle = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right), \quad n = 1, 2, \dots \right\}$$

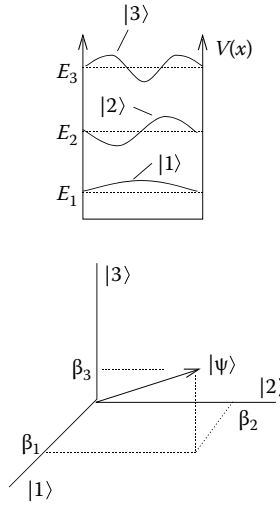


FIGURE 5.7 Three of the basis functions for the infinitely deep well.

These basis states are also the energy eigenstates

$$\hat{\mathcal{H}}|\phi_n\rangle = E_n|\phi_n\rangle \quad (5.15)$$

Figure 5.7 shows several allowed energy levels E_n corresponding to basis functions ϕ_n . In general, an electron in the well occupies a superposition state $|\psi\rangle$ in the Hilbert space

$$|\psi\rangle = \sum_n \beta_n(t) |\phi_n\rangle \quad (5.16)$$

Making a measurement of the energy causes the wave function $|\psi\rangle$ to collapse to a basis vectors $|\phi_n\rangle$ with probability $P(n) = |\beta_n|^2$. The bottom portion of Figure 5.7 indicates the electron occupies state $|\psi\rangle$ at time t . Making a measurement causes $|\psi\rangle$ to spontaneously degenerate to one of the basis vectors. A measurement of the energy causes the wave $|\psi\rangle$ to suddenly become one of the sine waves depicted in the top of Figure 5.7.

The quantum mechanical and classical waves behave amazingly different. Consider a string with both ends tied down. Imagine that someone plucks the string—maybe the wave looks like a triangular wave. The wave consists of the superposition of elementary sine waves. If the classical wave function could “collapse” when measuring an observable like energy or speed, then the triangular wave would suddenly become a perfectly defined sine wave! This does not at all agree with our experience. Disturbing the triangular wave might distort it but the wave does not suddenly become a perfect sine wave!

Let us discuss how we might mathematically represent the process of measuring an observable. So far, we claim to model the measurement process by applying a Hermitian operator to a state. However, we have shown the process only for eigenstates

$$\hat{\mathcal{H}}|\phi_n\rangle = E_n|\phi_n\rangle \quad (5.17)$$

In fact, the interpretation of Equation 5.17 does not match the processes of “measuring an observable” since we expect the results to be a number such as E_n and not the vector $E_n|\phi_n\rangle$.

How would we interpret the case when measuring an observable for a superposed wave function such as in Equation 5.16? If we apply $\hat{\mathcal{H}}$ to the vector $|\psi\rangle$ we find

$$\hat{\mathcal{H}}|\psi\rangle = \sum_n \beta_n(t) \hat{\mathcal{H}}|\phi_n\rangle = \sum_n \beta_n(t) E_n |\phi_n\rangle \quad (5.18)$$

This last equation attempts to measure the energy of a particle in state $|\psi\rangle$ at time t . So what is the result of the observation? While mathematically correct, this last equation does not accurately model the “act of observing!” Observing the superposition wave function must disturb it and cause it to collapse to one of the eigenstates! The process of observing a particle must therefore involve a projection operator! The collapse must introduce a time dependence beyond that in the coefficient $\beta_n(t)$. The interaction between the external measurement agent and the system introduces uncontrollable changes in time.

Let us show how the “observation act” might be modeled. Suppose for example, that the observation causes the wave function to collapse to state 2 (of course it could also collapse to states 1 or 3 with nonzero probability) for Figure 5.7. The mathematical model for the “act of observing” the energy state should include a projection operator $\hat{P}_2 = (1/\beta_2)\langle\phi_2|$ where $\hat{P}_2$ includes a normalization constant of $1/\beta_2$ for convenience (the symbol P should not be confused with the momentum operator and probability). The operator corresponding to the “act of observing” could be written as $\hat{P}_2\hat{\mathcal{H}}$. The results of the observation becomes

$$\hat{P}_2\hat{\mathcal{H}}|\psi\rangle = \sum_n \beta_n(t) \frac{1}{\beta_2} \langle\phi_2|\hat{\mathcal{H}}|\phi_n\rangle = E_2$$

However, we do not know a priori into which state the wave function will collapse and therefore cannot say $\hat{P}_2\hat{\mathcal{H}}$ represents the “act of making an observation” since we cannot rule out the quantity $\hat{P}_1\hat{\mathcal{H}}$ for example. We can only give the probability of the wave function collapsing into a particular state. One could define a measurement operator with basis states as the range that include the probability amplitude such as for $\hat{M}|\psi\rangle = \{\beta_1|1\rangle, \beta_2|2\rangle, \dots\}$ or perhaps for a range of eigenvalues with probabilities arranged as ordered pairs $\hat{M}|\psi\rangle = \{(E_1, \beta_1), (E_2, \beta_2), \dots\}$. The point is, one cannot assign a definite formula to $\hat{M}$ in the traditional sense since one cannot know a priori into which state the wave function will collapse. The probability of it collapsing into state $|n\rangle$ must be $|\beta_n|^2 = \beta_n^* \beta_n = |\langle\phi_n|\psi\rangle|^2$, which is obviously related to the expansion coefficients $\beta_n(t)$. We will find other interpretations for the measurement process and realize that quantities such as $\langle\psi|\hat{\mathcal{H}}|\psi\rangle$ give a single quantity E that represents an average energy. In fact, for the eigenstates we must find $\langle\phi_n|\hat{\mathcal{H}}|\phi_n\rangle = E_n$ where E_n must be a sharp value. The difference between $\langle\psi|\hat{\mathcal{H}}|\psi\rangle = \bar{E}$ and $\langle\phi_n|\hat{\mathcal{H}}|\phi_n\rangle = E_n$ has to do with the fact that the first one gives an average value (there must be a nonzero standard deviation lurking about) and the second one produces a sharp value (a standard deviation of zero).

5.1.8 INTERPRETATIONS OF THE COLLAPSE

So far in the discussion, we make a distinction between an undisturbed and a disturbed wave function. For the undisturbed wave function, the components in a generalized summation

$$|\psi\rangle = \sum_n \beta_n(t) |\phi_n\rangle \quad (5.19)$$

maintain their phase relation as the system evolves in time. In this case, the components $\beta_n(t)$ satisfy a differential equation (which implies the components must be continuous).

The undisturbed wave function follows the dynamics embedded in Schrödinger's equation. The general wave function satisfies

$$\hat{\mathcal{H}}|\psi\rangle = \sum_n \beta_n(t) \hat{\mathcal{H}}|\phi_n\rangle = \sum_n \beta_n(t) E_n |\phi_n\rangle \quad (5.20)$$

The collection of eigenvalues E_n make up the spectrum of the operator $\hat{\mathcal{H}}$. The coefficient β_n is the probability amplitude for the particle to be found in state ϕ_n with energy E_n .

Disturbing a wave function causes it to collapse (or make a transition) to one of the basis states at some point in time. The collapse does not affect the basis vectors in the generalized summation

$$|\psi\rangle = \sum_n \beta_n(t) |\phi_n\rangle$$

The components $\beta_n(t)$ must undergo catastrophic discontinuous behavior that the differential equation for the naturally evolving system cannot account. For example, if the wave function collapses as $|\psi\rangle \rightarrow |\phi_i\rangle$ then the coefficients must change according to $\beta_n(t) = \delta_{ni}$ since only the i th component remains afterward. Once the wave function collapses to one of the basis states, a randomizing process must be applied to the system for the wave function to move away from that basis state. The theory must be refined to account for the collapse; at the very least, we must incorporate the interaction between the observer and the system. Prior to the collapse, the coefficients give the probability $P(n) = |\beta_n|^2$ that the wave function collapses to the n th basis vector. Therefore, the coefficients give the probability of finding the energy E_n when making a measurement.

How can we physically picture the wave function and the collapse? We can imagine a number of different interpretations. For the first view, people sometimes view the wave function as a mathematical construct describing the probability amplitude. They assume that the particle occupies a particular state although they do not know which one. They make a measurement to determine the state the particle (or system) actually occupies. Before a measurement, they have limited information of the system. They know the probability

$$P(n) = |\beta_n|^2$$

that the particle occupies a given fundamental state (basis vector). Therefore, they know a wave function by the superposition of $\beta_n |\phi_n\rangle$. Making a measurement naturally changes the wave function because they then have more information on the actual state of the particle. After the measurement, they know for certain that the electron must be in state i for example. Therefore, they know $\beta_i = 1$ while all the other β must be zero. In effect, the wave function collapses from ψ to ϕ_i . With this first view, they ascribe any wave motion of the electron to the probability amplitude while implicitly assuming that the electron occupies a single state and behaves as a point particle. Making a measurement removes their uncertainty. The collapse refers to probability and nothing more.

As a second picture, and probably the most profound, let us view the collapse of the wave function as more related to physical phenomena. The Copenhagen interpretation (refer to Max Jammer's book) of a quantum particle in a superposed state

$$|\psi\rangle = \sum_n \beta_n(t) |\phi_n\rangle$$

views the particle as simultaneously existing in all of the fundamental states $|\phi_n\rangle$. In this case, we do not think of the particle as occupying a definite state $|\phi_i\rangle$. Somehow the particle simultaneously has all of the attributes of all of the fundamental states. A measurement of the particle forces it to

“decide” on one particular state. This second point of view requires some explanation using examples and it produces one of the most profound theorems of modern times—Bell’s theorem.

First let us consider the case of a particle described by the wave function $\psi(x)$. We will see later that this wave function can also be interpreted as the probability amplitude which means that the probability density of finding a particle at point x must be $\rho(x) = |\psi(x)|^2$. Recall that a general wave function can be expanded in a coordinate basis as

$$|\psi\rangle = \int dx |x\rangle \langle x|\psi\rangle \quad (5.21)$$

The components $\psi(x) = \langle x|\psi\rangle$ of the expansion must be the probability amplitudes according to Section 5.1.2. We must imagine that somehow the particle simultaneously occupies all states $|x\rangle$. We might picture the particle as a cloud extending over a large region of space. Suppose we make a measurement of the position of the particle. According to our second point of view, the wave function of the particle must collapse to a single point in space (or small volume). Pictorially, we imagine that the cloud suddenly condenses into this small region! Recall that the collapse should occur instantaneously. However, if we interpret the mass of the particle as somehow spread over space, then the collapse would violate special relativity since not even massless particles (like photons) can travel faster than light! Let us take another example connected with the Einstein–Podolsky–Rosen (EPR) paradox and related to Bell’s theorem.

Suppose a system of atoms can emit two correlated photons (entangled) in opposite directions. We require that the polarization of one to be tied with the polarization of the other. For example, suppose every time that we measure the polarization of photon A, we find photon B to have the same polarization. However, let us assume that each photon can be transversely polarized to the direction of motion according to

$$|\psi^a\rangle = \beta_1^a |1\rangle + \beta_2^a |2\rangle \quad (5.22)$$

where

- $|1\rangle, |2\rangle$ represent the x and y polarization directions
- a represents particle A or B

This last equation says that the wave moves along the z -direction but polarized partly along the x -direction and partly along the y -direction. We regard each photon as simultaneously existing in both polarized states $|1\rangle, |2\rangle$. If a measurement is made on photon A, and its wave function collapses to state $|1\rangle$, then the wave function for photon B simultaneously collapses to state $|1\rangle$ (for example). The collapse occurs even though the photons might be separated by several light years! Apparently the collapse of one can influence the other at speeds faster than light!

Some researchers are presently trying to find practical methods of making “faster than light” communicators. The ideas center on sending two correlated photons in opposite directions across the universe. If observer A wants to send a message to observer B, separated by many light years, then observer A arranges to have the polarization of one photon in state 1. The other photon, many light years away, will have its polarization as state 2 (for example). The states 1, 2 can represent yes or no answers to questions. When photon 1 is forced to collapse to state 1, it requires photon 2 to simultaneously collapse to state 2. Most commercial bookstores carry a number of “easy reading” accounts of this endeavor.

5.1.9 NONCOMMUTING OPERATORS AND THE HEISENBERG UNCERTAINTY RELATION

Consider two Hermitian operators $\hat{A}, \hat{B}$ corresponding to two observables. Figure 5.8 indicates that measuring $\hat{A}$ collapses the wave function $|\psi\rangle$ into one of many fundamental states. Suppose the wave

$$\begin{array}{c}
 |\psi\rangle \xrightarrow{\hat{A}} |a\rangle \xrightarrow{\hat{A}} |a\rangle \xrightarrow{\hat{A}} \\
 |\psi\rangle \xrightarrow{\hat{B}} |b\rangle \xrightarrow{\hat{B}} |b\rangle \xrightarrow{\hat{B}}
 \end{array}$$

FIGURE 5.8 Repeatedly applying an operator to a state gives the same number.

function collapses to the state $|a\rangle$. Repeated measurements of observable A produces the sequence a, a, a and so on. The dispersion (standard deviation) for the sequence must be zero. We see that once the wave function collapses, the operator $\hat{A}$ cannot change the state since it produces the same state $\hat{A}|a\rangle = a|a\rangle$. Similar comments apply to $\hat{B}$. Now we can see what happens when two operators do not influence each others eigenstates.

Let us suppose that the two observables $\hat{A}, \hat{B}$ can be measured at the same time without dispersion; this means we can repeatedly measure $\hat{A}, \hat{B}$ and find the same result each time. We will use the shortcut phrase of “simultaneous observables.” Let us assume that $|\phi\rangle$ characterizes the state of a particle such that $\hat{B}|\phi\rangle = b|\phi\rangle$ and $\hat{A}|\phi\rangle = a|\phi\rangle$. We can first apply $\hat{A}$ without affecting the results for $\hat{B}$. Applying $\hat{A}$ gives $\hat{A}|\phi\rangle = a|\phi\rangle$ and then applying $\hat{B}$ gives $\hat{B}\{\hat{A}|\phi\rangle\} = \hat{B}\{a|\phi\rangle\} = b\{a|\phi\rangle\} = b\{\hat{A}|\phi\rangle\}$. The result of observing $\hat{B}$ must be “b.” Therefore $\hat{A}$ does not affect the state of the particle (as far as concerns property B) and therefore does not disturb a measurement of $\hat{B}$. As a matter of generalizing the discussion, consider the following string of equalities.

$$\hat{A}\hat{B}|\phi\rangle = b\hat{A}|\phi\rangle = ab|\phi\rangle = a\hat{B}|\phi\rangle = \hat{B}a|\phi\rangle = \hat{B}\hat{A}|\phi\rangle \quad (5.23)$$

This relation must hold for every vector in the space since it holds for each basis vector. We can conclude

$$\hat{A}\hat{B} = \hat{B}\hat{A} \quad \rightarrow \quad 0 = \hat{A}\hat{B} - \hat{B}\hat{A} \equiv [\hat{A}, \hat{B}] \quad (5.24)$$

Therefore, simultaneous observables must correspond to operators that commute (refer to Section 3.12).

In this discussion, we say that we first apply $\hat{B}$ and then apply $\hat{A}$ according to their order in the product $\hat{A}\hat{B}$. We can make this explicit by perhaps imagining a time parameter t to indicate the time. For example,

$$\hat{A}(t_2)\hat{B}(t_1)|\psi\rangle \quad t_2 > t_1$$

In our case, the $\hat{A}, \hat{B}$ do not depend on time. We might say that we are making a measurement at the same time (simultaneous). We might think of the order $\hat{A}\hat{B}|\psi\rangle$ as a remnant of mathematical notation (involving t). Physically it does not matter if we write $\hat{A}\hat{B}$ or $\hat{B}\hat{A}$ because we require them to be measured at the same time. We expect to find the same answer if the operators correspond to simultaneous observables. Therefore we expect $\hat{A}\hat{B} = \hat{B}\hat{A}$ for simultaneous observables.

Now let us consider the situation where two operators $\hat{A}, \hat{B}$ interfere with the measurement of each other. Suppose $\hat{B}$ disturbs the eigenvector of $\hat{A}$ where the eigenvectors of $\hat{A}$ satisfy

$$\hat{A}|\phi_1\rangle = a_1|\phi_1\rangle \quad \hat{A}|\phi_2\rangle = a_2|\phi_2\rangle \quad (5.25)$$

Suppose that $\hat{B}$ disturbs the eigenstates of $\hat{A}$ according to

$$\hat{B}|\phi_1\rangle = |v\rangle \quad (5.26)$$

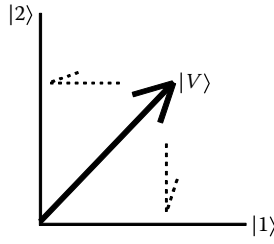


FIGURE 5.9 The vector collapses to either of two eigenvectors of A .

which appears in Figure 5.9. Assume that $|v\rangle$ has the expansion

$$|v\rangle = \beta_1|\phi_1\rangle + \beta_2|\phi_2\rangle \quad (5.27)$$

Now we can see that the order of applying the operators makes a difference. If we apply first $\hat{A}$ then $\hat{B}$, we find

$$\hat{B}\hat{A}|\phi_1\rangle = \hat{B}a_1|\phi_1\rangle = a_1|v\rangle \quad (5.28)$$

The reverse order $\hat{A}\hat{B}$ produces different behavior.

$$\hat{A}\hat{B}|\phi_1\rangle = \hat{A}|v\rangle = \hat{A}\{\beta_1|\phi_1\rangle + \beta_2|\phi_2\rangle\} = \beta_1a_1|\phi_1\rangle + \beta_2a_2|\phi_2\rangle \quad (5.29)$$

The results of the two orderings do not agree. We therefore surmise

$$\hat{A}\hat{B} \neq \hat{B}\hat{A}$$

Therefore, operators that interfere with each other do not commute. Further, the collapse of the wave function $|v\rangle$ under the action of $\hat{A}$ can produce either $|\phi_1\rangle$ or $|\phi_2\rangle$ so that the standard deviation for the measurements of $\hat{A}$ can no longer be zero.

Let us demonstrate how the noncommutivity of two observables might be imagined to produce the Heisenberg uncertainty relation. Assume a 2-D Hilbert space with two different basis sets $\{|\phi_1\rangle, |\phi_2\rangle\}$ and $\{|\psi_1\rangle, |\psi_2\rangle\}$ where $\hat{A}|\phi_n\rangle = a_n|\phi_n\rangle$ and $\hat{B}|\psi_n\rangle = b_n|\psi_n\rangle$. The relation between the basis vectors appears in Figure 5.10. We make repeated measurements of $\hat{B}\hat{A}$. Suppose we start with the wave function $|\phi_1\rangle$ and measure $\hat{A}$; we find the result a_1 . Next, let us measure $\hat{B}$. There is a 50–50 chance that $|\phi_1\rangle$ will collapse to $|\psi_1\rangle$ and a 50–50 chance it will collapse to $|\psi_2\rangle$. Let us assume that it

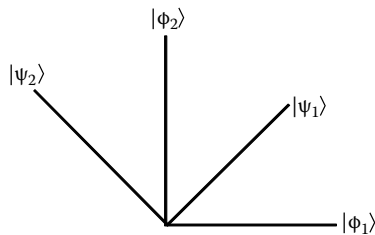


FIGURE 5.10 The two basis sets.

collapses to $|\psi_1\rangle$ and we find the value b_1 . Next we measure $\hat{A}$ and find that $|\psi_1\rangle$ collapses to $|\phi_2\rangle$ and we observe value a_2 , and so on. Suppose we find the following results for the measurements.

$$a_1 \quad b_1 \quad a_2 \quad b_1 \quad a_2 \quad b_2 \quad a_1 \quad b_1 \quad a_1 \quad b_2$$

Next lets sort this into two sets for the two operators

$$\begin{aligned} A &\rightarrow a_1 \quad a_2 \quad a_2 \quad a_1 \quad a_1 \\ B &\rightarrow b_1 \quad b_1 \quad b_2 \quad b_1 \quad b_2 \end{aligned}$$

We therefore see that both A and B must have a nonzero standard deviation. Section 3.12 shows how the observables must satisfy a relation of the form $\sigma_A \sigma_B \geq \text{constant} \neq 0$. We find a nonzero standard deviation when we measure two noncommuting observables and the wave function collapses to different basis vectors. Had we repeatedly measured A, we would have found $a_1 \quad a_1 \quad a_1 \quad a_1 \quad a_1$ which has zero standard deviation.

5.1.10 COMPLETE SETS OF OBSERVABLES

As previously discussed, we define the state of a particle or a system by specifying the values for a set of observables

$$\{\hat{O}_1, \hat{O}_2, \dots\}$$

such as $\hat{O}_1 = \text{energy}$, $\hat{O}_2 = \text{angular momentum}$, and so on. We know that each Hermitian operator induces a basis set. The direct product space has a basis set of the form $|o_1, o_2, \dots\rangle = |o_1\rangle|o_2\rangle \dots$ where the eigenvalue o_n occurs in the eigenvalue relation $\hat{O}_n|o_1 \dots o_n \dots\rangle = o_n|o_1 \dots o_n \dots\rangle$. These operators all share a common basis set. Knowing that the particle occupies the state $|o_1, o_2, \dots\rangle$ means that we exactly know the outcome of measuring the observables $\{\hat{O}_1, \hat{O}_2, \dots\}$.

How do we know which observables to include in the set? Naturally we include observables of interest to us. We make the set as large as possible without including Hermitian operators that do not commute. Because commuting operators produce a common basis set, we can make measurements of one without affecting the results of measuring another one. However, not all Hermitian operators commute and they therefore do not share common basis vectors. The case of the position $\hat{x}$ and momentum $\hat{p}$ operators provide a well-known example. This means that the measurements of “noncommuting” operators interfere with each other.

In quantum theory, we specify the basic states (i.e., basis states) of a particle or system by listing the observable properties. The particle might have a certain energy, momentum, angular momentum, polarization, etc. Knowing the value of all observable properties is equivalent to knowing the basis states of the particle or system. Each physical “observable” corresponds to a Hermitian operator $\hat{O}_i$ which induces a preferred basis set for the respective Hilbert space V_i (i.e., the eigenvectors of the operator comprises the “preferred” basis set). The multiplicity of possible observables means that a single particle can “reside” in many Hilbert spaces at the same time since there can be a Hilbert space V_i for each operator $\hat{O}_i$. The particle can therefore reside in the direct product space (see Chapters 2 and 3) given by

$$V = V_1 \otimes V_2 \otimes \dots$$

where

V_1 might describe the energy

V_2 might describe the spin, and so on

The basis set for the direct product space consists of the combination of the basis vectors for the individual spaces such as

$$|\Psi\rangle = |\phi, \eta, \dots\rangle = |\phi\rangle|\eta\rangle \dots$$

where we assume, for example, that the space spanned by $\{|\phi\rangle\}$ refers to the energy content and $\{|\eta\rangle\}$ refers to spin, etc.

The basis states can be most conveniently labeled by the eigenvalues of the commuting Hermitian operators. For example, $|E_i, p_j\rangle$ represents the state of the particle with energy E_i and momentum p_j assuming, of course, that the Hamiltonian and momentum commute. These two operators might represent all we care to know about the system.

5.2 FUNDAMENTAL OPERATORS AND PROCEDURES FOR QUANTUM MECHANICS

Quantum mechanics represents physical objects in terms of mathematics. As such, there must be well-defined symbols and procedures established to first translate the physical situation into the mathematics, provide for manipulation of the symbols, and then to interpret the results back in terms of the physical world. The Hilbert spaces have a close symbiotic relation with the quantum mechanics. The present section discusses usable forms of the operators and shows the Schrödinger wave equation (SWE) as the primary quantity of interest for determining the time evolution of quantum level particles and systems. The next section applies the formalism to examples of a 1-D infinitely deep and finitely deep quantum well.

5.2.1 SUMMARY OF ELEMENTARY FACTS

Electrons, holes, photons, and phonons can be pictured as particles or waves. Momentum and energy usually apply to particles while wavelength and frequency apply to waves. The momentum and energy relations provide a bridge between the two pictures

$$p = \hbar k \quad E = \hbar \omega \quad (5.30a)$$

where $\hbar = h/2\pi$ and “ h ” is Planck’s constant. For both massive and massless particles, the wave vector and angular frequency can be written

$$k = \frac{2\pi}{\lambda} \quad \omega = 2\pi\nu \quad (5.30b)$$

where λ and ν represent the wavelength and frequency (Hz). For massive particles, the momentum $p = mv$ can be related to the wavelength by

$$\lambda = \frac{h}{mv}$$

for mass m and velocity v .

“Hermitian operators” $\hat{O}$ represent observables, which are physically measurable quantities such as the momentum of a particle, temperature, electric field, and position in a laboratory. If Φ is an eigenvector (basis vector), then the eigenvector equation $\hat{O}\Phi = o\Phi$ gives the result of the observation when the particle occupies eigenstate Φ where o , a “real” constant, represents the results of a measurement. If for example, $\hat{O}$ represents the momentum operator, then o must be the momentum of the particle when the particle occupies state “ Φ .” We can write an eigenfunction equation for every observable. The result of every physical observation must always be an eigenvalue.

Quantum mechanics does not allow us to simultaneously “know” the values of all observables. For example, position and momentum of a particle cannot be “simultaneously” known with infinite accuracy for both quantities.

5.2.2 MOMENTUM OPERATOR

The mathematical theory of quantum mechanics admits many different forms for the operators. The “spatial-coordinate representation” (see Appendix L and Section 3.2.6) relates the momentum to the spatial gradient. To find an operator representing the momentum, consider the plane wave $\Phi = Ae^{i\vec{k}\cdot\vec{r}-i\omega t}$. The gradient gives

$$\nabla\Phi = i\vec{k}\Phi = i\frac{\vec{P}}{\hbar}\Phi$$

where $\vec{P} = \hbar\vec{k}$ is the momentum. We assume that this form holds for all eigenvectors of the momentum operator. Therefore, comparing both sides of the last equation, it appears reasonable to identify the momentum operator with the spatial derivative

$$\hat{P} = \frac{\hbar}{i}\nabla = \frac{\hbar}{i}\left[\tilde{x}\frac{\partial}{\partial x} + \tilde{y}\frac{\partial}{\partial y} + \tilde{z}\frac{\partial}{\partial z}\right] \quad (5.31)$$

The momentum operator has both a vector and operator character. The operator character comes from the derivatives in the gradient and the vector character comes from the unit vectors appearing in the gradient. We identify the individual components of the momentum as

$$\hat{P}_x = \frac{\hbar}{i}\frac{\partial}{\partial x} \quad \hat{P}_y = \frac{\hbar}{i}\frac{\partial}{\partial y} \quad \hat{P}_z = \frac{\hbar}{i}\frac{\partial}{\partial z}$$

The position operator $\hat{x}$ becomes the coordinate x in the coordinate representation. Sometimes it is more convenient to work with alternate notation

$$x_m = \begin{cases} x & m=1 \\ y & m=2 \\ z & m=3 \end{cases} \quad \hat{P}_m = \begin{cases} \hat{P}_x & m=1 \\ \hat{P}_y & m=2 \\ \hat{P}_z & m=3 \end{cases}$$

The position and momentum do not commute.

$$[x_m, \hat{P}_n] = i\hbar\delta_{mn}$$

In general, conjugate variables (i.e., $m=n$) refer to the same degree of freedom and do not commute.

5.2.3 HAMILTONIAN OPERATOR AND THE SCHRÖDINGER WAVE EQUATION

We can observe the total energy of a particle or a system (the word “system” usually denotes a collection of particles—not necessarily all of the same type). We know that there exists a Hermitian operator $\hat{\mathcal{H}}$ representing the total energy. Earlier sections in this book on classical mechanics develop the special mathematical properties of the classical Hamiltonian and associated Lagrangian. Quantum theory determines the fundamental states and allowed energies of a particle through an eigenvalue equation

$$\hat{\mathcal{H}}|\Phi\rangle = E|\Phi\rangle \quad \text{or} \quad \hat{\mathcal{H}}\Phi = E\Phi \quad (5.32)$$

where $|\Phi\rangle$ is an energy basis function for the particle. The eigenvector equation cannot easily be solved without more detail on the form of the operator. In general, we need a wave equation in order to find the wave motion associated with the probability of the quantum particles.

One can determine another form for the energy operator using a plane wave representation for the wave function of a particle. Even though we use a specific wave function, we require the partial differential equation to hold, in general, even for arbitrary wave functions. A plane wave traveling along the $+z$ -direction with phase velocity $v = \omega/k$ has the form

$$\Phi = Ae^{ikz - i\omega t}$$

Differentiating with respect to time and using $E = \hbar\omega$ gives us

$$\frac{\partial\Phi}{\partial t} = -i\omega\Phi = -i\frac{E}{\hbar}\Phi \quad \rightarrow \quad i\hbar\frac{\partial\Phi}{\partial t} = E\Phi \quad (5.33)$$

Comparing Equations 5.33 and 5.32, we are encouraged to write

$$\hat{\mathcal{H}}\Psi = i\hbar\frac{\partial\Psi}{\partial t} \quad (5.34)$$

The Schrödinger wave equation (SWE) in Equation 5.34 provides the dynamics for the motion of the quantum particles. The dynamics in the SWE can refer to a variety of motions including the motion of a particle through space or the evolution of the spin of a particle. One should expect this wide range applicability of the SWE since it essentially embodies the Hamiltonian H as related to that in Chapter 4. However, the quantum Hamiltonian is an operator and must operate on a wave function or vector. Any wave function solving Equation 5.34 can be Fourier expanded in the basis set. Equation 5.34 has only a first derivative in time contrary to the usual form of a classical wave equation (the wave equation for electromagnetics for example). The single time derivative leads to a probability interpretations for the wave function, and to the conservation of particle number (i.e., an equation of continuity for probability) in addition to the introduction of the complex number multiplying the time derivative.

We must specify the form of the energy operator in terms of other quantities related to the energy of the system. For a single particle (without electromagnetic fields for example), we know that the total energy can be related to the kinetic and potential energy. We must keep in mind throughout this procedure that $\hat{\mathcal{H}}$ is an operator; any expression for $\hat{\mathcal{H}}$ must therefore contain operators.

The usual procedure for finding the quantum mechanical Hamiltonian starts by writing the classical Hamiltonian (i.e., energy) and then substituting operators for the dynamical variables (i.e., observables). The operators are then required to satisfy commutation relations which determine whether or not the corresponding observables are simultaneously observable (i.e., the Heisenberg uncertainty relations must be satisfied).

The classical Hamiltonian for a particle with potential energy $V(\vec{r})$ can be written as

$$\mathcal{H} = ke + pe = \frac{p^2}{2m} + V(\vec{r}) \quad (5.35)$$

The quantum mechanical Hamiltonian can be found by replacing all dynamical variables, which consist of $\vec{r}$ and $\vec{p}$ in this case, with the equivalent operator. We will work in the spatial-coordinate representation (Appendix L and Section 3.2.6) so that we denote the position vector by $\vec{r}$ and we use Equation 5.31 for the momentum. The quantum mechanical Hamiltonian can be written as

$$\hat{\mathcal{H}} = \frac{\hat{p}^2}{2m} + V(\vec{r}) = \frac{1}{2m} \left(\frac{\hbar}{i} \nabla \right) \cdot \left(\frac{\hbar}{i} \nabla \right) + V(\vec{r}) = -\frac{\hbar^2}{2m} \nabla^2 + V(\vec{r}) \quad (5.36)$$

Question: If we cannot simultaneously and precisely measure both momentum and position for the Hamiltonian, how can the energy ever have an exact value? We resolve this apparent contradiction by noting that the Hamiltonian is well defined for an energy eigenfunction basis set even though momentum and position cannot be simultaneously exactly known.

As a note, the basis vectors by themselves do not solve the Schrödinger equation. Instead, the functions of the form $e^{Et/\hbar}|E\rangle$ and their superposition do solve the Schrödinger equation.

5.2.4 INTRODUCTION TO COMMUTATION RELATIONS AND HEISENBERG UNCERTAINTY RELATIONS

Elementary theory and experiment tell us that certain observables cannot be simultaneously measured with “infinite precision.” For example, position and momentum as conjugate variables must satisfy the Heisenberg uncertainty relation

$$\Delta x \Delta p \geq \hbar/2 \quad (5.37a)$$

here we interpret x and p as the observed values of the observables. The symbol “ Δ ” actually refers to the standard deviation found in probability theory. Equation 5.37a can be rewritten as

$$\sigma_x \sigma_p \geq \hbar/2 \quad (5.37b)$$

where the symbols σ_x, σ_p represent the standard deviation in the position and momentum (corresponding to the x -direction). Section 5.1 shows how to calculate the expectation values of operators. The standard deviations σ_x, σ_p are not operators since the expectation values have been calculated. The Heisenberg uncertainty relation tells us that repeated measurements of position and momentum yields a range of values for x and p . Using linear algebra, we can show that the Heisenberg uncertainty relation actually follows from properties of the corresponding operators (as shown later in the present section).

To say that momentum and position cannot be precisely measured at the same time is equivalent to writing

$$\hat{P}_x \hat{x} \Psi \neq \hat{x} \hat{P}_x \Psi$$

as discussed in Section 5.1. Introducing the commutator notation

$$[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}$$

we can write

$$[\hat{x}, \hat{P}_x] \neq 0$$

As an important note, we must always treat the commutator itself as an operator and it must therefore always operate on a function. For example, to say that an operator $\hat{A} = 0$ really means that for every function Ψ in the function space we must have $\hat{A}\Psi = 0$.

Let us evaluate the commutation relation for position and momentum using spatial coordinates as an introductory example. For this case, we already know (Appendix L and Section 3.2.6)

$$\hat{x} = x \quad \hat{P}_x = \frac{\hbar}{i} \frac{\partial}{\partial x}$$

and so the commutator can be evaluated as

$$[x, \hat{P}_x]\Psi = x \frac{\hbar}{i} \frac{\partial}{\partial x} \Psi - \frac{\hbar}{i} \frac{\partial}{\partial x} (x\Psi) = i\hbar\Psi$$

As a note, we only get the right answer because the commutator operates on the function Ψ . The reader is well advised to try the calculation without the function and verify that the wrong answer is obtained. We require this last relation to hold for every function in the vector space. We can therefore write the operator equation for the commutator as

$$[\hat{x}, \hat{P}_x] = i\hbar$$

We can also show two other relations (among many)

$$[\hat{x}, \hat{x}] = 0 \quad [\hat{y}, \hat{P}_x] = 0$$

In particular, notice that the y -position coordinate commutes with the momentum for the x -coordinate. Commuting operators corresponding to dynamical variables that can be simultaneously and precisely measured. A “nonzero” commutator produces an uncertainty relation as verified in the next section.

We can demonstrate other uncertainty relations. For example, the uncertainty relation between energy and time can be written as

$$\Delta E \Delta t \geq \frac{\hbar}{2} \quad (5.38)$$

We will see that only Gaussian distributions give the equality sign.

Example 5.4

Suppose two states are separated by energy E . Suppose that we know this difference in energy E to within ΔE , that is, the actual energy lies in the interval

$$E \pm \frac{\Delta E}{2}$$

We then know the amount of time required for the particle to make a transition to within

$$\Delta t \geq \frac{\hbar}{2\Delta E}$$

5.2.5 DERIVATION OF THE HEISENBERG UNCERTAINTY RELATION

Previous sections discuss how commuting operators correspond to dynamical variables that can be simultaneously and precisely measured. If two operators $\hat{A}, \hat{B}$ commute then there exists a simultaneous set of basis functions $|a, b\rangle = |a\rangle|b\rangle$ such that

$$\hat{A}|a, b\rangle = a|a, b\rangle \quad \text{and} \quad \hat{B}|a, b\rangle = b|a, b\rangle$$

and vice versa. We can show that if two operators do not commute then there exists a Heisenberg uncertainty relation between them.

We now show that two noncommuting Hermitian operators must always produce an uncertainty relation.

THEOREM 5.1

If two operators $\hat{A}, \hat{B}$ are Hermitian and satisfy the commutation relation $[\hat{A}, \hat{B}] = i\hat{C}$ then the observed values a, b of the operators must satisfy a Heisenberg uncertainty relation of the form $\sigma_a \sigma_b \geq \frac{1}{2} |\langle \hat{C} \rangle|$.

Proof: Consider the “real, positive number” defined by

$$\xi = \langle (\hat{A} + i\lambda\hat{B})\psi | (\hat{A} + i\lambda\hat{B})\psi \rangle$$

which we know to be a real and positive since the inner product provides the length of the vector. The vector, in this case, is defined by

$$|(\hat{A} + i\lambda\hat{B})\psi\rangle = (\hat{A} + i\lambda\hat{B})|\psi\rangle$$

We assume that λ is a real parameter. Now working with the number ξ and using the definition of adjoint, namely

$$\langle \hat{O}f | g \rangle = \langle f | \hat{O}^+ g \rangle$$

we find

$$\begin{aligned} \xi &= \langle \psi | (\hat{A} + i\lambda\hat{B})^+ (\hat{A} + i\lambda\hat{B}) | \psi \rangle = \langle \psi | (\hat{A} + i\lambda\hat{B})^+ (\hat{A} + i\lambda\hat{B}) | \psi \rangle \\ &= \langle \psi | (\hat{A}^+ - i\lambda\hat{B}^+) (\hat{A} + i\lambda\hat{B}) | \psi \rangle = \langle \psi | (\hat{A} - i\lambda\hat{B}) (\hat{A} + i\lambda\hat{B}) | \psi \rangle \end{aligned}$$

where the last step uses the Hermiticity of the operators $\hat{A}, \hat{B}$. Multiply the operator terms in the bracket expression and suppress the reference to the wave function (for convenience) to obtain

$$\xi = \langle \hat{A}^2 \rangle - \lambda \langle \hat{C} \rangle + \lambda^2 \langle \hat{B}^2 \rangle \geq 0$$

which must hold for all values of the parameter λ . The minimum value of the positive real number ξ is found by differentiating with respect to the parameter λ .

$$\frac{\partial \xi}{\partial \lambda} = 0 \quad \rightarrow \quad \lambda = \frac{\langle \hat{C} \rangle}{2\langle \hat{B}^2 \rangle}$$

The minimum value of the positive real number ξ must be

$$\xi_{\min} = \langle \hat{A}^2 \rangle - \frac{1}{4} \frac{\langle \hat{C} \rangle^2}{\langle \hat{B}^2 \rangle} \geq 0$$

Multiplying through by $\langle \hat{B}^2 \rangle$ to get

$$\langle \hat{A}^2 \rangle \langle \hat{B}^2 \rangle \geq \frac{1}{4} \langle \hat{C} \rangle^2 \quad (5.39)$$

We could have assumed the quantities $\langle \hat{A} \rangle = \langle \hat{B} \rangle = 0$ and we would have been finished at this point. However, the commutator $[\hat{A}, \hat{B}] = i\hat{C}$ holds for the two Hermitian operators defined by

$$\hat{A} \rightarrow \hat{A} - \langle \hat{A} \rangle \quad \hat{B} \rightarrow \hat{B} - \langle \hat{B} \rangle$$

As a result, Equation 5.39 becomes

$$\langle (\hat{A} - \langle \hat{A} \rangle)^2 \rangle \langle (\hat{B} - \langle \hat{B} \rangle)^2 \rangle \geq \frac{1}{4} \langle \hat{C} \rangle^2$$

However, the terms in the angular brackets are related to the standard deviations σ_a and σ_b , respectively. We obtained the proof to the theorem by taking the square root of the previous expression

$$\sigma_a \sigma_b \geq \frac{1}{2} |\langle \hat{C} \rangle|$$

Notice that this Heisenberg uncertainty relation involves the absolute value of the expectation value of the operator C . By its definition, the operator C must be Hermitian and its expectation value must be real.

5.2.6 PROGRAM

The most basic procedure for describing a system (such as a quantum well) by quantum mechanics consists of finding an orthonormal set of functions to form a basis set. Most often, people have the greatest interest in the energy basis set that describes the possible energy levels. For a particle within a 1-D quantum well for example, the basis set will consist of sinusoids with positive energy levels (when the bottom of the well has zero potential). Other basis sets are possible such as for momentum or angular momentum. We might have interest in one particular basis set such as for energy (over another) because we want to know the possible energy levels of an electron (for example) or the probability that an electron has a given energy (i.e., that the electron resides in the corresponding energy state). However, finding the energy basis has more significance than the linear algebra of basis sets reveals. The Hamiltonian also provides the “dynamics of the system” meaning it provides the evolution of the system with time (interestingly, energy and time are conjugate variables with an uncertainty relation). Basically to determine how any (observable) quantity $\hat{O}$ evolves in time, it is necessary to know the Hamiltonian and either solve Schrödinger’s wave equation and calculate statistical moments or at least find the commutator $[\hat{H}, \hat{O}]$ (as will be discussed later). If we have interest in the probability of an electron having a certain spin for example, then we will need to know the spin basis set and we will need to know the Hamiltonian in order to predict how the spin changes with time.

Obtaining the basis set (energy basis set in the next section) is the most fundamental result of the analysis. Suppose an operator $\hat{O}$ produces a basis set $\{|\phi_1\rangle, |\phi_2\rangle, \dots\}$ meaning the eigenvector relation holds

$$\hat{O}|\phi_i\rangle = o_i|\phi_i\rangle$$

In general, a wave function $\psi(x, t)$ representing the particle or system can be a superposition of these basis vectors. Projecting the wave function onto one of the basis vectors $|\phi_i\rangle$ provides the component of the vector which is also the “probability amplitude” of finding the particle in that state. For example, the probability amplitude P_A and the probability P of finding the particle in state $|\phi_i\rangle$ will be

$$P_A = \langle \phi_i | \psi(t) \rangle \quad P(i) = |P_A|^2 = |\langle \phi_i | \psi(t) \rangle|^2 \quad (5.40a)$$

Equivalently this asks for the probability that the value o_i will be observed. For our examples in the next section, we will be interested in the probability of finding the electron in energy state E_i in which case, we need to find the eigenfunctions of the Hamiltonian (as the total energy). Also keep in mind, that we might be interested in the probability of finding the electron at some location in space rather than in one of the energy levels. In the case of a 1-D system (described by the x -direction for example), the wave function is projected into the basis set of coordinates to find the probability amplitude P_A and probability density ρ (probability per length) as

$$P_A = \langle x | \psi(t) \rangle = \psi(x, t) \quad \rho = |P_A|^2 = |\psi(x, t)|^2 = \psi^* \psi \quad (5.40b)$$

The probability of finding the particle in the range (a, b) will then be the integral of the probability density over that range.

Finding the energy basis set will be central to the next section. We look for solutions to the “time-independent Schrödinger equation”

$$\hat{\mathcal{H}}|\phi_n\rangle = E_n|\phi_n\rangle \quad (5.41a)$$

where $\hat{\mathcal{H}}$ represents the Hamiltonian (note that basis vectors are always independent of time). We must start with the “time-dependent Schrödinger equation”

$$\hat{\mathcal{H}}|\psi\rangle = i\hbar\partial_t|\psi\rangle \quad (5.41b)$$

and describe a procedure by which to extract the time-independent equation. Here $|\psi\rangle$ represents the general motion of the particle and it can be expressed as a superposition of basis vectors. Schrödinger’s equation also requires one to specify the quantum mechanical Hamiltonian.

The quantum mechanical Hamiltonian $\hat{\mathcal{H}}$ can be found as follows:

1. Write a classical Hamiltonian. That is, write an expression for the total energy of the system in question (potential and kinetic energy for the present section). The Hamiltonian must always be expressed in terms of the momentum and conjugate position coordinates (rather than velocity and position). For the case of a single electron able to move along the x -direction (momentum P_x) but confined by potential energy $V(x)$, this will take the form

$$\mathcal{H} = \frac{P_x^2}{2m} + V(x)$$

2. Convert all classical dynamical variables to operators such as momentum and position.

$$\hat{\mathcal{H}} = \frac{\hat{P}_x^2}{2m} + V(\hat{x}) \quad (5.42)$$

3. Require the operators to satisfy commutation relations such as $[\hat{x}, \hat{P}_x] = i\hbar$ for example.

The energy eigenfunctions and eigenvalues for a 1-D problem such as the quantum wells discussed in the next section can be found as follows:

1. Start with the time-dependent Schrödinger equation and project it into the coordinate basis set (Appendix L and Section 3.2.6), which means that $\hat{P}_x \rightarrow i\hbar\partial_x$, $\hat{x} \rightarrow x$, and the Hamiltonian becomes

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x)$$

The SWE becomes

$$\frac{-\hbar^2}{2m} \frac{\partial^2 \psi(x, t)}{\partial x^2} + V(x) \psi(x, t) = i\hbar \frac{\partial \psi(x, t)}{\partial t}$$

The boundary and initial conditions must also be specified. In particular, $\psi(x, t=0) \sim |\psi(0)\rangle$ provides the initial superposition of basis vectors.

2. Identify the time-independent Schrödinger equation by separating variables in the SWE using $\psi(x, t) = X(x)T(t)$ and symbolizing the separation constant by E .

$$\left[\frac{-\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \right] X(x) = E X(x) \quad (5.43)$$

The time-independent Schrödinger equation forms part of the Sturm-Liouville problem for finding the eigenfunctions and eigenvalues. Two boundary conditions form the remaining portion of the problem. A second-order ordinary differential equation requires two boundary conditions to completely determine the solution (which would apply to Equation 5.43 if E were known). However, for Equation 5.43, not only is X a priori unknown but the value of E is unknown and we have only two boundary conditions. Therefore, we do not expect to find a single unique solution for X and E . For this reason, we expect to find a set of eigenfunctions and a set of eigenvalues.

The time portion of the original partial differential equation has the form

$$T'_n = i\hbar E_n T_n$$

but it does not represent an eigenvalue problem since the E_n were determined in the Sturm-Liouville problem for X . Generally, only spatial coordinates will be associated with the Sturm-Liouville problem since the basis sets (modes) refer most often to a spatial distribution. For example, the basic modes of a violin string or the modes in an optical Fabre-Perot cavity refer to spatial distributions of energy or electric field.

The full solution (prior to applying the initial conditions) has the form

$$\psi(x, t) = \sum_n T_n(t) X_n(x)$$

where the symbol T_n represents the components of the vector rather than the symbol β_n previously used. Once having applied the initial conditions, which eliminates constants in T_n , the full solution will be known and written as a summation over the energy eigenstates.

5.3 EXAMPLES FOR SCHRÖDINGER'S WAVE EQUATION

The simple examples for the Schrödinger wave equation (SWE) examine the situations where an electron is confined to infinitely deep and finitely deep wells. These examples highlight the fundamental concepts in quantum mechanics with the underlying application of linear algebra. Nature can form a variety of quantum-well-like structures that confine electrons to small regions of space such as for an atom, a molecular orbital or for traps within a semiconductor material. Human-made quantum wells have become center stage to modern technology as a result of new material growth and fabrication techniques that can tailor the material and hence the electrostatic potential to confine a particle. These artificial methods produce quantum well lasers having one to eight quantum wells with barriers, quantum nanostructures and devices such as single electron transistors, and other more exotic devices with exotic operating principles. Subsequent chapters will discuss conduction through quantum well material using the transfer matrix formalism and apply it to one of the simplest models, the Kronig-Penney model, to predict the occurrence of bands in semiconductor material.

5.3.1 DISCUSSION OF QUANTUM WELLS

The singly dimensioned (1-D) infinitely deep quantum well consists of a center region with constant potential V_c , normally assumed to be zero, and at either edge ($x=0, L$), an impenetrable barrier such as shown in Figure 5.11 (in vacuum and not semiconductor material). The total energy E of the particle must be the sum of kinetic $P_x^2/2m$ and potential energy V . Classically, the particle will only be found in those regions where the total energy E has a value at least as large as the potential energy. In the barrier regions $x < 0$ and $x > L$, the difference $E - V$ becomes negative which then requires an imaginary momentum through the conservation of energy as $E - V = P_x^2/2m$. Electron waves with “real” momentum produce sinusoidal waves of probability amplitude $\psi(x, t)$ whereas imaginary momentum converts the sinusoids into real exponentials that represent an exponential decay. The particle can only escape the infinitely deep well by acquiring an infinite amount of energy. On the other hand, if the well has finite barriers with potential V_b (as for the finitely deep well in the next section), the electron only needs to gain an energy on the order of V_b to escape the well. Even though classically speaking the particle cannot be found in the regions where $E < V_b$, the electron can be found in those regions for quantum mechanical reasons related to quantum tunneling (without receiving extra energy to surmount the barrier).

The potential barriers for the infinitely deep wells produce wave functions (probability amplitudes) confined to the well region. The infinitely deep quantum well (Figure 5.11) has eigenfunctions of the form $\sqrt{2/L} \sin(n\pi x/L)$ where $n = 1, 2, \dots, \infty$. Each basis state produces its own corresponding probability density function. This occurs because even quantum mechanically, the electron cannot penetrate into the infinite barrier as will later become clear from studying the finitely deep well. For this reason, the probability density $\psi^*\psi$ must be zero outside the well and therefore the wave function must also be zero. In this case, the wave function must satisfy the boundary conditions of $\psi(x \leq 0, t) = 0 = \psi(x \geq L, t)$. In particular, one chooses the condition of $\psi(0, t) = 0 = \psi(L, t)$ for the SWE. Consider the example for the $n = 2$ basis function shown in Figure 5.11. Based on the probability density $|\psi|^2$, the figure indicates that we would be least likely to find the electron at $x = 0, L/2$, and L , and most likely to find it at $x = L/4$ and $3L/4$.

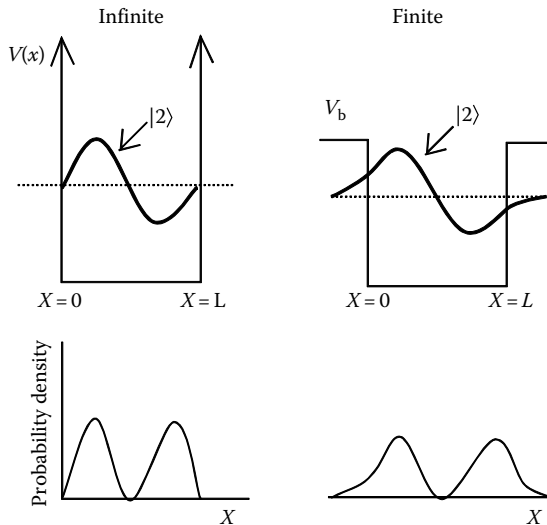


FIGURE 5.11 The infinitely (left) and finitely (right) deep well. The top diagrams shows an example wave function ($n=2$ in this case) in relation to the barriers. The dotted lines represent both the zero of the wave function and the energy E_2 corresponding to the shown wave function. The bottom diagrams show the probability (density) $\psi^*\psi$ of finding an electron at a specific location x .

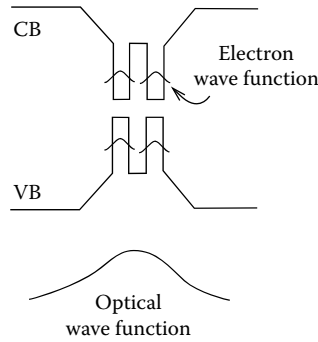


FIGURE 5.12 Well structure for a quantum well laser.

Two notes need to be mentioned. First, one speaks of the probability of finding the electron in a given spatial region and also of the probability of finding an electron in a given energy state. The probability amplitude of finding the particle in an eigenstate n is given by $\langle n | \psi(t) \rangle$ whereas the probability amplitude of finding the particle at a specific location x will be $\langle x | \psi(t) \rangle = \psi(x, t)$. Notice that even though the particle might be in a specific eigenstate so that repeated measurements of the energy produce the same “energy” value (zero variance), repeated measurements of “position” produce multiple values for position (spread out across 0 to L) and therefore produce nonzero variance for position. This behavior occurs because the SWE produces the eigenstates of energy and not the eigenstates of position as stated here. Second, it should be noted that confining the electron or particle to a small spatial region (because of the boundary conditions) produces the quantization of the energy and wave vector.

The quantum well has immediate application to the quantum well laser. The previous discussion focused on a quantum well formed in free space. Applying the quantum well to a material with conduction and valence bands can produce the structure shown in Figure 5.12 for the case of two wells separated by a barrier. Both the conduction and valence bands have quantum wells. The electrons will be confined to those wells for the conduction band while holes will be confined to those wells for the valence band. The two sets of wells appear inverted from each other since electron energy increase upward while hole energy increases in the downward direction. The wells are tailored by the type of semiconductors used since the barrier heights and well depths will depend on the material. For example, the barriers can be made of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ (where x represents the mole fraction) and the wells formed by GaAs as a result of the dependence of the band edge on mole fraction x . The wells can be grown by molecular beam epitaxy (MBE) for example. Connecting a battery to the structure will cause electrons to enter an electron state in the electron quantum well and a hole will enter a state in the hole well. Notice that the electron and hole cannot be at the bottom of the well as there are no quantum well states at the bottom. As a result, when the electron and hole recombine (in the AlGaAs system shown), the photon will have an energy somewhat larger than the band gap energy, which is the difference between the bottom of the electron well and the top of the hole well. Finally notice that the electron wave function has smaller “size” than the photon wave function shown in the figure. The wells confine the electrons and holes while the index of refraction confines the photon. In fact, the optical photon has such large wavelength for normal values of refractive index that it cannot be confined to such small regions as for the electron. However, it is interesting to ponder the fact that the confined electrons and holes (as well as those in atoms) can still produce the photons.

5.3.2 SOLUTIONS TO SCHRÖDINGER’S EQUATION FOR THE INFINITELY DEEP WELL

The present section solves Schrödinger’s equation for an electron confined to an infinitely deep well of width L . We will see that the SWE produces a basis set comprised of sine waves. Figure 5.11

shows the $n=2$ energy basis function and the corresponding probability density function $\psi^*\psi$. For the infinitely deep well shown, assume that the potential energy is zero at the bottom of the well (i.e., $V=0$ for $0 < x < L$). In this section, we outline the solution for Schrödinger's equation as applied to the infinitely deep well.

The boundary value problem consists of a partial differential equation for Schrödinger's time-dependent wave equation $\hat{H}|\Psi\rangle = i\hbar\partial_t|\Psi\rangle$ or using $\hat{H} = (\hat{p}^2/2m) + V(x)$ and substituting $\hat{p} = (\hbar/i)\partial_x$ and $V=0$ in the well region, we obtain

$$-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \Psi = i\hbar \frac{\partial \Psi}{\partial t} \quad (5.44a)$$

with boundary conditions

$$\Psi(0, t) = \Psi(L, t) = 0 \quad (5.44b)$$

where m is the mass of an electron. There should also be an initial condition (IC) for the time it should have the form $\Psi(x, 0) = f(x)$. The initial condition specifies the initial probability amplitude for each of the basis states (as can be seen by considering a Fourier series expansion of f). We are most interested in the basis states for now.

One can often use the technique for the separation of variables to find a solution to the partial differential equation. Set $\Psi(x, t) = X(x)T(t)$, substitute into the partial differential equation, and then divide both sides by Ψ to obtain

$$\frac{1}{x} \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} X \right) = i\hbar \frac{1}{T} \frac{\partial T}{\partial t} \quad (5.45a)$$

Both sides must be equal to a constant, called E . This last equation can be rewritten as

$$\frac{1}{x} \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} X \right) = E = i\hbar \frac{1}{T} \frac{\partial T}{\partial t} \quad (5.45b)$$

We now have two equations

$$-\frac{\hbar^2}{2m} \frac{\partial^2 X}{\partial x^2} = EX \quad (5.45c)$$

$$i\hbar \frac{\partial T}{\partial t} = ET \quad (5.45d)$$

Equation 5.45d provides

$$T(t) = \beta(0) \exp\left(\frac{E}{i\hbar}t\right) = \beta(0) \exp(-i\omega t) \quad (5.46)$$

where

$\beta(0)$ is an integration constant

$E = \hbar\omega$ as usual

Separation of variables also provides boundary conditions for $X(x)$ as follows

$$\Psi(0, t) = 0 = \Psi(L, t) \rightarrow X(0)T(t) = 0 = X(L)T(t) \rightarrow X(0) = 0 = X(L) \quad (5.47)$$

Next, we look for the basis set $\{X_n(x)\}$.

The Sturm-Liouville (SL) system of equations for finding the energy basis functions X includes the ordinary differential equation from Equation 5.45c and the boundary conditions from Equation 5.47.

$$-\frac{\hbar^2}{2m} \frac{d^2 X}{dx^2} = EX \quad (5.48a)$$

$$X(0) = 0 = X(L) \quad (5.48b)$$

Notice that Equation 5.48a has the form of the eigenvector equation.

$$\hat{\mathcal{H}} X(x) = EX(x)$$

The Hamiltonian $\hat{\mathcal{H}}$, a Hermitian operator, is the total energy but in this case, the potential energy $V=0$ in the well, and the Hamiltonian reduces to the kinetic energy.

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

Three ranges for the separation constant E must be considered because the sign of E determines the character of the solution. We can find real exponentials, linear functions, or sines depending on whether $E < 0$, $E = 0$, $E > 0$, respectively. All cases must be considered because the solution wave function will be a summation over all eigenfunctions with the eigenvalues as the index for the summation. We must be sure to include all eigenvectors in the set so that the set will be complete as a basis. The $E < 0$, $E = 0$ cases lead to trivial solutions and not eigenvectors. For example, consider $E = 0$. Equation 5.48a becomes $-(\hbar^2/2m)(\partial^2 X/\partial x^2) = 0$ with the general solution $X = c_1 x + c_2$. The boundary conditions on X lead to $c_1 = c_2 = 0$ and therefore we find only the trivial solution $X = 0$. The trivial solution cannot be classified as an eigenfunction since it would require the wave function $\Psi \sim XT$ to be zero and that would imply that the particle does not exist. A similar result is obtained for the $E < 0$ case.

Now consider the case $E > 0$. The equation for $X(x)$ provides a solution of the form

$$X(x) = A'e^{-ikx} + B'e^{ikx} = A \cos(kx) + B \sin(kx) \quad (5.49a)$$

where

$$k = \sqrt{2mE/\hbar^2} \quad (5.49b)$$

This last equation comes from substituting Equation 5.49a into Equation 5.48a. We have three unknowns A , B , k and only two boundary conditions in Equation 5.44. Clearly, we will not find values for all three parameters. The boundary conditions lead to multiple discrete values for k and hence for the energy E .

Next, let us determine the parameters A , B , k as much as possible. The first boundary condition of $\psi(0, t) = 0$ requires $X(0) = 0$ and therefore Equation 5.49a provides

$$X(x) = B \sin(kx) \quad (5.49c)$$

Consider the second boundary condition $\psi(L, t) = 0$ which requires $X(L) = 0$. The case of $B = 0$ should be avoided if at all possible since then only the trivial solution would be obtained. Therefore, look for values of k that provide

$$\sin(kL) = 0 \quad (5.50a)$$

If such k 's cannot be found or perhaps only $k = 0$, then one must conclude that either $B = 0$ or $k = 0$ which produce only the trivial solution. Equation 5.50a holds when

$$k = n\pi/L \quad \text{for } n = 1, 2, 3, \dots \quad (5.50b)$$

and therefore the electron wavelength must be given by

$$\lambda = 2\pi/k = 2L/n$$

which requires multiples of half wavelengths to fit in the width of the well. One usually interprets the “electron wavelength” to be the same as the “wavelength” of the probability amplitude ψ . The functions

$$X_n(x) = B \sin\left(\frac{n\pi x}{L}\right) \quad (5.50c)$$

are the eigenfunctions of the Hamiltonian, which is the kinetic energy operator for our case with $V = 0$. The basis set comes from normalizing the eigenfunctions. We require $\langle X_n | X_n \rangle = 1$ so that Equation 5.50c therefore provides

$$B = \sqrt{\frac{2}{L}}$$

The energy basis set must be

$$\left\{ X_n(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) \right\} \quad (5.50d)$$

These are also called “stationary solutions” because they do not depend on time. Stationary solutions satisfy the “time-independent” Schrödinger equation $\hat{H}X_n(x) = E_n X_n(x)$. So, because “solving” the time-independent Schrödinger equation is the same as solving the Sturm-Liouville problem, one sees that the time-independent Schrödinger equation provides the basis set as expected.

A solution of the partial differential equation corresponding to an allowed energy E_n must be

$$\Psi_n = X_n T_n = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) \beta_n(0) e^{-itE_n/\hbar} \quad (5.51)$$

As for the allowed energies, Equation 5.49b provides

$$k = \sqrt{2mE/\hbar^2} \quad \text{and} \quad k = n\pi/L \quad \text{for } n = 1, 2, 3, \dots$$

Substituting for the values of k found in Equation 5.50b then yields

$$E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{\hbar^2 \pi^2}{2mL^2} n^2 \quad (5.52)$$

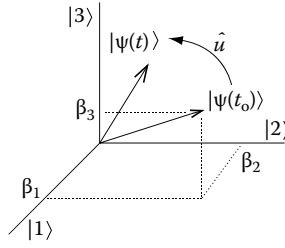


FIGURE 5.13 The full solution moves in Hilbert space which makes the components depend on time.

The full wave function must be a linear combination of these fundamental solutions

$$\Psi(x, t) = \sum_E \Psi_E = \sum_n T_n(t) X_n(t) \quad (5.53a)$$

which has the form of the summation over basis vectors with time-dependent components $T_n(t)$ (i.e., time-dependent probability amplitudes T_n). Substituting for X and T from Equations 5.46 and 5.50d, respectively, we find

$$\Psi(x, t) = \sum_n \beta_n(0) e^{-itE_n/\hbar} \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) \quad (5.53b)$$

The components of the vector must be

$$\beta_n(t) = \beta_n(0) e^{-itE_n/\hbar} \quad (5.53c)$$

where $\beta_n(0)$ are constants. The time-dependent components indicate motion in the Hilbert space as suggested by Figure 5.13.

Example 5.5

Suppose a student places an electron in the infinitely deep well at $t=0$ according to the prescription

$$\Psi(x, 0) = \frac{1}{\sqrt{L}} \sin\left(\frac{\pi x}{L}\right) + \frac{1}{\sqrt{L}} \sin\left(\frac{2\pi x}{L}\right) \quad (5.54)$$

The function $\Psi(x, 0)$ provides the initial condition and the reader should verify that it has unit length before proceeding. Find the full wave function.

SOLUTION

The full wave function appears in Equation 5.53

$$\Psi(x, t) = \sum_n \beta_n(0) \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) e^{-itE_n/\hbar} \quad (5.55)$$

We need the coefficients $\beta_n(0)$ which come from the wave function evaluated at the fixed time $t=0$. We have

$$\Psi(x, 0) = \sum_n \beta_n(0) \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi}{L}x\right) = \sum_n \beta_n(0) X_n(x)$$

We can find the coefficients by projecting the wave function onto the basis vectors

$$\beta_n(0) = \langle X_n | \Psi(x, 0) \rangle = \int_0^L dx X_n^*(x) \Psi(x, 0)$$

where $\Psi(x, 0)$ appears in Equation 5.54. Rather than do the integration, let us take a simple route for this problem. Notice that the initial condition can be written in terms of the basis vectors as

$$\Psi(x, 0) = \frac{1}{\sqrt{L}} \sin\left(\frac{\pi x}{L}\right) + \frac{1}{\sqrt{L}} \sin\left(\frac{2\pi x}{L}\right) = \frac{1}{\sqrt{2}} X_1(x) + \frac{1}{\sqrt{2}} X_2(x)$$

Therefore the expansion coefficients must have the form

$$\beta_n(0) = \langle X_n | \Psi(x, 0) \rangle = \langle X_n | \left\{ \frac{1}{\sqrt{2}} |X_1\rangle + \frac{1}{\sqrt{2}} |X_2\rangle \right\} \rangle = \frac{1}{\sqrt{2}} \delta_{1n} + \frac{1}{\sqrt{2}} \delta_{2n}$$

and the full wave function becomes

$$\Psi(x, t) = \sum_n \beta_n(0) \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) e^{-itE_n/\hbar} = \sum_n \left[\frac{1}{\sqrt{2}} \delta_{1n} + \frac{1}{\sqrt{2}} \delta_{2n} \right] \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) e^{-itE_n/\hbar}$$

which reduces to

$$\Psi(x, t) = \frac{1}{\sqrt{L}} \sin\left(\frac{\pi x}{L}\right) e^{-itE_1/\hbar} + \frac{1}{\sqrt{L}} \sin\left(\frac{2\pi x}{L}\right) e^{-itE_2/\hbar} = \frac{1}{\sqrt{2}} X_1 e^{-itE_1/\hbar} + \frac{1}{\sqrt{2}} X_2 e^{-itE_2/\hbar}$$

where Equation 5.52 gives

$$E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{\hbar^2 \pi^2}{2mL^2} n^2$$

Example 5.6

What is the probability of finding the particle in $n=2$ at time $t=1$ for the previous example?

SOLUTION

The full wave function has the form

$$\Psi(x, t) = \sum_n \beta_n(t) \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) = \sum_n \beta_n(t) X_n(x)$$

where $\beta_n(t) = \beta_n(0) e^{-itE_n/\hbar}$. At $t=1$, we find $\beta_n(1) = \beta_n(0) e^{-iE_n/\hbar}$. The probability is

$$P(n=2) = |\langle X_2 | \Psi(1) \rangle|^2 = \left| \left\langle X_2 \left| \sum_n \beta_n(1) X_n(x) \right. \right\rangle \right|^2 = |\beta_2(1)|^2$$

where Equation 5.53c provides $\beta_2(t) = \beta_2(0) e^{-itE_2/\hbar}$. Consequently, we find

$$P(n=2) = |\beta_2(0)|^2 = 0.5$$

Example 5.7

If the particles starts in the eigenstate X_1 at $t=0$, (a) find the probability that the electron will be found in the region $0 < x < L/2$ at $t=0$, (b) find the standard deviation σ_x , and (c) explain how a particle can be in an eigenstate and still have a nonzero variance σ_x^2 .

SOLUTION

(a) The wave function is $\psi(x, 0) = X_1(x)$ and the probability can be written as

$$P(0 < x < L/2) = \int_0^{L/2} dx \psi^* \psi = \int_0^{L/2} dx \frac{2}{L} \sin^2 \left(\frac{\pi x}{L} \right) = \frac{1}{2}$$

(b) The variance can be written as $\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$. The average position can be calculated as $\langle x \rangle = \langle \psi(x, 0) | x | \psi(x, 0) \rangle = \langle X_1(x) | x | X_1(x) \rangle = \int_0^L dx X_1(x) x X_1(x) = L/2$ and average of x^2 is $\langle x^2 \rangle = \frac{2\pi^2 - 8}{\pi^3} L^2$. The variance is approximately $\sigma_x^2 = 0.128L^2$ and the standard deviation is $\sigma_x = 0.36L$.

(c) The particle is in an energy eigenstate not a coordinate eigenstate.

5.3.3 FINITELY DEEP SQUARE WELL

The case of the finitely deep square well appears in Figure 5.14. The finite barrier heights significantly complicates the solution by dividing space into three regions. Each region requires a solution and then all three solutions must be made to agree at the two barriers through boundary conditions in addition to the behavior at infinite distances from the well. Once we find the general superposition of basis states, then the initial conditions can be applied. Assume that the potential energy $V(x)$ has the form given by

$$V(x) = \begin{cases} V_b & x < 0 \\ 0 & 0 < x < L \\ V_b & x > L \end{cases}$$

where the well has width L and barrier height V_b .

The SWE for this 1-D case can be written as

$$\hat{H}\Psi(x, t) = i\hbar \frac{\partial}{\partial t} \Psi(x, t) \quad \text{or} \quad -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x, t)}{\partial^2 x} + V(x)\Psi(x, t) = i\hbar \frac{\partial \Psi(x, t)}{\partial t}$$

In addition to the partial differential equation, we also need boundary and initial conditions. We require the wave functions Ψ to approach 0 for very large distances $x \rightarrow \pm\infty$. However, because we will solve the time-independent Schrödinger equation for three separate regions, namely ($x < 0$, $0 < x < L$, $L < x$), we need boundary conditions at the $x=0$ and $x=L$ interfaces. We will assume that the wave function and its first derivative are both continuous across each interface.

$$\begin{aligned} \Psi(0^-, t) &= \Psi(0^+, t) & \Psi(L^-, t) &= \Psi(L^+, t) \\ \frac{d}{dx} \Psi(0^-, t) &= \frac{d}{dx} \Psi(0^+, t) & \frac{d}{dx} \Psi(L^-, t) &= \frac{d}{dx} \Psi(L^+, t) \end{aligned}$$

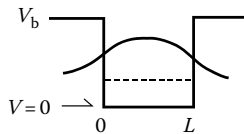


FIGURE 5.14 Lowest energy level for the finitely deep well.

where the superscripts $+$, $-$ stand for “slightly greater than” and “slightly less than,” respectively. We will see in upcoming chapters that the probability-current density must be continuous across interfaces. The continuity of the current density represents an area of research especially important for physical heterostructure. However at present, we consider the potentials in free space independent of matter.

We separate variables in Schrödinger’s wave equation using $\Psi(x, t) = X(x)T(t)$ to find

$$-\frac{\hbar^2}{2m} \frac{\partial^2 X_n(x)}{\partial^2 x} + V(x)X_n(x) = E_n X_n(x), \quad \frac{\partial T_n(t)}{\partial t} = \frac{E_n}{i\hbar} T_n(t) \quad (5.56)$$

Although we will consider three separate regions, there exists only one eigenfunction X_n and one eigenvalue E_n for each integer n . The eigenvalue E_n must be independent of the coordinate x . The solutions in the three separate regions must combine to give a single function X_n for each integer.

$$X_n(x) = \begin{cases} X_n^<(x) & \text{for } x < 0 \\ X_n^=(x) & \text{for } 0 < x < L \\ X_n^>(x) & \text{for } L < x \end{cases}$$

Notice the superscripts of $<$, $=$, and $>$ indicate the region to which a given function X applies.

Region 1: $0 < x < L$

The time-independent Schrödinger equation is

$$\frac{\partial^2 X_n^=(x)}{\partial^2 x} + \frac{2mE_n}{\hbar^2} X_n^=(x) = 0$$

We only consider the case $\frac{2mE_n}{\hbar^2} > 0$ since the energy of the particle must be larger than 0 at the bottom of the well. Setting $k_n^2 = \frac{2mE_n}{\hbar^2}$, we find

$$X_n^=(x) = B_n \cos(k_n x) + C_n \sin(k_n x) \quad (5.57)$$

The two constants B_n and C_n will be determined after considering the other two regions. As usual, there will be one remaining constant (A_n in this case) which can be determined by normalizing the wave function to 1. We cannot determine the energy levels E_n until first finding the eigenfunction $X_n(x)$. The value of k_n will differ from that of the infinitely deep well since the wave no longer needs to fit exactly in the length L .

Region 2: $x < 0$

The time-independent Schrödinger equation for this region can be written as

$$\frac{\partial^2 X_n^<(x)}{\partial^2 x} - \frac{2m(V_b - E_n)}{\hbar^2} X_n^<(x) = 0$$

Again we consider only the case, namely $\frac{2m(V_b - E_n)}{\hbar^2} > 0$ since we want the confined electron to have less energy than the top of the barrier ($V_b > E_n$) otherwise the electron would not be confined to the well. Defining $K_n^2 = \frac{2m(V_b - E_n)}{\hbar^2}$, we find

$$X_n^<(x) = A_n e^{K_n x} + A'_n e^{-K_n x}$$

(notice the capital K used for the wave vector). However, the fact that $X \rightarrow 0$ as $x \rightarrow -\infty$ requires that $A'_n = 0$. Therefore for this region, we have

$$X_n^<(x) = A_n e^{K_n x} \quad (5.58)$$

Region 3: $x > L$

For this region, the time-independent Schrödinger equation is identical to that for region 2. The boundary condition $X \rightarrow 0$ as $x \rightarrow \infty$ produces a function of the form

$$X_n^>(x) = D_n e^{-K_n(x-L)} \quad (5.59)$$

where to simplify later work, we have included the L in the argument of the exponential.

We must combine all the individual solutions for the three regions into the one eigenvector X_n . This means that we must determine all of the constants using the remaining boundary conditions. The boundary condition $\Psi(0^-, t) = \Psi(0^+, t)$ provides $X_n^<(0) = X_n^=(0)$ or equivalently (from Equations 5.57 and 5.58)

$$A_n e^{K_n 0} = B_n \cos(k_n 0) + C_n \sin(k_n 0) \rightarrow A_n = B_n \quad (5.60)$$

The boundary condition $\Psi(L^-, t) = \Psi(L^+, t)$ provides $X_n^=(L) = X_n^>(L)$ or using Equation 5.57 (with $A_n = B_n$) and Equation 5.59

$$A_n \cos(k_n L) + C_n \sin(k_n L) = D_n \quad (5.61)$$

The boundary condition $\frac{d}{dx}\Psi(0^-, t) = \frac{d}{dx}\Psi(0^+, t)$ provides $\frac{d}{dx}X_n^<(0) = \frac{d}{dx}X_n^=(0)$ or, using Equations 5.58 and 5.57, we find

$$A_n K_n e^{K_n 0} = -B_n k_n \sin(k_n 0) + C_n k_n \cos(k_n 0) \rightarrow C_n = A_n K_n / k_n \quad (5.62)$$

Finally the remaining boundary condition $\frac{d}{dx}\Psi(L^-, t) = \frac{d}{dx}\Psi(L^+, t)$ provides $\frac{d}{dx}X_n^=(L) = \frac{d}{dx}X_n^>(L)$ or equivalently (after using Equations 5.60 and 5.62)

$$A_n [-k_n \sin(k_n L) + K_n \cos(k_n L)] = -D_n K_n \quad (5.63)$$

We must solve three equations (not much fun)

$$A_n \cos(k_n L) + C_n \sin(k_n L) = D_n \quad (5.64a)$$

$$C_n = A_n K_n / k_n \quad (5.64b)$$

$$A_n [-k_n \sin(k_n L) + K_n \cos(k_n L)] = -D_n K_n \quad (5.64c)$$

Combining the first and second, and repeating the third gives the following set

$$A_n \left[\cos(k_n L) + \frac{K_n}{k_n} \sin(k_n L) \right] = D_n \quad (5.65a)$$

$$A_n [-k_n \sin(k_n L) + K_n \cos(k_n L)] = -D_n K_n \quad (5.65b)$$

Eliminating D between the two equations yields

$$\cos(k_n L) + \frac{K_n}{k_n} \sin(k_n L) = \frac{k_n}{K_n} \sin(k_n L) - \cos(k_n L) \quad (5.66)$$

Notice that we were unable to solve for A ; we can find this one by normalizing the final eigenfunction. Solving this last equation for $\tan(kL)$ gives us

$$\tan(k_n L) = \frac{2k_n K_n}{k_n^2 - K_n^2} \quad (5.67)$$

Both k and K depend on the eigenvalues E_n . Let us drop the n subscript for simplicity.

$$\tan(kL) = \frac{2kK}{k^2 - K^2} \quad (5.68)$$

As will be discussed next, solving Equation 5.68 for k, K provides the allowed energies in the well. One way to see this is to write both k and K in terms of E and keep in mind that E is independent of position and hence independent of regions 1, 2, and 3. The solutions E_n will then give the energies of the modes that can then be used to find the wave functions (composed of three parts). We can rewrite k and K in terms of E , or we can write $K^2 = \frac{2m(V_b - E)}{\hbar^2}$ in terms of k . We choose the latter method and find the allowed values of k and then find the allowed values of E through $k_n^2 = \frac{2mE_n}{\hbar^2}$.

$$K^2 = \frac{2m(V_b - E)}{\hbar^2} = \frac{2m}{\hbar^2} V_b - k^2 = k_m^2 - k^2 \quad (5.69a)$$

where we have defined a new symbol

$$k_m^2 = 2mV_b/\hbar^2 \quad (5.69b)$$

(for simplicity) that represents the maximum value for k since we must keep $E < V_b$ in order for the electron to remain bound in the well. Equation 5.68 becomes

$$\tan(kL) = \frac{2k\sqrt{k_m^2 - k^2}}{2k^2 - k_m^2} \quad (5.70a)$$

We know V_b and therefore also $k_m^2 = 2mV_b/\hbar^2$.

The allowed values of k in Equation 5.70a can be found by plotting both sides on the same set of axes. It is easiest to define a two new parameters $z = kL$ and the maximum value of $z_m = k_m L$. Equation 5.70a becomes

$$\tan(z) = \frac{2z\sqrt{z_m^2 - z^2}}{2z^2 - z_m^2} \quad (5.70b)$$

Now plot the two sides of this last equation on the same set of axes as

$$F(z) = \tan(z) \quad (5.71a)$$

$$G(z) = \frac{2z\sqrt{z_m^2 - z^2}}{2z^2 - z_m^2} \quad (5.71b)$$

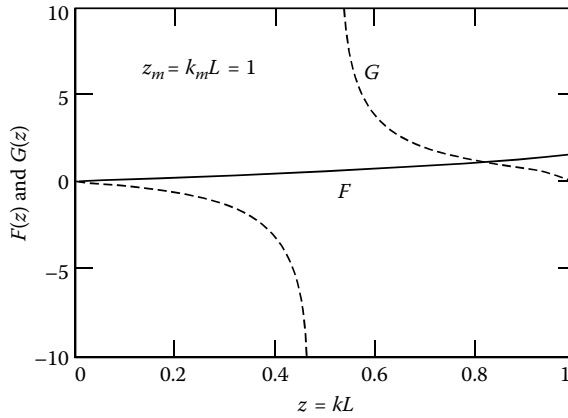


FIGURE 5.15 Plot of Equation 5.71 for $z_m = k_m L = 1$ shows only one intersection point at $k_1 = 0.814/L$.

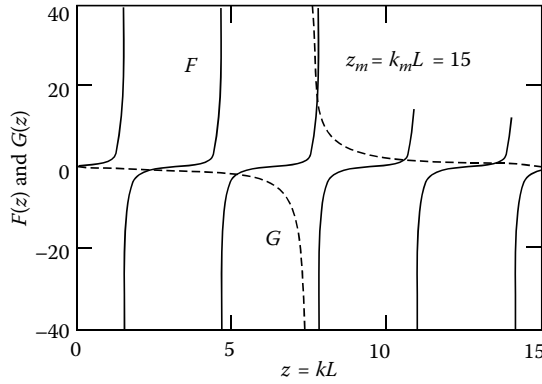


FIGURE 5.16 Plot of Equation 5.71 for $x_m = k_m L = 15$ shows five intersection points to produce k_1 through k_5 .

and find the intersection points. Although k_m and L have not been specified, we can still describe how the solution proceeds. Figures 5.15 and 5.16 show Equation 5.71 plotted on the same set of axes. For $k_m L = 1$ or equivalently $V_b = \hbar^2/(2mL^2)$ (Equation 5.69b), the single allowed k is $k_1 = 0.814/L$. The equations

$$K^2 = \frac{2m}{\hbar^2}(V_b - E) = k_m^2 - k^2, \quad k^2 = \frac{2mE}{\hbar^2}, \quad z = kL, \quad z_m = k_m L \quad (5.72)$$

providing an energy level of $E_1 = \hbar^2 k_1^2/(2m)$ and $K_1 = 0.581/L$. Similarly, Figure 5.16 shows the case of $k_m L = 15$ produces five values of kL (2.59, 5.13, 7.82, 10.5, 13.2) which produces five energy levels E_n in the well (see the problem set for more).

Now we know E_n , k_n , and K_n . We still need to find the normalization constant A_n . The eigenfunction X_n is

$$X_n(x) = \begin{cases} A_n e^{K_n x} & x < 0 \\ B_n \cos(k_n x) + C_n \sin(k_n x) & 0 < x < L \\ D_n e^{-K_n(x-L)} & L < x \end{cases}$$

or, substituting the values for the coefficients

$$X_n(x) = \begin{cases} A_n e^{K_n x} & x < 0 \\ A_n \cos(k_n x) + A_n \frac{K_n}{k_n} \sin(k_n x) & 0 < x < L \\ A_n \left[\cos(k_n L) + \frac{K_n}{k_n} \sin(k_n L) \right] e^{-K_n(x-L)} & L < x \end{cases} \quad (5.73)$$

The value of A_n can be found by requiring $1 = \langle X_n | X_n \rangle = \int_{-\infty}^{\infty} dx X_n^* X_n$. However, we must divide the integral into three pieces to match the three regions in the definition of X .

$$1 = \langle X_n | X_n \rangle = \int_{-\infty}^0 dx X_n^* X_n + \int_0^L dx X_n^* X_n + \int_L^{\infty} dx X_n^* X_n$$

We can drop the complex conjugate since the functions are all real. We must calculate

$$A_n^2 = \left\{ \int_{-\infty}^0 dx e^{2K_n x} + \int_0^L dx \left[\cos(k_n x) + \frac{K_n}{k_n} \sin(k_n x) \right]^2 + \int_L^{\infty} dx \left[\cos(k_n L) + \frac{K_n}{k_n} \sin(k_n L) \right]^2 e^{-2K_n(x-L)} \right\}^{-1} \quad (5.74)$$

The integrals depend on the values K_n and k_n which in turn depend on the value of V_b . For the case of the single energy level with $z_m = k_m L = 1$ as discussed above, substituting for K_1, k_1 provides the normalization as $A_1 = 0.536/\sqrt{L}$.

The wave function describing the situation can be written as

$$\Psi(x, t) = \sum_n T_n(t) X_n(x)$$

where T has the form

$$T_n(t) = \beta_n e^{-i \frac{E_n}{\hbar} t}$$

so that the

$$\Psi(x, t) = \sum_n \beta_n X_n(x) e^{-i \frac{E_n}{\hbar} t}$$

The constants β are found by knowing the wave function at some time—usually at $t = 0$. Usually the problem specifies the starting wave function $\Psi(x, 0)$. We must have

$$\Psi(x, 0) = \sum_n \beta_n X_n(x)$$

which we recognize as a generalized summation of basis functions in a Hilbert space. The β are therefore given by

$$\beta_n = \langle X_n(x) | \Psi(x, 0) \rangle$$

5.4 HARMONIC OSCILLATOR

The SWE describes the time evolution of the wave function. The Hamiltonian for the harmonic oscillator describes a particle of mass m in a quadratic potential. Displacing the mass from equilibrium produces a linear restoring force. We focus on the 1-D oscillator since a 3-D oscillator can be decomposed into three 1-D oscillators. Any coupling between the three 1-D oscillators can be included in the Hamiltonian later if desired (Figure 5.17).

The harmonic oscillator has important applications. Many systems have nonlinear potential functions. Expanding these nonlinear potentials in a Taylor series often produces a quadratic term as the lowest order approximation after the constant and linear ones. As an example, the periodic motion of atoms about their equilibrium position can be modeled with the quadratic potential. We know this motion must be related to phonons moving through the material. We will see that the zero point motion of the atom can be described by the quantum mechanical vacuum state.

The quadratic potential has surprising applications to electromagnetic theory in the area of quantum optics. The electromagnetic fields can be modeled by quadratic kinetic and potential terms. This section is best explored in the book *Physics of Optoelectronics* or in one devoted to quantum optics. For the present section, one should realize that the formalism associated with the harmonic oscillator has wide ranging applications.

5.4.1 INTRODUCTION TO CLASSICAL AND QUANTUM HARMONIC OSCILLATORS

For a harmonic oscillator, the quadratic potential produces a linear restoring force

$$V = \begin{cases} \frac{1}{2}kx^2 & \text{1-D} \\ \frac{1}{2}k|\vec{r}|^2 & \text{3-D} \end{cases}$$

where the equilibrium position occurs at the origin $x=0$, the “spring constant” must be positive $k > 0$ and it describes the curvature of the potential (i.e., magnitude of the force), and $|\vec{r}|^2 = x^2 + y^2 + z^2$.

The classical (1-D) Hamiltonian has the form

$$H_c = \frac{p^2}{2m} + \frac{1}{2}kx^2 \quad (5.75)$$

where one considers the dynamic variables x, p to be independent of one another. Newton’s second law can be demonstrated using Hamilton’s canonical equation (refer to Section 5.2).

$$\dot{p} = -\frac{\partial H_c}{\partial x} = -kx = F$$

The Lagrangian shows that the momentum p must be related to the velocity by $p = mv = m\dot{x}$.

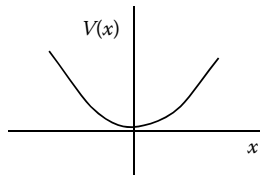


FIGURE 5.17 The quadratic potential.

We want to compare and contrast solutions for position (as a function of time t) between the classical and quantum harmonic oscillators. The classical Hamiltonian (the total energy) can be rewritten using Equation 5.75 and $p = m\dot{x}$

$$\frac{m}{2} \left(\frac{dx(t)}{dt} \right)^2 + \frac{1}{2} k(x(t))^2 = E \quad (5.76)$$

where

E represents the total energy of the oscillator

$x(t)$ represents the position of the electron parameterized by the time t

The solution has the form

$$x(t) = A \sin(\omega_0 t) \quad (5.77a)$$

The formula $\omega_0^2 = k/m$ relates the angular frequency of oscillation ω_0 to the “spring constant” k . Note, do not confuse this k with the wave vector used in previous sections. Substituting Equation 5.77a into Equation 5.76 provides

$$A = \sqrt{\frac{2E}{k}} = \sqrt{\frac{2E}{m\omega_0^2}} \quad (5.77b)$$

The amplitude A represents the points on the potential plot $V(x)$ where the kinetic energy becomes zero (see Figure 5.18)

$$E = \frac{1}{2} kx^2 \Big|_{x=A} \rightarrow A = \sqrt{\frac{2E}{k}}$$

Classically, the particle can only be found in the region $x \in [-A, A]$ and never outside that region. The probability density ρ (i.e., probability per unit x -distance) for finding the particle at a point x appears similar to a delta function near the endpoints of the motion; this behavior occurs because the particle slows down near those points and spends more time there.

Several differences exist between the classical and quantum mechanical harmonic oscillators. Figure 5.19 shows an example quantum mechanical solution to Schrödinger’s equation with the

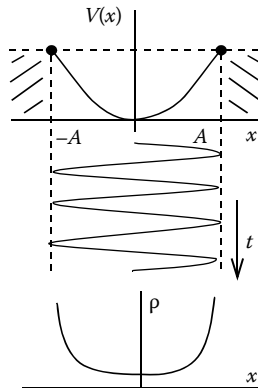


FIGURE 5.18 Motion of a harmonic oscillator. The probability density ρ shows the most likely position of finding the mass m is at the turning points where the oscillator momentarily comes to rest.

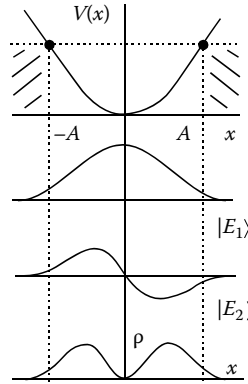


FIGURE 5.19 The first two quantum mechanical solutions to the harmonic oscillator. The probability density ρ for finding the particle at point x does not resemble the classical one.

quadratic potential. Unlike the classical particle, the quantum particle can be found in the classically forbidden region. The figure shows how the wave function exponentially decays in these classically forbidden regions. Classically, the particle does not have enough energy to enter the forbidden region. The basis functions have the form

$$\phi_n(x) = \left(\frac{\alpha}{\pi^{1/2} n! 2^n} \right)^{1/2} H_n(\alpha x) \exp\left(-\frac{\alpha^2 x^2}{2}\right) \quad (5.78)$$

where $\alpha^4 = (m\omega_0/\hbar)^2$. The exponential part of the solution ensures that the wave function decreases in the classically forbidden region. The Hermite polynomials H_n primarily control the behavior in the classically allowed region near the center. They can be conveniently found from a “generating function” according to

$$H_n(\xi) = (-1)^n \exp(\xi^2) \frac{d^n}{d\xi^n} \exp(-\xi^2) \quad (5.79a)$$

where $\xi = \alpha x$. The first three Hermite polynomials are

$$H_0(\xi) = 1 \quad H_1(\xi) = 2\xi \quad H_2(\xi) = 4\xi^2 - 2 \quad (5.79b)$$

Continuing with Figure 5.19, perhaps most striking of all, the probability density function for the quantum particle decreases to zero near the endpoints of motion and reaches its peak value (or values) near the center of the classical region $[-A, A]$. However, the classical probability of finding the classical particle assumes its minimum value near the origin.

Here is another difference between the classical and harmonic oscillator solutions. The classical oscillator energy can be increased by applying a driving force that increases the oscillation amplitude, and hence the energy $E = A^2 m \omega_0^2 / 2$. The angular oscillation frequency $\omega_0 = \sqrt{k/m}$ remains constant for a fixed spring constant k . The energy of the quantum oscillator increases also by absorbing energy

$$E_n = \hbar \omega_n = \hbar \omega_0 \left(n + \frac{1}{2} \right) \quad n = 0, 1, 2, \dots \quad (5.80)$$

The integer n can be interpreted as either the “basis function number” or as the number of quanta stored in the motion. The quantity ω_n of the quantum oscillator changes even though the value

angular frequency ω_0 remains fixed. The angular frequency does not appear to refer to the rate at which the quantum particle bounces from side to side. In the present case of Figure 5.19, we view the quantum particle as a stationary wave function. Larger numbers of quanta n result in larger “displacements” from equilibrium meaning the probability density has more peaks that move closer to the classically forbidden region.

We find similar plots for quantized electromagnetic (EM) waves. The energy of an EM oscillator (the EM waves) can be changed by changing the angular frequency (or wavelength) or by changing the amplitude (i.e., the number of quanta in the mode). Studies in quantum optics show that the “position x ” and “momentum p ” become the “in-phase” and “out-of-phase” electric fields. Therefore, the wave functions in the EM case describe the probability of finding a particular value of the electric field.

5.4.2 HAMILTONIAN FOR THE QUANTUM HARMONIC OSCILLATOR

The quantum mechanical Hamiltonians come from the classical ones by replacing the dynamical variables x, p with the corresponding operators $\hat{x}, \hat{p}$ in $H_c = p^2/2m + kx^2/2$ to find

$$\hat{\mathcal{H}}|\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad \left(\frac{\hat{p}^2}{2m} + \frac{1}{2} k \hat{x}^2 \right) |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad (5.81)$$

Operating with the “coordinate” projection operator $\langle x|$ produces $\hat{x} \rightarrow x$ and $\hat{p} \rightarrow \frac{\hbar}{i} \frac{\partial}{\partial x}$ (refer to Appendix L) to obtain the Schrödinger equation

$$\hat{\mathcal{H}} \Psi(x, t) = i\hbar \frac{\partial \Psi(x, t)}{\partial t} \quad \text{or} \quad \left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2} k x^2 \right) \Psi(x, t) = i\hbar \frac{\partial}{\partial t} \Psi(x, t) \quad (5.82a)$$

The boundary conditions for the SWE for the harmonic oscillator require the wave function to approach zero as x goes to infinity

$$\Psi(x \rightarrow \pm\infty, t) \rightarrow 0 \quad (5.82b)$$

We consider two methods for solving the Schrödinger equation for the harmonic oscillator. The first method uses a power series solution, which becomes very algebraically involved. The solution starts by separating variables in Equation 5.82a and using a power series to find the solutions to the Sturm-Liouville problem (the eigenvector problem). The second method uses the linear algebra of raising and lowering operators. We present the method of raising and lowering operators (commonly referred to as the algebraic approach). We will find the stationary solutions given in Equation 5.78 and the energy eigenvalues in Equation 5.80.

5.4.3 INTRODUCTION TO THE LADDER OPERATORS FOR THE HARMONIC OSCILLATOR

The operator approach (i.e., algebraic approach) to solving Schrödinger’s equation for the harmonic oscillator is simpler than the power series approach. In addition, it provides a great deal of insight into the mathematical structure of the quantum theory. The algebraic approach uses “raising” $\hat{a}^+$ and “lowering” $\hat{a}$ operators (i.e., ladder operators, or sometimes called promotion and demotion operators). We will later rewrite the Hamiltonian in terms of the raising and lowering operators in the form of the number operator $\hat{N} = \hat{a}^+ \hat{a}$.

The raising and lowering operators map one basis vector into another one. If one has the basis set

$$\{ |0\rangle, |1\rangle, |2\rangle \dots \} \quad (5.83)$$

then the raising operator $\hat{a}^+$ maps $|n\rangle$ into $|n+1\rangle$ and $\hat{a}$ maps $|n+1\rangle$ into $|n\rangle$. For the “lowest” basis vector, the lowering operator cannot produce a “lower” basis vector and so one requires $\hat{a}|0\rangle = 0$. For the Harmonic oscillator, the raising and lowering operators are adjoints of one another and have a specific normalization so as to make the Hamiltonian writeable in the form $\hat{\mathcal{H}} = \hbar\omega_0\hat{N} = \hbar\omega_0\hat{a}^+\hat{a}$. Where N is the so-called number operator (more on this later). The raising and lowering operators for the harmonic oscillator map the basis vectors according to

$$\hat{a}^+|n\rangle = \sqrt{n+1}|n+1\rangle \quad \hat{a}|n\rangle = \sqrt{n}|n-1\rangle \tag{5.84a}$$

as depicted by Figure 5.20, where one can show that the n take on values of $0, 1, 2, \dots$. The lowering operator produces zero when operating on the lowest possible basis state (i.e., the “vacuum state”) $\hat{a}|0\rangle = 0$. Examples for other raising operators (not harmonic oscillator ones) appear in Figure 5.21. One can interpret the effect of the raising operator as promoting an electron in state 1 to state 2.

The number operator has two interpretations for the harmonic oscillator. First, the number operator provides the basis state number, which corresponds to an energy state of a particle. The interpretation is based on the relation $\hat{N}|n\rangle = n|n\rangle$ that results from an application of Equation 5.84a

$$\hat{N}|n\rangle = \hat{a}^+\hat{a}|n\rangle = \sqrt{n}\hat{a}^+|n-1\rangle = n|n\rangle \tag{5.84b}$$

The number operator therefore tells us the number of the eigenstate occupied by a particle. Second, the number operator provides the number of energy quanta in the system as its second interpretation. A particle occupying one of the energy basis states $|n\rangle \in B_v = \{|0\rangle = |E_0\rangle, |1\rangle = |E_1\rangle, \dots\}$ has n quanta of energy according to $E_n = \hbar\omega_0(n + 1/2)$. Therefore the vacuum state $|0\rangle$ corresponds to a particle state without any quanta of energy $n=0$. The value of n being the number of quanta

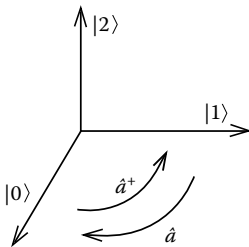


FIGURE 5.20 Raising and lowering operators move the harmonic oscillator from one state to another.

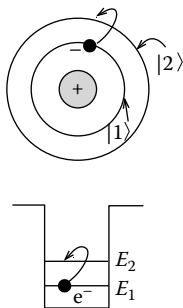


FIGURE 5.21 Physical examples showing the effect of a raising operators defined for an atom (top) and square well (bottom) rather than for the harmonic oscillator.

provides the concept of ordering basis vectors from “lowest” to “higher” – the ordering is determined by the number of quanta. Interestingly however, the vacuum state $|0\rangle$ still has an energy $E_0 = \hbar\omega_0/2$ which corresponds to zero point motion of atoms in a solid, for example. The atoms continue to move even though all of the extractable energy has been removed (i.e., $n=0$). Absolute zero can never be achieved since it is a classical concept corresponding to stationary atoms. Studies in quantum optics indicate that the electric field also experiences vacuum fluctuations; these fluctuations produce spontaneous emission from an ensemble of excited atoms.

In the next few sections, we wish to find the energy eigenvectors $B_V = \{|n\rangle\}$ and eigenvalues E_n for the harmonic oscillator. At times, it is convenient to use other notation for vectors in the energy basis set B_V to emphasize the relation between the basis vector and the eigenvalues E_n or the functional form $\phi_n(x)$ as in $B_V = \{|0\rangle = |E_0\rangle = |\phi_0\rangle, |1\rangle = |E_1\rangle = |\phi_1\rangle, \dots\}$. We assume nondegenerate eigenvalues E_n , which means that for each energy E_n , there corresponds exactly one eigenstate ϕ_n satisfying $\hat{H}|\phi_n\rangle = E_n|\phi_n\rangle$. We further assume an order for the energy levels $E_0 < E_1 < E_2 < \dots$. The operator approach must reproduce the results found with the power series approach.

We first show how the Hamiltonian incorporates the raising–lowering operators. We briefly discuss the mathematical description of the ladder operators and demonstrate the origin of their normalization constant. We then easily solve for the energy eigenvalues and eigenvectors. Two nice and important aspects of the operator approach are that it applies to many diverse systems without needing to rederive all of the results and it is easy to remember the results.

5.4.4 LADDER OPERATORS IN THE HAMILTONIAN

The Hamiltonian for the harmonic oscillator is

$$\hat{H} = \frac{\hat{p}^2}{2m} + \frac{m\omega_0\hat{x}^2}{2} \quad (5.85)$$

After considerable thought, one defines the lowering $\hat{a}$ and the raising $\hat{a}^+$ operators in terms of the position $\hat{x}$ and momentum operator $\hat{p}$.

$$\hat{a} = \frac{m\omega_0}{\sqrt{2m\hbar\omega_0}}\hat{x} + \frac{i\hat{p}}{\sqrt{2m\hbar\omega_0}} \quad (5.86a)$$

$$\hat{a}^+ = \frac{m\omega_0}{\sqrt{2m\hbar\omega_0}}\hat{x} - \frac{i\hat{p}}{\sqrt{2m\hbar\omega_0}} \quad (5.86b)$$

The operators defined by Equations 5.86 will later be seen as correctly defining the ladder operators. The raising operator in Equation 5.86b comes from taking the adjoint of the lowering operator in Equation 5.86a and using the fact that both $\hat{x}, \hat{p}$ must be Hermitian since they correspond to observables. Notice that the raising and lowering operators are not Hermitian $\hat{a} \neq \hat{a}^+$. These two equations for the lowering and raising operators can be solved for the position and momentum operators to find

$$\hat{x} = \sqrt{\frac{\hbar}{2m\omega_0}}(\hat{a} + \hat{a}^+) \quad \hat{p} = -i\sqrt{\frac{m\omega_0\hbar}{2}}(\hat{a} - \hat{a}^+) \quad (5.87)$$

The algebraic approach to finding the eigenfunctions and eigenvalues requires the Hamiltonian be written in terms of the ladder operators. One must first determine the commutation relations. We can

demonstrate that the raising operator commutes with itself as does the lowering operator while the raising operator does not commute with the lowering operator

$$[\hat{a}, \hat{a}] = 0 = [\hat{a}^+, \hat{a}^+] \quad [\hat{a}, \hat{a}^+] = 1 \quad (5.88)$$

These last two relations can be proven using the commutation relations between the position and momentum operators

$$[\hat{x}, \hat{x}] = 0 = [\hat{p}, \hat{p}] \quad [\hat{x}, \hat{p}] = i\hbar \quad (5.89)$$

We prove $[\hat{a}, \hat{a}^+] = 1$ by first substituting Equation 5.56.

$$[\hat{a}, \hat{a}^+] = \left[\frac{m\omega_0}{\sqrt{2m\hbar\omega_0}} \hat{x} + \frac{i\hat{p}}{\sqrt{2m\hbar\omega_0}}, \frac{m\omega_0}{\sqrt{2m\hbar\omega_0}} \hat{x} - \frac{i\hat{p}}{\sqrt{2m\hbar\omega_0}} \right]$$

Distributing the terms provides

$$[\hat{a}, \hat{a}^+] = \left(\frac{m\omega_0}{\sqrt{2m\hbar\omega_0}} \right)^2 [\hat{x}, \hat{x}] + \frac{[\hat{p}, \hat{p}]}{2m\hbar\omega_0} + \frac{i}{2\hbar} [\hat{p}, \hat{x}] - \frac{i}{2\hbar} [\hat{x}, \hat{p}]$$

Substituting the commutation relations from Equation 5.89, we find the desired results

$$[\hat{a}, \hat{a}^+] = 0 + 0 + \frac{i}{2\hbar} (-i\hbar) - \frac{i}{2\hbar} (i\hbar) = 1$$

As a side comment for the case of an ensemble of independent harmonic oscillators, each one has its own degrees of freedom $\hat{x}_i, \hat{p}_i$ that obey their own commutation relations.

$$[\hat{x}_i, \hat{x}_j] = 0 = [\hat{p}_i, \hat{p}_j] \quad [\hat{x}_i, \hat{p}_j] = i\hbar\delta_{ij}$$

As a result, there will be raising and lowering operators for each oscillator that satisfy

$$[\hat{a}_i, \hat{a}_j] = 0 = [\hat{a}_i^+, \hat{a}_j^+] \quad [\hat{a}_i, \hat{a}_j^+] = \delta_{ij}$$

Suppose there are two independent oscillators so that $i = 1$ and $j = 2$. The energy basis vectors for these independent degrees of freedom have the form $|m\rangle_1 |n\rangle_2 = |m, n\rangle$. Then for example, as discussed for direct product spaces in Chapter 3, one finds

$$\hat{a}_2^+ |m, n\rangle = |m\rangle_1 \hat{a}_2^+ |n\rangle_2 = |m\rangle_1 \sqrt{n+1} |n+1\rangle_2 = \sqrt{n+1} |m, n+1\rangle$$

Using the definitions of the position and momentum operators in Equation 5.87, the Hamiltonian for the single harmonic oscillator in Equation 5.81 can be rewritten as

$$\hat{\mathcal{H}} = \frac{\hat{p}^2}{2m} + \frac{1}{2} m\omega_0^2 \hat{x}^2 = \frac{1}{2m} \left[-i\sqrt{\frac{m\omega_0\hbar}{2}} (\hat{a} - \hat{a}^+) \right]^2 + \frac{1}{2} m\omega_0^2 \left[\sqrt{\frac{\hbar}{2m\omega_0}} (\hat{a} + \hat{a}^+) \right]^2 \quad (5.90a)$$

Squaring the constants provides

$$\hat{\mathcal{H}} = -\frac{\hbar\omega_0}{4} (\hat{a} - \hat{a}^+)^2 + \frac{\hbar\omega_0}{4} (\hat{a} + \hat{a}^+)^2$$

Squaring the operators and taking care not to commute them gives us

$$\hat{\mathcal{H}} = \frac{\hbar\omega_0}{4} \{-\hat{a}^2 + \hat{a}\hat{a}^+ + \hat{a}^+\hat{a} - \hat{a}^{+2} + \hat{a}^2 + \hat{a}\hat{a}^+ + \hat{a}^+\hat{a} + \hat{a}^{+2}\}$$

Combining the squared terms

$$\hat{\mathcal{H}} = \frac{\hbar\omega_0}{2} \{\hat{a}\hat{a}^+ + \hat{a}^+\hat{a}\} \quad (5.90b)$$

We must always use commutation relations to change the order of operators. Finally, by using the commutation relation $[\hat{a}, \hat{a}^+] = 1 \rightarrow \hat{a}\hat{a}^+ = 1 + \hat{a}^+\hat{a}$, the Hamiltonian becomes

$$\hat{\mathcal{H}} = \frac{\hbar\omega_0}{2} \{\hat{a}\hat{a}^+ + \hat{a}^+\hat{a}\} = \frac{\hbar\omega_0}{2} \{2\hat{a}^+\hat{a} + 1\}$$

As a result, the Hamiltonian for the single harmonic oscillator can be written as

$$\hat{\mathcal{H}} = \hbar\omega_0 \left(\hat{a}^+\hat{a} + \frac{1}{2}\right) \quad (5.91a)$$

We can define the number operator $\hat{N} = \hat{a}^+\hat{a}$ and rewrite Equation 5.91a as

$$\hat{\mathcal{H}} = \hbar\omega_0 \left(\hat{N} + \frac{1}{2}\right) \quad (5.91b)$$

5.4.5 PROPERTIES OF THE RAISING AND LOWERING OPERATORS

Next, we demonstrate the relations

$$\hat{a}^+|n\rangle = \sqrt{n+1}|n+1\rangle \quad \hat{a}|n\rangle = \sqrt{n}|n-1\rangle$$

by first showing $|n-1\rangle \sim \hat{a}|n\rangle$ and $|n+1\rangle \sim \hat{a}^+|n\rangle$ are eigenvectors of the number operator $\hat{N}$ corresponding to the eigenvalues $n-1$ and $n+1$, respectively. We next find the constants of proportionality. We will need two commutation relations. Using $[\hat{A}\hat{B}, \hat{C}] = \hat{A}[\hat{B}, \hat{C}] + [\hat{A}, \hat{C}]\hat{B}$ and $[\hat{A}, \hat{B}] = -[\hat{B}, \hat{A}]$ and Equation 5.88, we find

$$[\hat{N}, \hat{a}] = [\hat{a}^+\hat{a}, \hat{a}] = [\hat{a}^+, \hat{a}]\hat{a} = -\hat{a} \quad [\hat{N}, \hat{a}^+] = [\hat{a}^+\hat{a}, \hat{a}^+] = \hat{a}^+[\hat{a}, \hat{a}^+] = \hat{a}^+ \quad (5.92)$$

We now show $\hat{N} = \hat{a}^+\hat{a}$ and $\hat{\mathcal{H}} = \hbar\omega_0(\hat{N} + 1/2)$ have eigenvectors $|n-1\rangle \sim \hat{a}|n\rangle$ and $|n+1\rangle \sim \hat{a}^+|n\rangle$. Suppose $|n\rangle$ represents one eigenvector then

$$\hat{N}[\hat{a}|n\rangle] = \hat{N}\hat{a}|n\rangle = \{[\hat{N}, \hat{a}] + \hat{a}\hat{N}\}|n\rangle = \{-\hat{a} + \hat{a}\hat{N}\}|n\rangle = \{-\hat{a} + \hat{a}n\}|n\rangle = (n-1)[\hat{a}|n\rangle]$$

Therefore $\hat{a}|n\rangle$ must be an eigenvector of $\hat{N}$ with eigenvalue $(n-1)$. Because the basis vectors can be labeled by the corresponding eigenvalue, we must have $\hat{a}|n\rangle \sim |n-1\rangle$ or by including a constant of proportionality $\hat{a}|n\rangle = D_n|n-1\rangle$. We can similarly show that $\hat{N}[\hat{a}^+|n\rangle] = (n+1)[\hat{a}^+|n\rangle]$ (see the chapter review exercises). Therefore, we conclude $\hat{a}^+|n\rangle = C_n|n+1\rangle$ and $\hat{a}|n\rangle = D_n|n-1\rangle$ since the eigenvalues are not degenerate where C_n and D_n denote constants of proportionality.

The eigenvalues of $\hat{N} = \hat{a}^+ \hat{a}$ and $\hat{\mathcal{H}} = \hbar\omega_o(\hat{N} + \frac{1}{2})$ must be real because $\hat{N} = \hat{a}^+ \hat{a}$ is Hermitian according to $\hat{N}^+ = (\hat{a}^+ \hat{a})^+ = \hat{a}^+ \hat{a} = \hat{N}$. Further the eigenvalues n must be greater than or equal to zero since the length of a vector must always be positive $n = \langle n | \hat{N} | n \rangle = \langle n | \hat{a}^+ \hat{a} | n \rangle = \|\hat{a} | n \rangle\|^2 \geq 0$. We can also show that only integers represent the eigenvalues n .

Next, we find the normalization constants C_n and D_n occurring in the relations.

$$\hat{a}^+ |n\rangle = C_n |n+1\rangle \quad \hat{a} |n\rangle = D_n |n-1\rangle$$

Let us work with the lowering operator. To find D_n , consider the string of equalities

$$\begin{aligned} D_n^* D_n \langle n-1 | n-1 \rangle &= [D_n |n-1\rangle]^+ [D_n |n-1\rangle] = [\hat{a} |n\rangle]^+ [\hat{a} |n\rangle] \\ &= \langle n | \hat{a}^+ \hat{a} | n \rangle = \langle n | \hat{N} | n \rangle = \langle n | n \rangle = n \langle n | n \rangle \end{aligned}$$

Now use the fact that all eigenvectors are normalized to one so that

$$\langle n-1 | n-1 \rangle = 1 = \langle n | n \rangle$$

Therefore, the coefficient D_n must be

$$|D_n|^2 = n \quad \rightarrow \quad D_n = \sqrt{n}$$

where a phase factor has been ignored. Similarly, an expression for C_n can be developed

$$\begin{aligned} C_n^* C_n \langle n+1 | n+1 \rangle &= [C_n |n+1\rangle]^+ [C_n |n+1\rangle] = [\hat{a}^+ |n\rangle]^+ [\hat{a}^+ |n\rangle] \\ &= \langle n | \hat{a} \hat{a}^+ | n \rangle = \langle n | \hat{a}^+ \hat{a} + 1 | n \rangle = \langle n | \hat{N} + 1 | n \rangle = \langle n | n+1 \rangle = (n+1) \langle n | n \rangle \end{aligned}$$

where a commutator has been used for the fifth term. Once again using the eigenvector normalization conditions and comparing both sides of the last equation

$$|C_n|^2 = n+1 \quad \rightarrow \quad C_n = \sqrt{n+1}$$

as expected. We therefore have the required relations.

$$\hat{a}^+ |n\rangle = \sqrt{n+1} |n+1\rangle \quad \hat{a} |n\rangle = \sqrt{n} |n-1\rangle \quad (5.93)$$

The set of eigenvectors

$$|0\rangle, |1\rangle, \dots$$

can be obtained by repeatedly using the relation $\hat{a}^+ |n\rangle = \sqrt{n+1} |n+1\rangle$ as

$$|1\rangle = \frac{\hat{a}^+}{\sqrt{1}} |0\rangle, \quad |2\rangle = \frac{\hat{a}^+}{\sqrt{2}} |1\rangle = \frac{(\hat{a}^+)^2}{\sqrt{2}\sqrt{1}} |0\rangle, \dots, \quad |n\rangle = \frac{(\hat{a}^+)^n}{\sqrt{n!}} |0\rangle, \dots \quad (5.94)$$

Some Commutation Relations

1. $[\hat{\mathcal{H}}, \hat{a}] = \hbar\omega_o[\hat{a}^+ \hat{a}, \hat{a}] = \hbar\omega_o \hat{a}^+ [\hat{a}, \hat{a}] + \hbar\omega_o [\hat{a}^+, \hat{a}] \hat{a} = -\hbar\omega_o \hat{a}.$
2. $[\hat{\mathcal{H}}, \hat{a}^+] = \hbar\omega_o \hat{a}^+ [\hat{a}, \hat{a}^+] = \hbar\omega_o \hat{a}^+.$
3. $[\hat{N}, \hat{a}] = -\hat{a} [\hat{N}, \hat{a}^+] = \hat{a}^+.$

5.4.6 ENERGY EIGENVALUES

The Hamiltonian for the harmonic oscillator can be written in terms of the ladder operators as given in Equation 5.91b.

$$\hat{\mathcal{H}} = \hbar\omega_o \left(\hat{N} + \frac{1}{2} \right) \quad (5.95)$$

We already know the eigenvalues of the number operator to be $\hat{N}|n\rangle = n|n\rangle$. The allowed energy values can be found as follows:

$$\hat{\mathcal{H}}|n\rangle = \hbar\omega_o \left(\hat{N} + \frac{1}{2} \right) |n\rangle = \hbar\omega_o \left(n + \frac{1}{2} \right) |n\rangle$$

Therefore the energy values must be

$$E_n = \hbar\omega_o \left(n + \frac{1}{2} \right) \quad (5.96)$$

5.4.7 ENERGY EIGENFUNCTIONS

The energy eigenvectors can be listed in the sequence

$$|1\rangle = \frac{\hat{a}^+}{\sqrt{1}}|0\rangle, \quad |2\rangle = \frac{\hat{a}^+}{\sqrt{2}}|1\rangle = \frac{(\hat{a}^+)^2}{\sqrt{2}\sqrt{1}}|0\rangle, \dots, \quad |n\rangle = \frac{(\hat{a}^+)^n}{\sqrt{n!}}|0\rangle, \dots \quad (5.97)$$

from Equation 5.94. However, we would like to know the functional form of these abstract vectors. There exists a simple method for finding the energy eigenfunctions for the harmonic oscillator using the ladder operators. Starting with

$$0 = \hat{a}|0\rangle$$

operate on both sides using the bra operator $\langle x|$ and insert the definition for the lowering operator

$$0 = \langle x|\hat{a}|0\rangle = \langle x| \frac{m\omega_o \hat{x}}{\sqrt{2\hbar m\omega_o}} + \frac{i\hat{p}}{\sqrt{2\hbar m\omega_o}} |0\rangle = \langle x| \frac{m\omega_o \hat{x}}{\sqrt{2\hbar m\omega_o}} |0\rangle + \langle x| \frac{i\hat{p}}{\sqrt{2\hbar m\omega_o}} |0\rangle \quad (5.98)$$

Factor out the constants from the brackets and use the relations (c.f., Appendix L)

$$\langle x|\hat{x}|0\rangle = x\langle x|0\rangle = x\phi_0(x) \quad \langle x|\hat{p}|0\rangle = \frac{\hbar}{i} \frac{\partial}{\partial x} \langle x|0\rangle = \frac{\hbar}{i} \frac{\partial}{\partial x} \phi_0(x)$$

where $\langle x|0\rangle = \phi_0(x)$ is the first energy eigenfunction in the set of eigenfunctions given by

$$\{ \phi_0(x), \phi_1(x), \dots \}$$

Equation 5.98 now provides

$$0 = \langle x|a|0\rangle = \frac{m\omega_o x}{\sqrt{2\hbar m\omega_o}} \phi_0(x) + \frac{\hbar}{\sqrt{2\hbar m\omega_o}} \frac{\partial}{\partial x} \phi_0(x)$$

which is a simple first-order differential equation

$$\frac{d\phi_0}{dx} + \frac{m\omega_0}{\hbar} x\phi_0 = 0$$

One can easily find the solution

$$\phi_0(x) = \phi_0(0) \exp\left(-\frac{m\omega_0}{2\hbar} x^2\right)$$

which represents the first energy eigenfunction. The normalization constant $\phi_0(0)$ is found by requiring the wave function to have unit length $1 = \langle \phi_0(x) | \phi_0(x) \rangle$ which gives

$$\phi_0(x) = \left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4} \exp\left(-\frac{m\omega_0}{2\hbar} x^2\right)$$

Now the other eigenfunctions can be found such as $\phi_1(x)$ by using the raising operator

$$\phi_1(x) = \langle x | 1 \rangle = \langle x | \frac{a^+}{\sqrt{1}} | 0 \rangle = \langle x | \frac{m\omega_0 \hat{x}}{\sqrt{2\hbar m\omega_0}} - \frac{i\hat{p}}{\sqrt{2\hbar m\omega_0}} | 0 \rangle$$

where the constants can be factored out and the coordinate representation can be substituted for the operators to get

$$\phi_1(x) = \frac{m\omega_0 x}{\sqrt{2\hbar m\omega_0}} \langle x | 0 \rangle - \frac{\hbar}{\sqrt{2\hbar m\omega_0}} \frac{\partial}{\partial x} \langle x | 0 \rangle = \frac{m\omega_0 x}{\sqrt{2\hbar m\omega_0}} \phi_0(x) - \frac{\hbar}{\sqrt{2\hbar m\omega_0}} \frac{\partial}{\partial x} \phi_0(x)$$

Notice that we do not need to solve a differential equation to find the eigenfunctions $\phi_1, \phi_2, \dots$ in the basis set. Differentiating $\phi_0(x)$ provides

$$\frac{\partial \phi_0}{\partial x} = \frac{\partial}{\partial x} \left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4} \exp\left(-\frac{m\omega_0}{2\hbar} x^2\right) = -\frac{m\omega_0 x}{\hbar} \phi_0(x)$$

Consequently the $n = 1$ energy eigenfunction becomes

$$\begin{aligned} \phi_1(x) &= \frac{m\omega_0 x}{\sqrt{2\hbar m\omega_0}} \phi_0(x) + \frac{\hbar}{\sqrt{2\hbar m\omega_0}} \frac{m\omega_0 x}{\hbar} \phi_0(x) = \frac{2\sqrt{m\omega_0}}{\sqrt{2\hbar}} x \phi_0(x) \\ &= \frac{2\sqrt{m\omega_0}}{\sqrt{2\hbar}} \left(\frac{m\omega_0}{\pi\hbar}\right)^{1/4} x \exp\left(-\frac{m\omega_0 x^2}{2\hbar}\right) \end{aligned}$$

The $n = 2$ energy eigenfunctions can be found by repeating the procedure using

$$\phi_2(x) = \frac{\hat{a}^+}{\sqrt{2}} \phi_1(x)$$

Notice that the above procedure requires only the relation between the ladder operators and the momentum/position operators. At this point, the energy eigenvalues can be found using the time-independent Schrödinger equation.

Special Integrals

The raising and lowering operators can be used to show the following integrals.

1. $\int_{-\infty}^{\infty} dx \phi_n(x) \frac{d}{dx} \phi_m(x) = \alpha \left[\delta_{m,n+1} \sqrt{\frac{n+1}{2}} - \delta_{m,n-1} \sqrt{\frac{n}{2}} \right]$ since $\frac{d}{dx} = \frac{i}{\hbar} \hat{p}$.
2. $\int_{-\infty}^{\infty} dx \phi_n(x) x \phi_m(x) = \sqrt{\frac{\hbar}{2m\omega_0}} [\delta_{m,n+1} \sqrt{n+1} + \delta_{m,n-1} \sqrt{n}]$ where Problem 5.37 was combined with integral (1) above and with $E_{n+1} - E_n = \hbar\omega_0$.
3. $\int_{-\infty}^{\infty} dx \phi_n(x) x^2 \phi_m(x) = \delta_{m,n} \frac{n+1}{2\alpha^2} + \delta_{m,n\pm 2} \frac{\sqrt{(n+1)(n+2)}}{2\alpha^2}$. The closure relation can be used to prove this last one.

5.5 INTRODUCTION TO ANGULAR MOMENTUM

Angular momentum plays a very important role in the description of nature. Orbital and spin angular momentum help explain the emission spectra from atoms and materials. Angular momentum represents an extension of the concept of linear momentum. The areas of spintronics, quantum computing, and quantum teleportation provide example technological applications for the angular momentum.

As discussed in the present section, a system for which the Hamiltonian is independent of angle (i.e., there are no rotational forces applied to the system) conserves angular momentum. The invariance of the system to rotations leads to a zero commutator between the Hamiltonian and the angular momentum. The fundamental states of the particle can then be labeled by the energy and angular momentum; these states form a common basis set to describe energy and angular momentum. The basis set consists of the spherical harmonic functions which depend only on the angular position coordinates (θ and φ).

The present section first reviews the classical expression for the angular momentum and Newton's relations between angular momentum and torque. The quantum theory of the angular momentum is obtained by replacing the dynamical variables with operators and using abstract vectors in Hilbert space to represent the specific properties of the angular momentum. The classical value of the angular momentum can be recovered from the quantum one by averaging the operators for a specific wave function. Once having introduced the operators, the eigenvalues and eigenfunctions can be introduced. These have special significance for systems with rotational invariance since the most fundamental states of the particle can most conveniently be labeled by angular momentum eigenvalues. The rotational invariance ensures the conservation of angular momentum.

5.5.1 CLASSICAL DEFINITION OF ANGULAR MOMENTUM

A point mass rotating about an axis has angular momentum

$$\vec{L} = \vec{r} \times \vec{p} \quad (5.99)$$

with a direction given by the right-hand rule. The simplest example appears in Figure 5.22 for a point particle confined to 2-D motion. Classically, the magnitude of the angular momentum for circular motion must be $L = rp = I\omega$ where I represents the moment of inertia $I = mr^2$.

The angular momentum vector for a point particle can be divided into components according to

$$\vec{L} = \vec{r} \times \vec{p} = \tilde{x}L_x + \tilde{y}L_y + \tilde{z}L_z \quad (5.100)$$

where the components have the form

$$L_x = yp_z - zp_y \quad L_y = zp_x - xp_z \quad L_z = xp_y - yp_x \quad (5.101a)$$

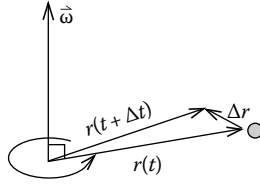


FIGURE 5.22 Rotating point particle a fixed distance r from the origin.

and where $\tilde{x}, \tilde{y}, \tilde{z}$ represent the usual Euclidean basis vectors. The angular momentum can be written in terms of a determinant as

$$\vec{L} = \vec{r} \times \vec{p} = \begin{vmatrix} \tilde{x} & \tilde{y} & \tilde{z} \\ x & y & z \\ p_x & p_y & p_z \end{vmatrix} \quad (5.101b)$$

The antisymmetric tensor ε_{ijk} can be used to provide a more convenient and compact notation

$$L_i = \sum_{jk} \varepsilon_{ijk} x_j p_k \quad (5.101c)$$

The antisymmetric tensor produces zero if two indices have the same value $\varepsilon_{iji} = 0$ (etc.), and interchanging indices produces a minus sign $\varepsilon_{ijk} = -\varepsilon_{jik}$. Consequently, all of the elements can be generated given the single value $\varepsilon_{123} = 1$. Often, the Einstein repeated-index sum convention is used to write $L_i = \varepsilon_{ijk} x_j p_k$. The indices j, k are repeated and therefore must be summed.

Recall from classical mechanics how the torque $\vec{\tau} = \vec{r} \times \vec{F}$ produces a change in the angular momentum.

$$\vec{\tau} = \vec{r} \times \vec{F} = \vec{r} \times \dot{\vec{p}} = \frac{d}{dt} \vec{r} \times \vec{p} = \dot{\vec{L}} \quad (5.102)$$

where we have used $\dot{\vec{r}} \times \vec{p} = \vec{v} \times m\vec{v} = 0$. Equation 5.102 shows that the angular momentum must be conserved ($\vec{L} = \text{constant}$) in the absence of an external torque $\vec{\tau} = 0$.

5.5.2 ORIGIN OF ANGULAR MOMENTUM IN QUANTUM MECHANICS

The quantum theory of angular momentum replaces the classical dynamical variables (such as x and P_x) with operators and uses vectors in an abstract Hilbert space to represent the specific angular momentum properties of a particle. This section first shows a semiclassical example for why the angular momentum should be quantized.

The need for quantizing the angular momentum can be appreciated from the example of an electron orbiting about a nucleus in the Bohr atom (Figure 5.23). Considering the wave nature of the particle, we require an exact number of wavelengths to fit the circumference $C = 2\pi r$ in order to form a standing wave. If the electron were to orbit the nucleus without a definite phase relation with respect to the path, then one can speculate that without a standing wave, the wave function would time-average to zero and the particle would be nonexistent. For the standing wave, the wavelength can be no longer than $\lambda = C$. In general, m complete wavelengths must fit in the circumference so that $\lambda = C/m$. Assuming that the particle orbits in the x - y plane, the angular momentum must be along the z -direction with a magnitude given by

$$L_z = rp = r\hbar k = \frac{2\pi r\hbar}{\lambda} = \frac{2\pi r m\hbar}{C} = \frac{2\pi r m\hbar}{2\pi r} = m\hbar \quad (5.103)$$

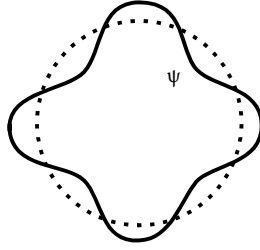


FIGURE 5.23 An orbit in the Bohr atom with an integral number of de Broglie wavelengths.

where m denotes an integer. In this quasiclassical case, the total angular momentum points along the z -direction. The simple model shows that the electron wave function must form a standing wave in order that phase variations not produce a zero average value.

Equation 5.103 indicates that the total angular momentum must be along the z -direction perpendicular to the plane of rotation. However, a measurement of the angular momentum for a microscopic system shows that the observed z -component will be smaller than the observed total magnitude. The observed angular momentum does not appear to align itself with any specific axis. The following discussion shows that the operators representing the three components of angular momentum do not commute with each other. Therefore, we cannot surmise the angular momentum aligns completely with the z -axis since then we must be certain the other two components are precisely zero. That is, we then simultaneously and precisely know the three components of the angular momentum which would require the three corresponding operators to all commute with each other in contradiction to the provable noncommutivity of those operators.

5.5.3 ANGULAR MOMENTUM OPERATORS

The quantum mechanical description of angular momentum replaces the classical dynamical variables with operators. We have primary interest in finding the values of the angular momentum and the fundamental states (i.e., basis states) in which the particles will be found upon observation. We will find that the angular momentum can be no better described than by specifying the length (i.e., magnitude) and the z -component of the angular momentum which leads to the basis states $|\ell, m\rangle$ with the observed values $L = \sqrt{\ell(\ell+1)}\hbar$ and $L_z = m\hbar$ where $\ell = 0, 1, 2, \dots$ and $m = -\ell, -\ell+1, \dots, 0, \dots, \ell-1, \ell$. Measurements sensitive to the total angular momentum L and the z -component L_z will produce exactly one of the values for L and L_z . Mathematically, one defines raising and lowering operators to step from one z -state to another. These ladder operators will help to prove the relations for L and L_z . First, however, we focus on defining the angular momentum operators and their commutation relations, and then show a physical picture.

Equation 5.101a provides the classical expressions for the components of the angular momentum, repeated as

$$L_x = yp_z - zp_y \quad L_y = zp_x - xp_z \quad L_z = xp_y - yp_x$$

The total angular momentum vector can be written in the quantum theory as $\vec{L} = \hat{L}_x\tilde{x} + \hat{L}_y\tilde{y} + \hat{L}_z\tilde{z}$ where $\tilde{x}, \tilde{y}, \tilde{z}$ represent the usual Euclidean basis vectors. The components in the quantum theory take the form of operators

$$\hat{L}_1 = \hat{x}_2\hat{p}_3 - \hat{x}_3\hat{p}_2 \quad \hat{L}_2 = \hat{x}_3\hat{p}_1 - \hat{x}_1\hat{p}_3 \quad \hat{L}_3 = \hat{x}_1\hat{p}_2 - \hat{x}_2\hat{p}_1 \quad (5.104a)$$

where $i = 1, 2, 3$ corresponds to x, y, z , respectively. The (square of the) total angular momentum must be

$$\hat{L}^2 = \hat{\vec{L}} \cdot \hat{\vec{L}} = \hat{L}_1^2 + \hat{L}_2^2 + \hat{L}_3^2 \quad (5.104b)$$

The commutation relations tell us which quantities can be simultaneously observed and therefore, which have common eigenvectors. We will show the relations

$$[\hat{L}_i, \hat{L}_j] = i\hbar \varepsilon_{ijk} \hat{L}_k \quad (\text{sum convention}) \quad [\hat{L}_3, \hat{L}^2] = 0 \quad (5.105)$$

where ε_{ijk} is the totally antisymmetric tensor as discussed in Section 5.5.1. Notice that the “Einstein summation convention” has been used which means repeated indices must be summed. Based on the second of Equation 5.105, one can surmise that the z -component and the length of the angular momentum form the most complete set of operators from which to construct a basis set.

The first of Equation 5.105 can be shown using the elemental commutation relations between position and momentum.

$$[\hat{x}_i, \hat{x}_j] = 0 \quad [\hat{p}_i, \hat{p}_j] = 0 \quad [\hat{x}_i, \hat{p}_j] = i\hbar \delta_{ij} \quad (5.106)$$

Calculate

$$[\hat{L}_1, \hat{L}_2] = [\hat{x}_2 \hat{p}_3 - \hat{x}_3 \hat{p}_2, \hat{x}_3 \hat{p}_1 - \hat{x}_1 \hat{p}_3] = [\hat{x}_2 \hat{p}_3, \hat{x}_3 \hat{p}_1] + [-\hat{x}_3 \hat{p}_2, -\hat{x}_1 \hat{p}_3] = \hat{x}_2 [\hat{p}_3, \hat{x}_3] \hat{p}_1 + \hat{x}_1 [\hat{x}_3, \hat{p}_3] \hat{p}_2$$

which can be simplified to

$$[\hat{L}_1, \hat{L}_2] = -i\hbar \hat{x}_2 \hat{p}_1 + i\hbar \hat{x}_1 \hat{p}_2 = i\hbar \hat{L}_3$$

The other commutators embodied in the first of Equation 5.105 similarly hold. The components do not commute. And yet, we must have as large as possible the set of commuting operators so that we might specify the basic states as much as possible. We chose one such operator as $\hat{L}_3 = \hat{L}_z$ and will show that $\hat{L}^2$ provides the only other one to complete the set of observables for angular momentum (excluding spin). The eigenvalues of these two operators provide the possible observed values.

The second of Equation 5.105 can be demonstrated.

$$[\hat{L}_3, \hat{L}^2] = [\hat{L}_3, \hat{L}_1^2 + \hat{L}_2^2 + \hat{L}_3^2] = [\hat{L}_3, \hat{L}_1^2] + [\hat{L}_3, \hat{L}_2^2] = 0 \quad (5.107)$$

Therefore the most complete description of the angular momentum uses the simultaneous eigenfunctions of the z -component of the angular momentum and the magnitude of the angular momentum (actually, the magnitude squared since it's easier to use in calculations).

In a similar manner, one can show the relations (Einstein sum convention)

$$[\hat{L}_i, \hat{r}_j] = i\hbar \varepsilon_{ijk} \hat{r}_k \quad [\hat{L}_i, \hat{p}_j] = i\hbar \varepsilon_{ijk} \hat{p}_k \quad (5.108)$$

where $r_1 = x$, $r_2 = y$, $r_3 = z$. For example, one can calculate the following

$$[L_x, y] = [y \hat{p}_z - z \hat{p}_y, y] = [y, z \hat{p}_y] = z i\hbar$$

5.5.4 PICTURES FOR ANGULAR MOMENTUM IN QUANTUM MECHANICS

As previously mentioned but yet demonstrated, the magnitude of the angular momentum and z -component operators have the following eigenvalues.

$$L = \sqrt{\ell(\ell+1)} \hbar \quad \text{where } \ell = 0, 1, 2, \dots \quad (5.109a)$$

$$L_z = m \hbar \quad \text{where } m = -\ell, -\ell+1, \dots, 0, \dots, \ell-1, \ell \quad (5.109b)$$

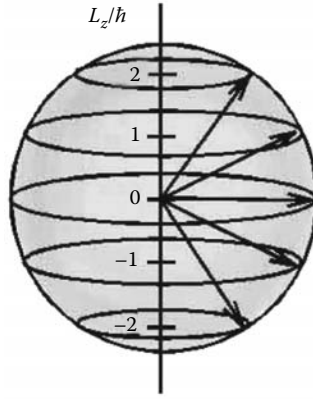


FIGURE 5.24 $L = \sqrt{6}\hbar$ for $\ell = 2$ and the various possible states for the L_z . The total angular momentum corresponds to the arrows and L represents their lengths.

where L denotes the total angular momentum. Upon measurement on a system, these will be the possible observed values. One can see that the total angular momentum L is larger than the z -component (for nonzero angular momentum) as a result of commutation relations. The x -, y -, z -components do not commute with each other. Therefore it is not possible to specify with certainty that $L_x = L_y = 0$ and $L_z = m\hbar$ ($m > 0$) in order to make $L = L_z$.

Figure 5.24 shows that the total angular momentum (vector) corresponding to $\ell = 2$ produces only five possible values for the z -component. The length of the total angular momentum vector and the z -component are well defined as indicated by the sphere of definite radius and the circles at a definite z -coordinate. The circles around the z -axis indicate that the x - and y -components are not precisely known. While a measurement of the x - and y -components can only produce five different values, the average could take on any value along the circles (assuming that the z -component remains fixed). The figure shows that the values of the z -component agree with the simple formula in Equation 5.103 and in addition shows the relation to the total magnitude. The quantized total angular momentum can be represented as a collection of concentric spheres of definite radii given by $\hbar\sqrt{\ell(\ell+1)}$. Based on Equation 5.109, spheres with larger radii (i.e., larger total angular momentum) have a greater number of circles around the z -axis.

The complete set of vectors can be written as

$$\{|\ell, m\rangle \quad \ell = 0, 1, \dots, \infty; \quad m = 0, \pm 1, \pm 2, \dots, \ell\}$$

where the “magnitude” of the momentum has eigenvalues

$$\hat{L}|\ell, m\rangle = \hbar\sqrt{\ell(\ell+1)}|\ell, m\rangle \quad (5.110a)$$

and the eigenvalues of $\hat{L}_z$ are

$$\hat{L}_z|\ell, m\rangle = m\hbar|\ell, m\rangle \quad (5.110b)$$

The basis vectors $|\ell, m\rangle$ span a vector space. For each ℓ , there exists a subspace spanned by the basis vectors $|m\rangle$. The full space spanned by $S_1 = Sp\{|\ell, m\rangle\}$ can be viewed as a direct product space $S_1 = Sp\{|\ell\rangle|m\rangle\}$ where the size of the second space $\{|m\rangle\}$ depends on the value of ℓ . In terms of energy degeneracy, the subspace represents degenerate states in the absence of some mechanism to lift the degeneracy. So for $\ell = 2$ there would be five degenerate states unless, for example, a magnetic field is applied. As a final note, separating variables in the Schrödinger equation for a system exhibiting rotational symmetry (such as for the atom) places limitations on

the range of l in terms of the principle quantum number n that describes the radial component of the wave function. In such a case, the range of n is infinite, but for each n , the number l must be limited to integers in the range $0 \leq l < n$. The limitation imposed by the principal quantum number n should remind the reader of the periodic table and the spectroscopic notation of 1S, 2S, 2P, and so on.

5.5.5 ROTATIONAL SYMMETRY AND CONSERVATION OF ANGULAR MOMENTUM

A system has rotational symmetry if “rotating” the Hamiltonian through an arbitrary angle θ does not change it. Hamilton’s equations from the previous chapter

$$\dot{q} = \frac{\partial H}{\partial p} \quad \text{and} \quad \dot{p} = -\frac{\partial H}{\partial q} \quad (5.111)$$

(where the generalized coordinate $q = \theta$, and the conjugate momentum $p = L_z$ for example) indicate that invariance with respect to angle then requires the angular momentum to be conserved. For a system with conserved energy and angular momentum, one would expect to use these values to describe the basic states of the system. Also as might be expected based on the correspondence between the classical Poisson brackets

$$\frac{dA}{dt} = [A, H] + \frac{\partial A}{\partial t} \quad \text{or more precisely} \quad \frac{dA}{dt} = \sum_i \left[\frac{\partial A}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial A}{\partial p_i} \frac{\partial H}{\partial q_i} \right] + \frac{\partial A}{\partial t} \quad (5.112)$$

and the quantum mechanical commutator, the time dependence of an operator in the quantum mechanics can be obtained from its commutator with the Hamiltonian. Therefore, an explicitly time-independent dynamical variable that commutes with the quantum mechanical Hamiltonian will be conserved. One expects to list the basis states using the largest set of commuting operators that are also conserved.

The chapter on the linear algebra of operators shows the rotation operator

$$\hat{R} = e^{i\vec{\alpha} \cdot \vec{L}/\hbar} \quad (5.113)$$

maps the wave function of a system into another one corresponding to rotated position coordinates. Here, the angular momentum is $\vec{L} = \hat{L}_x \tilde{x} + \hat{L}_y \tilde{y} + \hat{L}_z \tilde{z}$, and $\vec{\alpha} = \alpha_x \tilde{x} + \alpha_y \tilde{y} + \alpha_z \tilde{z}$ denotes a rotation angle. In the 3-D case, $|\vec{\alpha}|$ represents the rotation angle about the unit axis $\vec{\alpha}/|\vec{\alpha}|$. Often the direction of $\vec{\alpha}$ aligns with the z -axis and we can then work with the rotation operator in the form $\hat{R} = e^{i\theta_0 \hat{L}_z/\hbar}$ where the z -component of angular momentum can be written in any number of forms $\hat{L}_z = \hat{x}\hat{p}_y - \hat{y}\hat{p}_x = \frac{\hbar}{i}(x\partial_y - y\partial_x) = \frac{\hbar}{i}\frac{\partial}{\partial \theta} = \frac{\hbar}{i}\partial_\theta$. Recall the “rotation operator” rotates functions as shown in Figure 5.25 (refer to Chapter 3).

$$\psi'(\theta) = \psi(\theta - \theta_0) = \hat{R}(\theta_0) \psi(\theta) \quad (5.114)$$

Figure 5.25 shows how the rotation moves the function in the direction of a positive angle or rotates the coordinates in the negative direction.

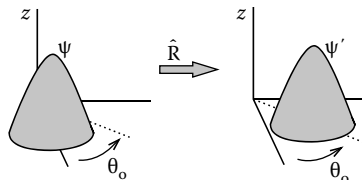


FIGURE 5.25 Rotating the function through an angle.

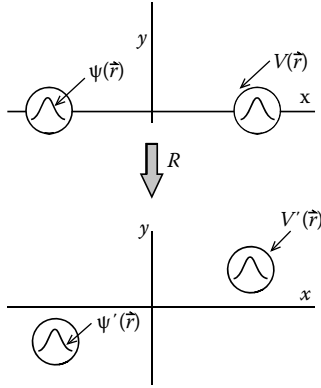


FIGURE 5.26 Rotating the system.

Suppose a system is rotated through an arbitrary angle θ_0 as shown in Figure 5.26. Rotating the example system shows that the wave function must have a new functional form and the Hamiltonian could in principle produce a different value for the energy. The upper part of the figure attempts to indicate that the wave function and potential mostly depend on the x -coordinate, whereas the lower portion attempts to indicate that the wave function and potential for the rotated system depend on both x - and y -coordinates.

Suppose to find conditions for the invariance of Hamiltonian under rotation, one requires the average measured energy to be invariant. That is, suppose one requires the expected energy to be the same for either configuration.

$$\langle \psi | \hat{H} | \psi \rangle = \langle \psi' | \hat{H}' | \psi' \rangle \quad (5.115)$$

Using the definition of the rotated wave function $|\psi'\rangle = R|\psi\rangle$, the rotated Hamiltonian must have the form

$$\hat{H}' = \hat{R} \hat{H} \hat{R}^\dagger \quad (5.116)$$

Notice one could simply write Equation 5.116 based on the notion of rotating an operator as presented in Chapter 3, which then directly leads to the invariance of expected energy in Equation 5.115.

If the system exhibits rotational symmetry then $\hat{H}' = \hat{H}$ for all angles. Equation 5.116 then indicates the Hamiltonian and the rotation operator must commute.

$$\hat{H} = \hat{R} \hat{H} \hat{R}^\dagger \rightarrow [\hat{H}, \hat{R}] = 0 \quad (5.117)$$

Equation 5.116 can also be written using Equation 5.113 (in particular, $\hat{R} = e^{i\theta_0 \hat{L}_z / \hbar}$) for a small angle ϵ and then expanded as

$$\hat{H}' = \hat{R} \hat{H} \hat{R}^\dagger = e^{i\epsilon \hat{L}_z / \hbar} \hat{H} e^{-i\epsilon \hat{L}_z / \hbar} \cong \hat{H} - \frac{i\epsilon}{\hbar} [\hat{H}, \hat{L}_z] + O(\epsilon^2)$$

Moving $\hat{H}$ to the left, dividing through by ϵ , and taking the limit $\epsilon \rightarrow 0$ produces the results

$$\frac{\partial \hat{H}}{\partial \theta} = \frac{1}{i\hbar} [\hat{H}, \hat{L}_z] \quad (5.118)$$

The rotational invariance requires the Hamiltonian to be independent of angle so that $\partial_\theta \hat{H} = 0$. In such a case, the angular momentum $\hat{L}_z$ commutes with the Hamiltonian and the two operators therefore share a common basis set describing the fundamental states of motion.

The rotational invariance also requires the angular momentum to be conserved as can be seen by two classical methods and a third quantum mechanical one. The simplest method uses Hamilton's classical canonical equation $\dot{L}_z = -\partial_\theta H$ from Chapter 4 where θ and L_z represent the angular canonical variables instead of x and p_x . The rotational invariance of the Hamiltonian then requires $\dot{L}_z = -\partial_\theta H = 0$ so that L_z must be a constant of the motion (independent of time). The second classical method uses Equation 5.102 in addition to the definition of force derived from a potential function V given by $-\vec{F} = \nabla V = \hat{r} \partial_r V + \frac{\partial}{\partial r} \partial_\theta V$ where the angle refers to a plane perpendicular to the z -direction. Then the torque will be $\vec{\tau} = \vec{r} \times \vec{F} = -\hat{z} \partial_\theta V$. However, the potential energy must be independent of angle and therefore the torque must be zero.

The third method uses quantum mechanical formalism. Consider the time evolution operator $\hat{u}(t) = \exp\left(\frac{i\hat{H}t}{\hbar}\right)$ for a closed system as described in Section 5.1 (c.f., Equation 5.12) and also Section 5.8 regarding quantum mechanical representations. Then making the expansion for the time-evolved operator provides

$$\hat{L}'_z = \hat{u} \hat{L}_z \hat{u}^\dagger = e^{-i\hat{H}\varepsilon/\hbar} \hat{L}_z e^{i\hat{H}\varepsilon/\hbar} \cong \hat{L}_z - \frac{i\varepsilon}{\hbar} [\hat{H}, \hat{L}_z] + O(\varepsilon^2)$$

where ε represents an infinitesimal change in time. As before, in the limit, one finds

$$\frac{\partial \hat{L}_z}{\partial t} = \frac{1}{i\hbar} [\hat{H}, \hat{L}_z] \quad (5.119a)$$

If z -component of the angular momentum commutes with the Hamiltonian then the z angular momentum is conserved.

$$[\hat{H}, \hat{L}_z] = 0 \quad \rightarrow \quad \hat{L}_z = \text{Constant} \quad (5.119b)$$

Interestingly, comparing Equations 5.118 and 5.119a, one can see how the rotational invariance links with the conservation of angular momentum

$$\frac{\partial \hat{H}}{\partial \theta} = \frac{\partial \hat{L}_z}{\partial t}$$

A symmetry leads to a conservation law. Regardless of the method and the size-scale of the system, one can see that central potentials produce conservation of angular momentum.

A system invariant to rotations about all three axes must conserve the total angular momentum. The Hamiltonian in Equation 5.118 must then commute with all components of the angular momentum $[\hat{H}, \vec{\hat{L}}] = 0$ where the total angular momentum has the form $\vec{\hat{L}} = \hat{x}\hat{L}_x + \hat{y}\hat{L}_y + \hat{z}\hat{L}_z$. However, the fundamental states (basis states) cannot be described by all components of angular momentum since they do not all commute with each other as shown in Equation 5.105. The best a person can do is to describe the basis states in terms of energy and of the angular momentum magnitude and one of the components of the angular momentum (Equation 5.110).

5.5.6 EIGENVALUES AND EIGENVECTORS

We previously mentioned that the complete set of vectors can be written as

$$\{|l, m\rangle \quad l = 0, 1, \dots, \infty; \quad m = 0, \pm 1, \pm 2, \dots, l\} \quad (5.120a)$$

where the “magnitude” of the momentum $\hat{L} = \sqrt{\hat{L}^2} = \sqrt{\hat{L} \cdot \hat{L}}$ has eigenvalues

$$\hat{L}|\ell, m\rangle = \hbar\sqrt{\ell(\ell+1)}|\ell, m\rangle \quad (5.120b)$$

and the eigenvalues of $\hat{L}_z$ are

$$\hat{L}_z|\ell, m\rangle = m\hbar|\ell, m\rangle \quad (5.120c)$$

Now we demonstrate these eigenvalues. The next section shows the function form of the eigenvectors in terms of the spherical harmonics.

We will need the following relations (refer to the chapter problems). Define the raising $\hat{L}_+$ and lowering $\hat{L}_-$ operator for the L_z component $\hat{L}_\pm = \hat{L}_x \pm i\hat{L}_y$ where $\hat{L}_-^+ = \hat{L}_+$.

$$[\hat{L}_i, \hat{L}_j] = i\hbar\epsilon_{ijk}\hat{L}_k \quad (\text{sum convention}) \quad (5.121a)$$

$$\hat{L}_+\hat{L}_- = \hat{L}^2 - \hat{L}_z^2 + \hbar\hat{L}_z, \quad \hat{L}_-\hat{L}_+ = \hat{L}^2 - \hat{L}_z^2 - \hbar\hat{L}_z \quad (5.121b)$$

$$[\hat{L}_+, \hat{L}_-] = 2\hbar\hat{L}_z \quad (5.121c)$$

$$[\hat{L}_z, \hat{L}_-] = -\hbar\hat{L}_-, \quad [\hat{L}_z, \hat{L}_+] = \hbar\hat{L}_+ \quad (5.121d)$$

The complete set of commuting observables for angular momentum is $\{\hat{L}^2, \hat{L}_z\}$ since $[\hat{L}^2, \hat{L}_z] = 0$. The component $\hat{L}_z$ can be replaced with one of the other two components (i.e., $\hat{L}_x$ or $\hat{L}_y$) if desired. The set of basis vectors must have the form $|\ell, m\rangle$ where ℓ, m refer to $\hat{L}^2$ and $\hat{L}_z$, respectively.

We first show that the eigenvalues for the total angular momentum $\hat{L}^2$ must be larger than zero in a manner similar to that for the harmonic oscillator. Let $|\lambda\rangle$ be an eigenfunction of the Hermitian operator $\hat{L} = \sqrt{\hat{L} \cdot \hat{L}}$. We have

$$\lambda = \langle\lambda|\hat{L}^2|\lambda\rangle = \{\langle\lambda|\hat{L}\rangle\{\hat{L}|\lambda\rangle\} = \|\hat{L}|\lambda\rangle\|^2 \geq 0 \quad (5.122)$$

since the magnitude of a vector must always be nonnegative by the definition of inner product. As will be shown next, the eigenvalues have the values given in Equation 5.120b and c. Equation 5.122 then indicates $\ell(\ell+1) \geq 0$ so that $\ell \geq 0$. The values ℓ (and m) are integers for orbital angular momentum.

To find the eigenvalues and eigenvectors of the angular momentum $\hat{L}^2$ and $\hat{L}_z$, define the raising and lowering operators for the z -component as given above $\hat{L}_\pm = \hat{L}_x \pm i\hat{L}_y$. We need to show four relations.

$$\hat{L}_-|\ell, m\rangle = c_-|\ell, m-1\rangle \quad c_- = \hbar\sqrt{\ell(\ell+1) - m(m-1)} \quad (5.123a)$$

$$\hat{L}_+|\ell, m\rangle = c_+|\ell, m+1\rangle \quad c_+ = \hbar\sqrt{\ell(\ell+1) - m(m+1)} \quad (5.123b)$$

For a given value of ℓ , the raising and lowering operators map one basis vector into another. We will demonstrate Equation 5.123a and leave the other one for the exercises at the end of the chapter. Consider

$$\hat{L}_z[\hat{L}_-|\ell, m\rangle] = (\hat{L}_-\hat{L}_z - \hbar\hat{L}_-)|\ell, m\rangle = \hbar(m-1)[\hat{L}_-|\ell, m\rangle]$$

where the first step follows from the commutator relation in Equation 5.121d. This last equation shows that vector $\hat{L}_-|\ell, m\rangle$ must be the eigenvector corresponding to the eigenvalue $\hbar(m-1)$ and

therefore $\hat{L}_-|\ell, m\rangle = c_-|\ell, m-1\rangle$. Now to find the value of c_- consider the string of equalities that make use of the relations in Equation 5.121

$$|c_-|^2 = \{\langle \ell, m | \hat{L}_+ \rangle \{\hat{L}_- | \ell, m \rangle\} = \langle \ell, m | (\hat{L}^2 - \hat{L}_z^2 + \hbar \hat{L}_z) | \ell, m \rangle = \hbar^2 \ell(\ell+1) - \hbar^2 m^2 + \hbar^2 m$$

Ignoring an arbitrary phase factor we have the results required for Equation 5.123a

$$c_- = \hbar \sqrt{\ell(\ell+1) - m(m-1)}$$

Next, in order to show the values $m=0, \pm 1, \dots, \pm \ell$, we need two more results

$$0 \leq \langle \ell, m | \hat{L}_+ \hat{L}_- | \ell, m \rangle = \hbar^2 [\ell(\ell+1) - m^2 + m] \rightarrow \left(\ell + \frac{1}{2}\right)^2 \geq \left(m - \frac{1}{2}\right)^2 \quad (5.124a)$$

$$0 \leq \langle \ell, m | \hat{L}_- \hat{L}_+ | \ell, m \rangle = \hbar^2 [\ell(\ell+1) - m^2 - m] \rightarrow \left(\ell + \frac{1}{2}\right)^2 \geq \left(m + \frac{1}{2}\right)^2 \quad (5.124b)$$

We show the first one and leave the second one for the end-of-chapter exercises. The inequality follows from the definition of inner product similar to Equation 5.122. Then since $(\hat{L}_+)^+ = \hat{L}_-$, one finds

$$0 < \langle \ell, m | \hat{L}_+ \hat{L}_- | \ell, m \rangle = \langle \ell, m | \hat{L}^2 - \hat{L}_z^2 + \hbar \hat{L}_z | \ell, m \rangle = \hbar^2 \ell(\ell+1) - \hbar^2 m^2 + \hbar^2 m$$

This last equation then gives the inequality $\ell^2 + \ell \geq m^2 - m$ and adding $1/4$ to each side gives the desired result of $\left(\ell + \frac{1}{2}\right)^2 \geq \left(m - \frac{1}{2}\right)^2$.

For each value of ℓ there must be a range of m -values because of the ladder operators. We already know $\ell \geq 0$. Equation 5.124 must simultaneously hold so that $\underbrace{-\ell-1}_b < \underbrace{-\ell}_a \leq m \leq \underbrace{+\ell}_b \leq \underbrace{\ell+1}_a$

where the underline indicates whether the term came from Equation 124a or b. However, both equations must hold simultaneously so one must choose the inner limits on either side. As a result, we conclude $-\ell \leq m \leq +\ell$. The equality holds only if ℓ is an integer since the ladder operators change m by an integer.

Therefore, we conclude the eigenvectors must be

$$|\ell, m=0\rangle, |\ell, m=\pm 1\rangle, \dots, |\ell, m=\pm \ell\rangle$$

as required.

5.5.7 EIGENVECTORS AS SPHERICAL HARMONICS

Spherical harmonics represent the typical orthonormal set of functions arising from Laplace's and Poisson's equations. A geometry with spherical symmetry suggests that a boundary value problem should be stated in spherical coordinates. The quantum theory uses the spherical harmonics as part of the wave functions for an electron bound by the potential of an atomic nucleus as well as for the eigenfunctions of the angular momentum operator.

The angular momentum operator can be written in spherical coordinates using the basic definitions given in Figure 5.27. The components in Equation 5.105 become

$$\hat{L}_x = i\hbar(\sin\phi \partial_\theta + \cot\theta \cos\phi \partial_\phi) \quad \hat{L}_y = i\hbar(-\cos\phi \partial_\theta + \cot\theta \sin\phi \partial_\phi) \quad \hat{L}_z = \frac{\hbar}{i} \partial_\phi \quad (5.125a)$$

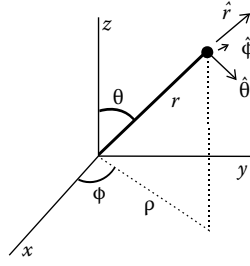


FIGURE 5.27 The spherical coordinates.

and the total magnitude (squared) becomes $\hat{L}^2 = \hat{L} \cdot \hat{L}$

$$\hat{L}^2 = -\frac{\hbar^2}{\sin \theta} \frac{\partial}{\partial \theta} \sin \theta \frac{\partial}{\partial \theta} - \frac{\hbar^2}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \quad (5.125b)$$

We look for the eigenfunctions $Y_l^m(\theta, \phi)$, the spherical harmonics, and eigenvalues of $\hat{L}_z, \hat{L}^2$. For solutions of partial differential equations, the spherical harmonics arise as solutions of Laplace's equation $\nabla^2 u(r, \theta, \phi) = 0$ where we require $u(r, \theta, \phi + 2\pi) = u(r, \theta, \phi)$. In this case, the function u is separated into radial and angular functions $u(r, \theta, \phi) = R(r)Y(\theta, \phi)$ where $Y(\theta, \phi)$ will become the spherical harmonics.

We look for functions such that $\hat{L}^2 Y(\theta, \phi) = \lambda Y(\theta, \phi)$ where λ represents an eigenvalue or equivalently, a separation constant for the partial differential equation discussed above. The eigenvalue equation can be written as

$$\left[-\frac{\hbar^2}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) - \frac{\hbar^2}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right] Y(\theta, \phi) = \lambda Y(\theta, \phi) \quad (5.126)$$

We need to find λ and Y . There are two eigenvalue equations as can be seen by further dividing the function Y into $Y(\theta, \phi) = P(\theta)Q(\phi)$ and substituting it into Equation 5.126.

$$\frac{\hbar^2 Q}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial P}{\partial \theta} \right) + \frac{\hbar^2 P(\theta)}{\sin^2 \theta} \frac{\partial^2 Q(\phi)}{\partial \phi^2} = -\lambda P(\theta)Q(\phi)$$

Multiplying both sides by $\sin^2 \theta / (\hbar^2 P(\theta)Q(\phi))$, collecting terms and defining a second separation constant ν provides

$$\frac{\sin \theta}{P(\theta)} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial P}{\partial \theta} \right) + \frac{\lambda}{\hbar^2} \sin^2 \theta = \nu = -\frac{1}{Q(\phi)} \frac{\partial^2 Q(\phi)}{\partial \phi^2} \quad (5.127)$$

where ν is independent of the spherical coordinates r, θ, ϕ .

1. The first Sturm-Liouville problem for Q .

$$\frac{\partial^2 Q}{\partial \phi^2} + \nu Q = 0 \quad \text{and} \quad Q(\phi + 2\pi) = Q(\phi)$$

As usual, we must consider three ranges for the separation constant $\nu < 0$, $\nu \geq 0$. The range $\nu < 0$ produces only trivial solutions for $Q(\phi)$. For $\nu \geq 0$, we have

$$Q(\phi) = c_1 e^{i\sqrt{\nu}\phi} + c_2 e^{-i\sqrt{\nu}\phi} \quad (5.128)$$

Working with $e^{i\sqrt{v}\varphi}$ and the boundary condition $Q(\varphi + 2\pi) = Q(\varphi)$, we find $v = m^2$ for $m = 0, 1, 2, \dots$ (the range of m will be modified by the P solutions). The eigenfunctions in Equation 5.128 become $Q_m(\varphi) = c_1 e^{im\varphi} + c_2 e^{-im\varphi}$. Now, let's allow m to be negative so that Q_m can be written as

$$Q_m(\varphi) = c_1 e^{im\varphi} \quad m = 0, \pm 1, \pm 2, \dots \quad (5.129)$$

where the angle φ has the range $\varphi \in [0, 2\pi)$. But notice that v is still written as $v = m^2$. We need to normalize Q_m such that

$$1 = \langle Q_m(\varphi) | Q_m(\varphi) \rangle = \int_0^{2\pi} d\varphi Q_m^*(\varphi) Q_m(\varphi) = \int_0^{2\pi} d\varphi c_1^2 e^{-im\varphi} e^{im\varphi} \rightarrow c_1 = \frac{1}{\sqrt{2\pi}}$$

The orthonormal functions and eigenvalues are

$$Q_m(\varphi) = \frac{e^{im\varphi}}{\sqrt{2\pi}} \quad v = m^2 \quad m = 0, \pm 1, \pm 2, \dots \quad \varphi \in [0, 2\pi) \quad (5.130)$$

The allowed values of m are modified in the next part for the function P .

We now have the functional form of the eigenvectors of the z -component of the angular momentum $\hat{L}_z$ which satisfy

$$\hat{L}_z Q_m(\phi) = \frac{\hbar}{i} \frac{\partial}{\partial \phi} \left(\frac{e^{im\varphi}}{\sqrt{2\pi}} \right) = m\hbar Q_m(\phi) \quad (5.131)$$

2. Next, we solve the second Sturm-Liouville problem associated with $P(\theta)$. Using Equation 5.127

$$\frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial P}{\partial \theta} \right) - \frac{v}{\sin \theta} P = -\frac{\lambda}{\hbar^2} \sin \theta P \quad (5.132)$$

The “associated Legendre equation” defined by Equation 5.132 is in self-adjoint form. Making the substitution $x = \cos \theta$ eliminates the weight function $w(\theta) = \sin \theta$. Using terms such as $\partial_\theta = -(\sin \theta) \partial_x$ allows Equation 5.132 to be written as

$$-\hbar^2 \frac{d}{dx} \left[(1-x^2) \frac{dP(x)}{dx} \right] + \frac{\hbar^2 v_m}{1-x^2} P(x) = \lambda P(x) \quad (5.133)$$

where $v_m = m^2$, $m = 0, \pm 1, \pm 2, \dots$, and $\theta \in [0, \pi]$ requires $x \in [-1, 1]$. The eigenvalues for Equation 5.133 are λ and not v_m . Each operator

$$\hat{L}'_m = -\hbar^2 \frac{d}{dx} (1-x^2) \frac{d}{dx} + \frac{m^2 \hbar^2}{1-x^2} \quad (5.134)$$

is self-adjoint and gives rise to the eigenfunction equation $\hat{L}'_m P_\lambda(x) = \lambda P_\lambda(x)$. We are looking for the eigenvalues λ and the corresponding eigenfunctions $P_\lambda(x)$ —the so-called associated Legendre polynomials.

Here is an important point. For each m , there is an operator $\hat{L}'_m$ where $\hat{L}'_0 \neq \hat{L}'_1 \neq \hat{L}'_2 \dots$. For each m , there exists a basis set since $\hat{L}'_m = \hat{L}'_m{}^+$. Let the “script P ”, namely $\mathcal{P}$, be the normalized

eigenfunctions (i.e., the normalized $P_\lambda(x)$). For $m = 0$, the basis set is $\{\mathcal{P}_\lambda^{(1)}\}$ and the eigenvalues are $\{\lambda^{(1)}\}$. For $m = 1$, the basis set is $\{\mathcal{P}_\lambda^{(2)}\}$ and the eigenvalues are $\{\lambda^{(2)}\}$. The two sets of basis vectors are not necessarily the same since $\hat{L}_m \neq \hat{L}_n$ for $m \neq n$. We already know that the eigenvalues must be $\lambda_\ell = \ell(\ell + 1)\hbar^2$. We label the eigenvectors as $\mathcal{P}_\ell^m$.

We can find the eigenvectors and eigenvalues in several ways. The simplest method consists of setting $m = 0$, solve Legendre's equation and then apply the ladder operators to find the z -components. Refer to the chapter exercises. The orthonormal eigenfunctions are

$$\mathcal{P}_\ell^m(\theta) = N_{\ell m} P_\ell^m(\theta) = \sqrt{\frac{2\ell+1}{2}} \left[\frac{(\ell-m)!}{(\ell+m)!} \right]^{1/2} P_\ell^m(\theta) \quad (5.135)$$

where $\ell = 0, 1, 2, \dots$ and $m = 0, \pm 1, \pm 2, \dots, \pm \ell$. The “spherical harmonics” become

$$Y_\ell^m(\theta, \varphi) = (-1)^m \mathcal{P}_\ell^m(\theta) Q_m(\varphi) = (-1)^m [N_{\ell m} P_\ell^m(\theta)] \frac{e^{im\varphi}}{\sqrt{2\pi}} \quad (5.136)$$

the additional Condon–Shortley phase factor of $(-1)^m$ does not need to be included but appears to be conventional in honor of the authors of a classic text on atomic spectroscopy, Condon and Shortley. Some example spherical harmonics are given in Table 5.2. The normalization constant is $N_{\ell m} = \sqrt{\frac{2\ell+1}{2}} \left[\frac{(\ell-m)!}{(\ell+m)!} \right]^{1/2}$ such that $\langle \mathcal{P}_\ell^m | \mathcal{P}_{\ell'}^m \rangle = \delta_{\ell'\ell}$. In general, Rodrigues' formula can be incorporated into a generating function to give an implicit definition of the $\mathcal{P}_\ell^m(\theta)$

$$P_\ell^m(x) = \frac{1}{2^\ell \ell!} (1-x^2)^{m/2} \frac{d^{\ell+m}}{dx^{\ell+m}} (x^2-1)^\ell \quad (5.137)$$

with $x = \cos(\theta)$.

As with any basis set, we have orthonormality relations

$$\delta_{\ell'\ell} = \langle \sin \theta \mathcal{P}_\ell^m(\theta) | \mathcal{P}_{\ell'}^m(\theta) \rangle = \int_0^\pi d\theta \sin \theta N_{\ell m} P_\ell^m(\theta) N_{\ell' m} P_{\ell'}^m(\theta) \quad (5.138a)$$

$$\delta_{mm'} = \langle Q_m | Q_{m'} \rangle = \int_0^{2\pi} d\varphi Q_m^*(\varphi) Q_{m'}(\varphi) \quad (5.138b)$$

TABLE 5.2
List of Some Spherical Harmonics

ℓ	m	$N_{\ell m}$	$\mathcal{P}_\ell^m$	Basis $\mathcal{P}_\ell^m = N_{\ell m} P_\ell^m$	Spherical Harmonics $Y_\ell^m = (-1)^m \mathcal{P}_\ell^m Q_m$
0	0	$N_{00} = \frac{1}{\sqrt{2}}$	$P_0^1 = 1$	$\mathcal{P}_0^1 = \frac{1}{\sqrt{2}}$	$Y_0^0 = \frac{1}{\sqrt{4\pi}}$
1	1	$N_{11} = \sqrt{\frac{3}{4}}$	$P_1^1 = \sin \theta$	$\mathcal{P}_1^1 = \sqrt{\frac{3}{4}} \sin \theta$	$Y_1^1 = -\sqrt{\frac{3}{8\pi}} \sin \theta e^{i\varphi}$
1	0	$N_{10} = \sqrt{\frac{3}{2}}$	$P_1^0 = \cos \theta$	$\mathcal{P}_1^0 = \sqrt{\frac{3}{2}} \cos \theta$	$Y_1^0 = \sqrt{\frac{3}{4\pi}} \cos \theta$
1	-1	$N_{1-1} = \sqrt{\frac{3}{4}}$	$P_1^{-1} = -\sin \theta$	$\mathcal{P}_1^{-1} = -\sqrt{\frac{3}{4}} \sin \theta$	$Y_1^{-1} = \sqrt{\frac{3}{8\pi}} \sin \theta e^{-i\varphi}$

where the azimuthal angle θ requires the weight function in the inner product. Likewise, the spherical harmonics must satisfy a orthonormality relation that can be found from the previous two.

$$\delta_{m'm} \delta_{l'l} = \langle \langle \sin \theta Y_l^m(\theta, \varphi) | Y_{l'}^{m'}(\theta, \varphi) \rangle \rangle = \int_0^{2\pi} d\varphi \int_0^\pi d\theta \sin \theta [Y_l^m(\theta, \varphi)]^* [Y_{l'}^{m'}(\theta, \varphi)] \quad (5.138c)$$

5.6 INTRODUCTION TO SPIN AND SPINORS

Spin describes the intrinsic angular momentum property of a particle, which from a classical point of view, represents its rotation around an internal axis. However, the classical angular momentum is not well defined for point particles such as electrons. Further, measuring the spin of the electron along a given direction produces only one of two possible values—the spin is quantized. As a result, the quantum mechanical model for electron spin uses a 2-D Hilbert space with a complex number field. The complete set of operators to describe the fundamental spin states consists of the magnitude and z -component of the spin. We cannot simultaneously specify the spin angular momentum along more than one of the spatial axes.

5.6.1 BASIC IDEA OF SPIN

Classically, we picture the electron as a small particle spinning at fixed speed about an axis. The rotating mass has angular momentum $\vec{S}$ with a direction given by the right-hand rule. Moving charge produces a magnetic field. Because the electron has negative charge, the magnetic field (of the electron) $\vec{B}_e$ and hence its dipole moment $\vec{\mu}$ must point in a direction opposite to the angular momentum at the position of the electron. Figure 5.28 illustrates the relationship. Keep in mind that $\nabla \cdot \vec{B} = 0$ holds so that the magnetic field forms continuous loops and therefore does not everywhere point along $\vec{B}_e$ shown.

Any change in angular momentum must be related to an applied torque $\vec{\tau} = \vec{\mu} \times \vec{B}$. The magnitude of the electron spin cannot change, but its direction can change. The magnetic moment $\vec{\mu}$ of the electron comes from the spinning electron and the references relate it to the spin angular momentum

$$\vec{\mu} = -2 \frac{\mu_B}{\hbar} \vec{S}$$

where $\mu_B = \frac{e\hbar}{2m}$ represents the Bohr magneton.

As usual, we are most interested in the dynamics of a spin system. We therefore need the Hamiltonian that describes the interaction of the spin with an applied field. The word “interaction”

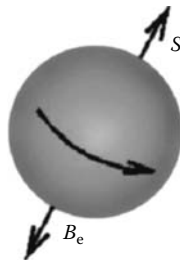


FIGURE 5.28 Classical picture of the rotating electron.

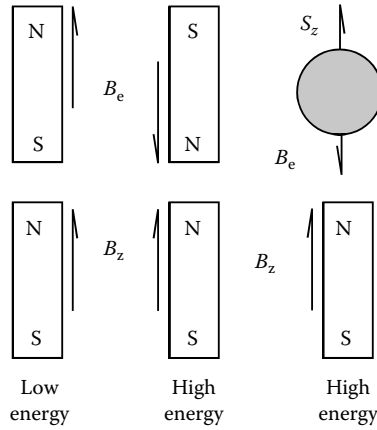


FIGURE 5.29 Upper row represents electrons as bar magnets and bottom shows the applied external fields along the z -direction due to bar magnets. “High energy” states are produced by the mutual repulsion between N poles of magnets (or equivalently, between S poles).

refers to the interaction between magnetic poles and external magnetic fields that gives the old adage: “like” poles of two magnets repel and “unlike” poles attract. The interaction energy of the electron (considering only the spin) with the external magnetic field must be related to the relative orientation between the electron spin and the external fields. As shown in Figure 5.29, the lowest energy comes from placing the electron “south pole” next to the external field “north pole” (or vice versa). With unlike poles together, the potential energy must be a minimum since a force must be applied to separate the magnets. The high-energy state occurs when “like” poles are pushed together which means the two magnetic fields must be antiparallel to each other. We therefore expect the total energy of the spinning electron to have the form $\hat{\mathcal{H}}_s \sim -\vec{B}_e \cdot \vec{B}$ where B represents the externally applied magnetic field. However, we also know that the magnetic field of the electron points in the opposite direction of its spin $\vec{S}$ and so the classical energy must be given by $\mathcal{H}_s \sim \vec{B} \cdot \vec{S}$. Including the necessary factors, the spin energy can be written as

$$\mathcal{H}_s = -\frac{q}{m} \vec{B} \cdot \vec{S} \quad (5.139)$$

where

m is the free mass of the electron

$\mu_B = \frac{e\hbar}{2m}$ is the Bohr magneton

$q = -e$ for electrons

Now we should examine some basic differences between classical and quantum spin. A macroscopically sized object can have any value of angular momentum and its direction can be arbitrarily close to the z -direction (or any other direction). As the object size decreases, it becomes apparent that measurements of the angular momentum produce discrete values and the total angular momentum never equals its measured value along one of the axis. Measurement of the spin produces only one value for the magnitude and two possible values for the z -component. Making measurements of the electron spin always produces one of two values typically called “spin up” and “spin down.” The spin direction refers to a direction set by the observer usually taken as the z -direction in the laboratory. A person making the measurement looks for the z -component of the spin. However, the person can equally well make a measurement along the laboratory x - or y -axes. Again, the person will find spin parallel or antiparallel to the axis. We usually take the “detection” direction as the z -direction. Figure 5.30 shows the lowest order approximate pictorial

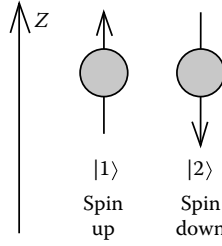


FIGURE 5.30 The basic idea of an electron having a “spin up” and “spin down.” In actuality, the spin vectors make an angle with the z -axis.

representations of the electron spin. The representation is only approximate because the figure implies that the magnitude of the spin S has the same value as the z -component S_z , which is not the case. We will discuss this last point more. In general, the spin wave function will be a linear combination of the spin-up and spin-down z -components (but the magnitude remains fixed). Making a measurement of the spin along a given direction causes the wave function to collapse so that the electron has either spin-up or spin-down (but not both).

The classical value of the spin, denoted by the vector $\vec{S}$, corresponds to the expectation value of the quantum mechanical operator $\hat{S}$

$$\vec{S} = \langle \psi_s | \hat{S} | \psi_s \rangle \quad (5.140)$$

where $|\psi_s\rangle$ represents the spin wave function. As will be discussed further in a subsequent section, the wave function represents the probability amplitude (“square” it to find the probability) that the electron will have spin “up” or “down” when measured. Further, as with other wave functions, it provides a superposition of basis states so that the electron simultaneously occupies the basis states. The superposition then provides an average as in Equation 5.140.

Although individual measurements of the z -component of spin yield discrete values, the average can assume a continuous range of values. For example, a large number of measurements on the same particle with identical preparation might yield the z -components of “up, up, down, up, down, . . .” Clearly the average of these values is not exactly “up” nor exactly “down” but somewhere in between the two values. The situation is similar to flipping a coin with one side having a value of 0 (i.e., “tails”) and a value of 1 (i.e., “heads”). After a very large number of “coin flip” experiments, the average of the sequence of 0’s and 1’s would be able to have any value between 0 and 1 despite the fact that the coin only has two values. For spin, the average itself comes from the mix of fundamental spin states that make up the spin wave function $|\psi_s\rangle$. We have seen how starting with the wave function, one can find the average. However, one can also start with the classical average and essentially deduce the wave function. The key point concerns the connection between the classical value of spin and the discrete quantum mechanical values.

We want to model the spin using Hilbert space. To start, consider a 2-D Hilbert space. Similar to the angular momentum discussed in the previous section, the basic spin states can be written as $|s, m\rangle = |s\rangle|m\rangle$ where the magnitude of the spin (squared) and z -component have eigenvalues

$$\hat{S}^2 |s, m\rangle = s(s+1)\hbar^2 |s, m\rangle \quad \hat{S}_z |s, m\rangle = m\hbar |s, m\rangle \quad \text{for } -s \leq m \leq s \quad (5.141)$$

However, the spin has only one value for s , specifically $s = 1/2$; the rate of spin of the quantum particle cannot be changed, only its direction! Therefore, parameter m describing the z -component can assume only the two values of $\pm 1/2$. The magnitude of the spin angular momentum $\sqrt{3}\hbar/2$ (from the

first of Equation 5.141) has a larger value than does the magnitude of the z -component $\hbar/2$ as expected. Because s does not change for a single electron, the notation for the fundamental states can be simplified by dropping the s parameter and using $|m\rangle$. That is, $|m\rangle$ represents a short-hand notation for the full vector $|m\rangle = |s, m\rangle = |s\rangle|m\rangle$. This observation also leads to the notion of “spin up” and “spin down” to represent the two states of the spin angular momentum (even though the z -components never have the same value as the magnitude of the total spin).

The spin-space basis vectors have the following definition. The basis vector $|1\rangle$ represents spin up and the basis vector $|2\rangle$ represents spin down. Basically two levels can represent the spin system (similar to two level atoms). We can write the basis vectors in several alternative forms

$$|1\rangle = \left| +\frac{1}{2} \right\rangle = |\uparrow\rangle \quad \text{and} \quad |2\rangle = |-1\rangle = \left| -\frac{1}{2} \right\rangle = |\downarrow\rangle \quad (5.142)$$

We will often switch between the various expressions of the basis vectors depending on the aspect being emphasized. The important point is that we represent the two physical directions in physical space by a basis vector in Hilbert space.

5.6.2 LINK BETWEEN PHYSICAL SPACE AND HILBERT SPACE

We want to know how to represent a 3-D object (spin vector) in a 2-D Hilbert space spanned by $\{|1\rangle, |2\rangle\}$. While the reduction in dimensions might appear to be “untenable,” the situation is further complicated by the fact that ket $|1\rangle$ refers to spin along the positive z -axis (1-D) and $|2\rangle$ represents the spin along the negative z -axis. For now we consider the spin as a classical object without regard to quantization. The present discussion links the average of the quantum mechanical average of spin $\langle \hat{S} \rangle \equiv \vec{S}$ with the Hilbert space wave function. The average of an operator in a state gives the classical value.

Consider the mathematical relation for the situation depicted in Figure 5.31. When $\vec{S}$ aligns with the z -axis, we expect the wave function $|\psi\rangle$ to align with $|1\rangle$ by definition of the basis vector. Likewise, $\vec{S}$ aligning with the negative z -axis requires the wave function $|\psi\rangle$ to align with $|2\rangle$. Therefore, when the angle θ varies from 0 to π , the angle γ in Hilbert space must vary from 0 to $\pi/2$. We conclude that $\gamma = \theta/2$. We know that the most general expression for ψ in Hilbert space has the form

$$\psi = \beta_1|1\rangle + \beta_2|2\rangle \quad (5.143)$$

where the coefficients β_i can be complex. The figure leads us to conclude

$$|\psi\rangle = |1\rangle \cos(\theta/2) + |2\rangle \sin(\theta/2) \quad (5.144)$$

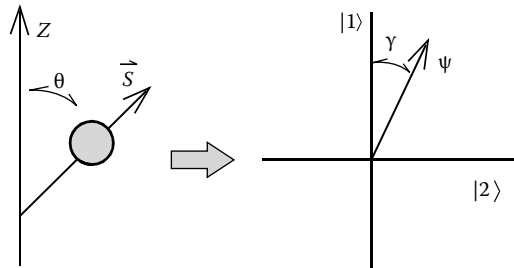


FIGURE 5.31 Mapping physical space (left) into Hilbert space (right). Note the basis vectors on the right have been switched for simpler comparison.

This wave function represents an electron with an average spin angular momentum pointing at angle θ with respect to the z -direction (although it still requires another parameter to describe the angle with respect to x - and y -directions). The validity of this last statement will become clear after discussing the Pauli operators. For now, it stands to reason that if $|1\rangle$ corresponds to the average spin at angle 0 and $|2\rangle$ corresponds to 180° , then a linear combination as in Equation 5.144 must correspond to an average spin pointing with an angle θ between 0° and 180° . Making a measurement causes the wave function to collapse to either $|1\rangle$ or $|2\rangle$ with corresponding probability of collapse of $\cos^2\theta$ and $\sin^2\theta$, respectively.

Sometimes matrix notation proves simpler than the Dirac notation. Using the matrix methods from linear algebra shows that the unit vectors in Equation 5.142 and the wave function in Equation 5.143 can be represented by column matrices

$$\text{Spin-up} = |1\rangle \leftrightarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{Spin-down} = |2\rangle \leftrightarrow \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad |\psi\rangle \leftrightarrow \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad (5.145)$$

These 2-D matrix column vectors are sometimes called spinors.

What about the components of the spin vector along the x - and y -directions? Although at first it might appear counterintuitive to represent 3-D vectors in a 2-D space, we can use the basis states $|1\rangle$ and $|2\rangle$ by making an intuitive use of the four parameters inherent to the two complex numbers β_1 and β_2 . Three of the parameters specify direction while the remaining one normalizes the wave function to 1. The full definitions for the coefficients β_i can be found by considering a specific example. Suppose the electron has its spin vector fully in the x - y plane (see Figure 5.32). In this case, the physical angle has the value $\theta = 90$ and so the angle in Hilbert space must be $\gamma = \frac{\theta}{2} = 45$.

Equation 5.144 leads us to write $|\psi\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. We still need to add another parameter while maintaining a normalization of 1. A phase factor involving a phase difference between the two components of the column vector can be included without affecting the normalization. Consider

$$|\psi\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ e^{i\varphi} \end{pmatrix}$$

The lower matrix element describes the physical x - y components of the classical spin vector $\vec{S}$ when it makes an angle φ with respect to the physical x -axis for the case shown in Figure 5.32. The angle φ in physical space matches the angle in Hilbert space (without angle doubling).

The full spinor can now be written for a spin vector pointing along arbitrary θ , φ as

$$\frac{1}{\sqrt{2}} \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\varphi} \sin \frac{\theta}{2} \end{pmatrix} \quad (5.146a)$$

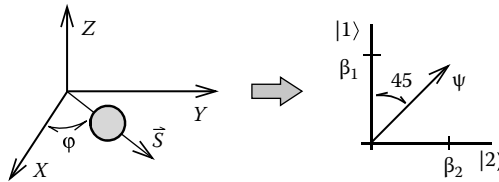


FIGURE 5.32 A particle with spin in the x - y plane (left) places the wave function at 45° with respect to the $|1\rangle$ axis in Hilbert space. The angle φ only affects the phase of the components.

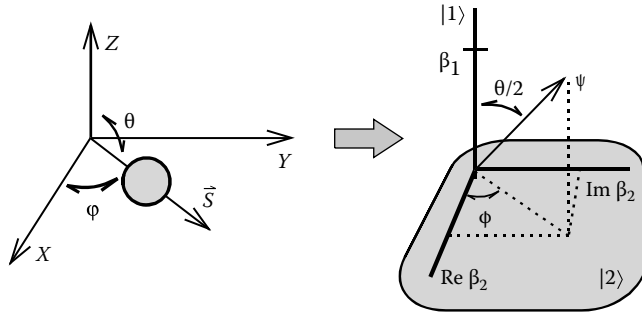


FIGURE 5.33 An intuitive map between the classical spin (average) and the quantum mechanical wave function.

or writing it as the vector in Hilbert space

$$|\psi\rangle = |1\rangle \cos \frac{\theta}{2} + |2\rangle e^{i\varphi} \sin \frac{\theta}{2} \quad (5.146b)$$

The Hilbert space vector can also be defined in a more symmetrical fashion as

$$|\psi\rangle = |1\rangle e^{-i\varphi/2} \cos \frac{\theta}{2} + |2\rangle e^{i\varphi/2} \sin \frac{\theta}{2} \quad (5.146c)$$

This wave function produces an “average” (i.e., classical) spin along the radial direction with angular coordinates (θ, φ) . We will use the first form in Equation 5.146b.

A figure helps make Equation 5.146b more intuitive although somewhat inaccurate from a mathematical point of view. Suppose the complex component $\beta_2|2\rangle$ is divided into a “real” axis and an “imaginary” axis as shown in Figure 5.33. The basis vector $|2\rangle$ then refers the complex plane represented by $\text{Re}(\beta_2) + i\text{Im}(\beta_2)$. That is, $|2\rangle$ simultaneously refers to the two “directions” described by the “real” and “imaginary” axes. Notice that the $\text{Re}\beta_2$ axis takes the place of the x -axis while the $\text{Im}\beta_2$ takes the place of the y -axis. Further recall from studies of complex variables that $e^{i\varphi}$ refers to a vector with unit length in the complex plane that makes an angle φ with respect to the real axis.

Figure 5.33 now has the following interpretation. The projection of the abstract vector $|\psi\rangle$ has the component $\cos(\theta/2)$ along the $|1\rangle$ axis. The component of $|\psi\rangle$ in the β_2 -plane (i.e., the $|2\rangle$ plane) is $\sin(\theta/2)$. Then Figure 5.33 shows $\text{Re}\beta_2 = \sin(\theta/2)\cos\varphi$, which is thought of as the physical x -component of the spin vector, and $\text{Im}\beta_2 = \sin(\theta/2)\sin\varphi$, which is thought of as the physical y -component. The value of β_2 is now written by combining the real and imaginary parts as $\beta_2 = \sin(\theta/2)e^{i\varphi}$ which agrees with Equation 5.146b. To say this another way, the classical vector $\vec{S}$ maps into the quantum mechanical wave function $|\psi\rangle$ simply by changing the angle θ to $\theta/2$.

$$\begin{aligned} \vec{S} &= \tilde{z} \cos(\theta) + \tilde{x} \sin(\theta) \cos(\varphi) + \tilde{y} \sin(\theta) \sin(\varphi) \\ \Rightarrow |\psi\rangle &\sim |1\rangle \cos(\theta/2) + [\sin(\theta/2) \cos(\varphi) (\text{Re-axis}) + i \sin(\theta/2) \sin(\varphi) (\text{Im-axis})] |2\rangle \end{aligned}$$

Or, to be more mathematically correct,

$$\begin{aligned} |\psi\rangle &= |1\rangle \cos(\theta/2) + [\sin(\theta/2) \cos(\varphi) + i \sin(\theta/2) \sin(\varphi)] |2\rangle \\ &= \cos(\theta/2) |1\rangle + \sin(\theta/2) e^{i\varphi} |2\rangle \end{aligned}$$

Using these relations, an expected classical value of the physical spin vector can be mapped into the corresponding Hilbert space vector and vice versa. The use of the Pauli operators will next be used to show how the wave functions produce the required average value of the spin vector.

5.6.3 PAULI SPIN MATRICES

We can use the link between the Hilbert space and the classical spin to establish the spin operator. We look for operators $\hat{S} = \tilde{x}\hat{S}_x + \tilde{y}\hat{S}_y + \tilde{z}\hat{S}_z$ that allow us to deduce the classical particle spin $\vec{S} = \langle \hat{S} \rangle$ as given in the previous section. Conventionally, the operators are defined in terms of the Pauli spin operators to remove pesky factors of Planck's constant.

$$\hat{S}_x = \frac{\hbar}{2}\hat{\sigma}_x \quad \hat{S}_y = \frac{\hbar}{2}\hat{\sigma}_y \quad \hat{S}_z = \frac{\hbar}{2}\hat{\sigma}_z \quad \rightarrow \quad \hat{S} = \frac{\hbar}{2}\vec{\sigma} = \frac{\hbar}{2}(\tilde{x}\hat{\sigma}_x + \tilde{y}\hat{\sigma}_y + \tilde{z}\hat{\sigma}_z) \quad (5.147)$$

The objective consists of finding the Pauli matrices.

The operator measuring the spin along the z -direction $\hat{\sigma}_z$ can be built from the eigenvalues and eigenvectors.

$$\hat{\sigma}_z|1\rangle = +1|1\rangle \quad \text{and} \quad \hat{\sigma}_z|2\rangle = -1|2\rangle \quad (5.148a)$$

In the usual manner prescribed in linear algebra, the operator can be written in the basis vector expansion as

$$\hat{\sigma}_z = |1\rangle\langle 1| - |2\rangle\langle 2| \quad \text{or} \quad \underline{\sigma}_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (5.148b)$$

We can now demonstrate the Pauli spin matrices for the x - and y -directions. We demonstrate the Pauli y -spin matrix and leave the one for the x -direction to the problems. Using Equation 5.146a with $\theta = 90$ and $\varphi = 90$ (i.e., the classical spin points along the y -axis in the laboratory), we want the following conditions to hold

$$\underline{\sigma}_y \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} = + \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} \quad \underline{\sigma}_y \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} = - \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} \quad (5.149)$$

That is, we want the column vector $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}$ to represent an average spin along the physical y -direction. We need to find an operator, $\hat{\sigma}_y$ in this case, that can be used as a type of “indicator” that tells us when a wave function produces a classical component parallel to the physical y -direction. We postulate that an eigenvalue of $+1$ indicates the component along $+y$, while -1 indicates a component along $-y$.

To deduce the Pauli y -spin operator $\hat{\sigma}_y$, we make an eigenvector expansion of $\hat{\sigma}_y$ by using these last two equations. Defining the eigenvector notation as

$$|e_1\rangle \equiv \begin{pmatrix} 1 \\ i \end{pmatrix} \quad \text{and} \quad |e_2\rangle \equiv \begin{pmatrix} 1 \\ -i \end{pmatrix} \quad (5.150a)$$

where the normalization has been temporarily suppressed for clarity.

These two column eigenvectors can be alternatively written as vectors in terms of the basis vectors as

$$|e_1\rangle = \frac{1}{\sqrt{2}}|1\rangle + \frac{i}{\sqrt{2}}|2\rangle \quad |e_2\rangle = \frac{1}{\sqrt{2}}|1\rangle - \frac{i}{\sqrt{2}}|2\rangle \quad (5.150b)$$

which now includes the normalization.

The eigenvector expansion of the y -spin operator

$$\hat{\sigma}_y = (+1)|e_1\rangle\langle e_1| + (-1)|e_2\rangle\langle e_2| \quad (5.151a)$$

can now be rewritten using Equation 5.150b

$$\hat{\sigma}_y = \frac{1}{2}(|1\rangle + i|2\rangle)(\langle 1| - i\langle 2|) - \frac{1}{2}(|1\rangle - i|2\rangle)(\langle 1| + i\langle 2|) \quad (5.151b)$$

Simplifying the equation provides

$$\hat{\sigma}_y = 0|1\rangle\langle 1| - i|1\rangle\langle 2| + i|2\rangle\langle 1| + 0|2\rangle\langle 2| \quad (5.151c)$$

which gives the Pauli y-spin matrix. The derivation of the Pauli matrix $\hat{\sigma}_x$ can be found in chapter problems. The three Pauli spin matrices are

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (5.152a)$$

The corresponding operators are

$$\hat{\sigma}_x = |1\rangle\langle 2| + |2\rangle\langle 1| \quad \hat{\sigma}_y = -i|1\rangle\langle 2| + i|2\rangle\langle 1| \quad \hat{\sigma}_z = |1\rangle\langle 1| - |2\rangle\langle 2| \quad (5.152b)$$

The magnitude of the spin can be deduced from the fact that the spin projected onto each physical axis must be $\pm\hbar/2$. The magnitude-squared of the total spin vector $|\vec{S}| = S$ must be

$$S^2 = \left(\frac{\hbar}{2}\right)^2 + \left(\frac{\hbar}{2}\right)^2 + \left(\frac{\hbar}{2}\right)^2 = \frac{3}{4}\hbar^2 \quad (5.153)$$

The magnitude of the spin S can be written similar to that for the orbital angular momentum

$$S = \hbar\sqrt{\frac{3}{4}} = \hbar\frac{\sqrt{3}}{2}$$

The magnitude S is larger than the z -component of the spin $|S_z| = \hbar/2$. The Pauli spin operators produce the same result

$$\hat{S}^2 = \left(\frac{\hbar}{2}\right)^2 \left\{ \hat{\sigma}_x^2 + \hat{\sigma}_y^2 + \hat{\sigma}_z^2 \right\} = \left(\frac{\hbar}{2}\right)^2 \{3\hat{1}\}$$

The magnitude operator is therefore

$$\hat{S} = \sqrt{\frac{3}{4}} \hbar \hat{1} \quad (5.154)$$

A classical picture of the relation between the spin and z -component appears in Figure 5.34. Since the magnitude of the spin is larger than the z -component, the spin vector must actually point away from the z -axis. The projection of the spin vector onto the z -axis gives S_z .

The problems show the commutation relations

$$[\hat{\sigma}_i, \hat{\sigma}_j] = 2i\varepsilon_{ijk}\hat{\sigma}_k \quad (\text{sum convention}) \quad \sum_i \hat{\sigma}_i^2 = 3\hat{1}$$

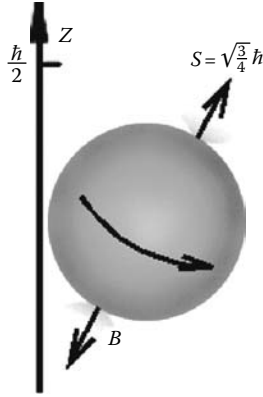


FIGURE 5.34 The total spin does not point along the z-axis.

Example 5.8

Use the Pauli y -spin matrix to show that $\theta = 90$ and $\varphi = 90$ in Equation 5.146 produces average spin along the y -axis.

SOLUTION

Operate on the column vector $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ e^{i\varphi} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}$ to find

$$\sigma_y \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix} = +1 \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}$$

So the column vector is an eigenvector of the y -Pauli operator, which is proportional to the y -spin operator, and so the column vector represents a particle with the average spin oriented along the y -direction.

Example 5.9

Find the average value of the spin operator $\hat{S} = \tilde{x}\hat{S}_x + \tilde{y}\hat{S}_y + \tilde{z}\hat{S}_z$ for the state $|\psi\rangle = |1\rangle \cos \frac{\theta}{2} + |2\rangle e^{i\varphi} \sin \frac{\theta}{2}$.

SOLUTION

This can be worked using either matrices or operators. Let us use the matrices.

$$\begin{aligned} \langle \psi | \hat{S} | \psi \rangle &= \frac{\hbar}{2} \begin{pmatrix} \cos \frac{\theta}{2} & e^{-i\varphi} \sin \frac{\theta}{2} \end{pmatrix} \left[\tilde{x} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \tilde{y} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} + \tilde{z} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right] \begin{pmatrix} \cos \frac{\theta}{2} \\ e^{i\varphi} \sin \frac{\theta}{2} \end{pmatrix} \\ &= \frac{\hbar}{2} [\tilde{x} \cos \varphi \sin \theta + \tilde{y} \sin \varphi \sin \theta + \tilde{z} \cos \theta] \end{aligned}$$

In this case, the average spin will point in the direction specified by the two angles θ, ϕ which correspond to angles with respect to the z - and x -axes, respectively.

5.6.4 ROTATIONS

Recall that the rotation of the function $|\psi\rangle$ in Hilbert space corresponding to a rotation in 3-D space can be affected by the rotation operator $\hat{R} = e^{-i\vec{\alpha}\cdot\vec{L}/\hbar}$. We use the rotation (hence the

“−” sign) that rotates the functions (or objects) in the positive angle direction. This means that we can find the spinors (as an alternate method to that described in the previous section) corresponding to a particle with x and y spin components as follows:

$$|e_x\rangle = \hat{R}\left(\theta = \frac{\pi}{2}\right) |1\rangle = e^{-i\theta\hat{S}_y/\hbar} |1\rangle = e^{-i\theta\hat{\sigma}_y/2} |1\rangle$$

It is easier to work with the matrix form of this equation

$$|e_x\rangle \rightarrow e^{-i\theta\hat{\sigma}_y/2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \left\{ 1 - i\frac{\theta}{2}\hat{\sigma}_y + \frac{1}{2}\left(-i\frac{\theta}{2}\hat{\sigma}_y\right)^2 + \dots \right\} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cos\frac{\theta}{2} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \sin\frac{\theta}{2} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

where we used $\hat{\sigma}_y\hat{\sigma}_y = \hat{1}$. We see that every operation in physical space has a corresponding operation in the Hilbert space.

5.6.5 DIRECT PRODUCT SPACE FOR A SINGLE ELECTRON

The complete description of the electron requires both the spin and translational states. The spin and translational states refer to two separate aspects of the electron. The translational part refers to the scalar wave function obtained, for example, for the infinitely deep well or the traveling plane wave. The spin and translational motion are described by a direct product space. The Hamiltonian can link the motion in the two spaces. If $|s_z\rangle$ (where $s_z = \pm 1$) represents the spin basis vectors and $|n\rangle$ represents the translational states such as the sinusoidal solutions for the infinitely deep well, then the basis vectors representing both aspects of the electron span a direct product space and have the form

$$\{|\phi_n\rangle = |s_z n\rangle = |s_z\rangle|n\rangle\} = \{|s_z = 1\rangle|n = 1\rangle, |s_z = 2\rangle|n = 1\rangle, |s_z = 1\rangle|n = 2\rangle, \dots\} \quad (5.155)$$

which includes all possible combinations of the basis vectors in each individual space.

The general wave function in the direct product Hilbert space must be a summation over the basis vectors in the direct product space.

$$|\psi\rangle = \sum_{s_z, n} \beta_{s_z, n} |s_z\rangle|n\rangle \quad (5.156a)$$

The Hilbert space vector can be written in several other forms

$$|\psi\rangle = \sum_{s_z, n} \beta_{s_z, n} |s_z\rangle|n\rangle = |1\rangle \sum_n \beta_{1, n} |n\rangle + |2\rangle \sum_n \beta_{2, n} |n\rangle \quad (5.156b)$$

where however, the coefficients β_{ab} should not be divided into two factors $\beta_{ab} = \beta_a \beta_b$ since it cannot be done uniquely and therefore would be meaningless. In matrix notation, the wave function would be

$$\psi = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \sum_n \beta_{1, n} |n\rangle + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \sum_n \beta_{2, n} |n\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \alpha_1 + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \alpha_2 = \begin{pmatrix} \alpha_1(\vec{r}) \\ \alpha_2(\vec{r}) \end{pmatrix} \quad (5.156c)$$

where

$$\alpha_1(\vec{r}) = \sum_n \beta_{1, n} u_n(\vec{r}) \quad \alpha_2(\vec{r}) = \sum_n \beta_{2, n} u_n(\vec{r}) \quad (5.156d)$$

and $u_n(\vec{r})$ is the translational state in the coordinate representation. One should keep in mind those equalities in Equation 5.146c hold by virtual of the isomorphism among kets, matrices and coordinate representations.

Example 5.10

At $t=0$ a particle is in a superposed state with spin up but without any classical x - and y -component. The particle is in the lowest level of an infinitely deep quantum well. Find the full wave function at $t=0$.

SOLUTION

The wave function can be written in a number of ways.

$$|s_z = 1, n = 1\rangle \sim \begin{pmatrix} 1 \\ 0 \end{pmatrix} \frac{1}{\sqrt{2}} \sin \frac{\pi x}{L} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sin \frac{\pi x}{L} \\ 0 \end{pmatrix}$$

5.6.6 SPIN HAMILTONIAN

We are most interested in the time-development of a spin system, its interaction with other types of systems, and the available energy states. The starting point consists of writing a suitable Hamiltonian. The energy of an electron magnetic dipole depends on its orientation with respect to an applied magnetic field $\vec{B}$. For spin $\hat{S} = \frac{\hbar}{2} [\tilde{x}\hat{\sigma}_x + \tilde{y}\hat{\sigma}_y + \tilde{z}\hat{\sigma}_z]$, the Hamiltonian can be written as

$$\hat{\mathcal{H}}_s = -\frac{q}{m} \vec{B} \cdot \hat{S} = \mu_B \vec{B} \cdot \hat{\sigma} \quad (5.157)$$

where

m is the free mass of the electron

$q = -e$ is its charge

$\mu_B = \frac{e\hbar}{2m}$ is the Bohr magneton for the electron

$\hat{\sigma} = \tilde{x}\hat{\sigma}_x + \tilde{y}\hat{\sigma}_y + \tilde{z}\hat{\sigma}_z$ is the Pauli spin vector (Figure 5.35)

The Hamiltonian in Equation 5.157 can be rewritten in a variety of notations. The matrix notation appears to be the most common. The notation combines the components of B , the externally applied field, and the matrices for the spin.

$$\hat{\mathcal{H}}_s = \mu_B \vec{B} \cdot \hat{\sigma} = \mu_B (B_x \hat{\sigma}_x + B_y \hat{\sigma}_y + B_z \hat{\sigma}_z)$$

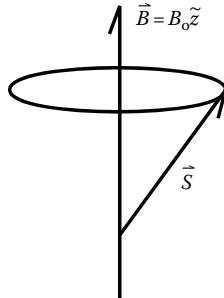


FIGURE 5.35 The spin vector precesses about the magnetic field.

We are treating the spin dynamical variables as operators but not the externally applied field B . Using the definitions for the Pauli spin matrices in Equation 5.152, we find

$$\hat{\mathcal{H}}_s = \mu_B \begin{pmatrix} B_z & B_x - iB_y \\ B_x + iB_y & -B_z \end{pmatrix} \quad (5.158)$$

The Schrödinger wave equation (SWE) for spin becomes

$$\hat{\mathcal{H}}_s \Psi = i\hbar \frac{\partial}{\partial t} \Psi, \quad \mu_B \begin{pmatrix} B_z & B_x - iB_y \\ B_x + iB_y & -B_z \end{pmatrix} \begin{pmatrix} \alpha_1(t) \\ \alpha_2(t) \end{pmatrix} = i\hbar \frac{\partial}{\partial t} \begin{pmatrix} \alpha_1(t) \\ \alpha_2(t) \end{pmatrix} \quad (5.159)$$

Example 5.11

Find the general solution to the SWE when the externally applied magnetic field has only a z -component and assume that the Hamiltonian does not include any translational component.

SOLUTION

The Schrödinger equation can be separated into two parts as usual

$$\hat{\mathcal{H}}_s |V\rangle = E|V\rangle \quad i\hbar \frac{\partial}{\partial t} T = ET \quad \text{where } \Psi = |V\rangle T(t)$$

and $|V\rangle$ refers to the Hilbert space spinor.

Before continuing with the solution, notice the separation of variables in $\hat{\mathcal{H}}_s |V\rangle T = i\hbar \partial_t |V\rangle T$ or equivalently $T \hat{\mathcal{H}}_s |V\rangle = i\hbar |V\rangle \partial_t T$ might be thought of (but never written) as $\hat{\mathcal{H}}_s |V\rangle / |V\rangle = i\hbar (\partial_t T / T)$ or using a separation constant E , $\hat{\mathcal{H}}_s |V\rangle / |V\rangle = E = i\hbar (\partial_t T / T)$; however, this division by a vector has not been mathematically defined. For the case of eigenvectors, one instead might define division through an inner product by applying $\langle V|$ to find $T \langle V| \hat{\mathcal{H}}_s |V\rangle = i\hbar \langle V|V\rangle \partial_t T$. Using the fact that the eigenvectors are normalized to unity and using a separation constant E , one can write $\langle V| \hat{\mathcal{H}}_s |V\rangle = E = i\hbar \partial_t T / T$. Now the differential equation for T is obtained in the normal manner. The vector portion can be rewritten by using the fact that for eigenvectors, $\langle V| \hat{\mathcal{H}}_s |V\rangle = E$ is equivalent to $\hat{\mathcal{H}}_s |V\rangle = E|V\rangle$.

Continuing now with the solution, the Hamiltonian only has diagonal components

$$\begin{pmatrix} \mu_B B_z & 0 \\ 0 & -\mu_B B_z \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = E \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \quad (5.160)$$

If the Hamiltonian is not diagonal, one finds the eigenvectors using techniques from Chapter 3. This is equivalent to diagonalizing the Hamiltonian and then applying the inverse transformation to the resulting column vector. The normalized eigenvalues and eigenvectors must be $E_1 = \mu_B B_z$ for spin up (i.e., when $\alpha_2 = 0$) and $E_2 = -\mu_B B_z$ for spin down (i.e., $\alpha_1 = 0$), and

$$|1\rangle = |\uparrow\rangle = |E_1\rangle = |\mu_B B_z\rangle \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad |2\rangle = |\downarrow\rangle = |E_2\rangle = |-\mu_B B_z\rangle \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Consequently, the general solution to the SWE must be

$$\Psi = |V\rangle T \rightarrow \Psi(t) = \beta_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{\frac{E_1 t}{\hbar}} + \beta_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{\frac{E_2 t}{\hbar}} \quad (5.161)$$

Example 5.12

For the previous example, find the solution when the particle initially has spin up.

SOLUTION

The initial wave vector has the form

$$\Psi(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \sim |\uparrow\rangle$$

While the time-dependent solution has the form

$$\Psi(t) = \beta_1 |\uparrow\rangle e^{\frac{E_1 t}{\hbar}} + \beta_2 |\downarrow\rangle e^{\frac{E_2 t}{\hbar}} \rightarrow \Psi(0) = \beta_1 |\uparrow\rangle + \beta_2 |\downarrow\rangle$$

Therefore, projecting $\Psi(0) = \beta_1 |\uparrow\rangle + \beta_2 |\downarrow\rangle$ on to the basis vectors $|\uparrow\rangle, |\downarrow\rangle$, produces the constants β_i which have the values

$$\begin{aligned} \beta_1 &= \langle \uparrow | \Psi(0) \rangle = 1 \quad \text{or} \quad 1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}^+ \left\{ \beta_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} = \beta_1 \\ \beta_2 &= \langle \downarrow | \Psi(0) \rangle = 0 \quad \text{or} \quad 0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^+ \left\{ \beta_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} = \beta_2 \end{aligned}$$

Example 5.13

Find solution to the SWE

$$\hat{\mathcal{H}}\Psi = i\hbar \frac{\partial}{\partial t} \Psi$$

describing both translation and spin in an infinitely deep well of width L and with the bottom of the well at $V=0$ for $x=0$ to $x=L$.

SOLUTION

The Hamiltonian has the form

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_s + \hat{\mathcal{H}}_t \quad (5.162)$$

where the translational Hamiltonian has the coordinate form

$$\hat{\mathcal{H}}_t = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2}$$

Separating variables with

$$\psi = |V\rangle X(x) T(t) \quad (5.163)$$

gives

$$(\hat{\mathcal{H}}_s + \hat{\mathcal{H}}_t) |V\rangle X(x) = E |V\rangle X(x) \quad i\hbar \frac{\partial}{\partial t} T = ET \quad (5.164)$$

where E represents the total energy. Since the two Hamiltonians in Equation 5.164 do not share any coordinates, we can separate variables again

$$\hat{\mathcal{H}}_s|V\rangle = E^{(s)}|V\rangle \quad \hat{\mathcal{H}}_tX(x) = E^{(t)}X(x)$$

where $E = E^{(s)} + E^{(t)}$

We already know the solution to each time-independent equation. The spin produces the basis vectors $|\uparrow\rangle, |\downarrow\rangle$ with corresponding energy $E_m^{(s)} = (-1)^{m+1}\mu_B B_z$ and $m = 1, 2$. The infinitely deep well has basis vectors $X_n(x) = \sqrt{\frac{2}{L}}\sin(k_n x)$ with energy eigenvalues $E_n^{(t)} = \frac{\hbar^2 k_n^2}{2m}$. The basis set must be the direct product set

$$\{|v_m\rangle = |\uparrow\rangle, |\downarrow\rangle\} \otimes \left\{X_n(x) = \sqrt{\frac{2}{L}}\sin(k_n x)\right\} = \{|v_m\rangle X_n = |\uparrow\rangle X_n, |\downarrow\rangle X_n \quad \text{for } n = 1, 2, \dots\}$$

and total energy

$$E_{m,n} = E_m^{(s)} + E_n^{(t)} = (-1)^{m+1}\mu_B B_z + \frac{\hbar^2 k_n^2}{2m}$$

The general solution to the SWE must be

$$\Psi = \sum_{m,n} \beta_{mn} |v_m\rangle X_n e^{\frac{E_{mn}t}{\hbar}}$$

Example 5.14

Show that the electron dipole precesses around the direction of the magnetic field assumed to be along the z-axis ($\vec{B} = B_0 \hat{z}$). Assume that the spin starts along the x-axis

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

SOLUTION

A number of methods can be used to solve this problem.

1. A method similar to that used in the previous problems starts by looking at the Schrödinger equation for spin

$$\begin{aligned} \hat{\mathcal{H}}_s \begin{pmatrix} a \\ b \end{pmatrix} &= i\hbar \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} \rightarrow \mu_B B_z \hat{\sigma}_z \begin{pmatrix} a \\ b \end{pmatrix} = i\hbar \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} \rightarrow \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix} = \frac{\mu_B B_z}{i\hbar} \begin{pmatrix} a \\ -b \end{pmatrix} \\ \begin{pmatrix} a \\ b \end{pmatrix} &= \frac{1}{\sqrt{2}} \begin{pmatrix} e^{-i\Omega t/2} \\ e^{i\Omega t/2} \end{pmatrix} \quad \text{where } \Omega = 2\mu_B B_z/\hbar \end{aligned}$$

This gives a vector rotating around the z-axis.

2. A simpler method consists of applying the evolution operator to the initial spin wave function. A spin-only evolution operator is

$$\hat{U}(t) = e^{-i\hat{\mathcal{H}}_s t/\hbar} = e^{-i\mu_B B_0 \hat{\sigma}_z t/\hbar}$$

Applying this to the initial state produces

$$e^{-i\mu_B B_0 \hat{\sigma}_z t/\hbar} \left(\frac{1}{\sqrt{2}}|1\rangle + \frac{1}{\sqrt{2}}|2\rangle \right) = \frac{1}{\sqrt{2}}e^{-i\mu_B B_0 t/\hbar}|1\rangle + \frac{1}{\sqrt{2}}e^{+i\mu_B B_0 t/\hbar}|2\rangle$$

where we have used results similar to

$$e^{\lambda \hat{\mathcal{H}}_s}|1\rangle = \sum_m \frac{1}{m!} (\lambda \hat{\mathcal{H}}_s)^m |1\rangle = \sum_m \frac{1}{m!} (\lambda E_1)^m |1\rangle = e^{\lambda E_1}|1\rangle$$

We therefore find that the two methods agree.

3. A third method uses the Heisenberg representation of $\hat{\sigma}_x$ and expands using the Baker–Campbell–Hausdorff theorem

$$\hat{\sigma}_{xh}(t) = e^{-i\mu_B B_0 t \sigma_z/\hbar} \hat{\sigma}_x e^{i\mu_B B_0 t \sigma_z/\hbar} = \hat{\sigma}_x + (-ia)[\hat{\sigma}_z, \hat{\sigma}_x] + \frac{(-ia)^2}{2} [\hat{\sigma}_z, [\hat{\sigma}_z, \hat{\sigma}_x]] + \dots$$

Evaluating the commutators and separating the x and y spin matrices provides

$$\hat{\sigma}_{xh}(t) = e^{-i\mu_B B_0 t \sigma_z/\hbar} \hat{\sigma}_x e^{i\mu_B B_0 t \sigma_z/\hbar} = \hat{\sigma}_x \cos\left(\frac{2\mu_B B_0}{\hbar} t\right) + \hat{\sigma}_y \sin\left(\frac{2\mu_B B_0}{\hbar} t\right)$$

The components of the spin vector switch identity from x to y and back. We would find similar behavior in the y -component. The z -component remains stationary. Therefore, the spin vector must rotate about the z -direction with an angular rate of $\Omega = 2\mu_B B_0/\hbar$.

5.7 ANGULAR MOMENTUM FOR MULTIPLE SYSTEMS

Many dynamical variables and system parameters for the physical world use mathematical addition to find a total such as for momentum or energy. We already know that the scalar quantity of energy can be added. What about vector quantities such as angular momentum? Obviously the linear operators can be added. The real question concerns the basis states for two (or more) particles. Each particle exists in its own angular momentum Hilbert space. We can combine the complete set of operators for both spaces and use the direct product of basis states for each space. However, the total angular momentum provides a type of coupling for the independent basis states.

5.7.1 ADDING ANGULAR MOMENTUM

Consider two particles in a box distinguished by the subscripts 1 and 2. Assume that each one has angular momentum $\vec{J}_i$ ($i = 1, 2$). We use the symbol $\vec{J}$ for angular momentum so as to include either orbital angular momentum or spin. The spin can be classified according to half integral spin for fermions (such as electrons) or integral spin for bosons. The z -component will be denoted by $\hat{M}_i$ for notational convenience.

The angular momentum for each particle can be represented similar to Figure 5.36. The complete set of operators consists of union of sets $\{\hat{J}_1^2, \hat{M}_1\} \cup \{\hat{J}_2^2, \hat{M}_2\}$. Therefore, the basis states for the combined systems must span a direct product space according to

$$|j_1 m_1\rangle |j_2 m_2\rangle = |j_1 j_2 m_1 m_2\rangle \quad (5.165)$$

We often refer to this basis set as the uncoupled set. We already know the eigenvalues for the operators.

$$\hat{J}_i^2 |j_1 j_2 m_1 m_2\rangle = \hbar^2 j_i(j_i + 1) |j_1 j_2 m_1 m_2\rangle \quad \hat{M}_i |j_1 j_2 m_1 m_2\rangle = m_i \hbar |j_1 j_2 m_1 m_2\rangle \quad (5.166)$$

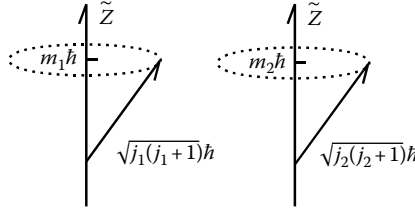


FIGURE 5.36 Two particles with separate angular momentum.

where $\{j_i = \frac{1}{2}, \frac{3}{2}, \dots\}$ for fermions, $\{j_i = 0, 1, \dots\}$ for bosons, and $-j_i \leq m_i \leq j_i$. The usual commutation rules hold for each space.

$$[\hat{J}_i^2, \hat{M}_i] = 0 \quad [\hat{J}_{(i)a}, \hat{J}_{(i)b}] = 2i\epsilon_{abc}\hat{J}_{(i)c} \quad (5.167a)$$

Note the “Einstein sum convention” should be applied to repeated indices in the last term. In addition, all operators between spaces commute such as

$$[\hat{J}_1^2, \hat{J}_2^2] = 0 \quad (5.167b)$$

since they refer to separate spaces.

Suppose we want to find the total angular momentum for a particle. The total angular momentum is $\hat{\mathbf{J}} = \hat{\mathbf{J}}_1 + \hat{\mathbf{J}}_2$ with a z -component of $\hat{M} = \hat{M}_1 + \hat{M}_2$. The magnitude (squared) of the total angular momentum is $\hat{J}^2 = \hat{\mathbf{J}} \cdot \hat{\mathbf{J}} = \hat{J}_1^2 + \hat{J}_2^2 + 2\hat{J}_1 \cdot \hat{J}_2$. The decoupled state $|j_1 j_2 m_1 m_2\rangle$ is not an eigenstate of $\hat{J}^2$ because of the x - and y -components occurring in the $\hat{J}_1 \cdot \hat{J}_2$ part of $\hat{J}^2$. We need to find different basis states instead of the uncoupled ones.

First, identify the complete set of operators since they lead to the alternate basis set. The magnitude (squared) of the angular momentum $\hat{J}^2$ commutes with its z -component as appropriate for angular momentum $[\hat{J}^2, \hat{M}] = 0$. We can also show these operators commute with $\{\hat{J}_1^2, \hat{J}_2^2\}$ and no others. The complete set of operators $\{\hat{J}_1^2, \hat{J}_2^2, \hat{J}^2, \hat{M}\}$ produces a complete description of the angular momentum of the two particles and induces the coupled basis set. This operator set must be equivalent to the uncoupled set $\{\hat{J}_1^2, \hat{J}_2^2, \hat{M}_1, \hat{M}_2\}$. Both sets produce basis vectors that span the same Hilbert space and therefore must be rotations of one another. The coupled basis set must be given by $\{|j, m, j_1, j_2\rangle\}$.

We can convert from one basis set to another using the closure relation.

$$\sum_{\substack{m'_1, m'_2 \\ j'_1, j'_2}} |m'_1 m'_2 j'_1 j'_2\rangle \langle m'_1 m'_2 j'_1 j'_2| = \hat{1} \quad (5.168)$$

The conversion between basis sets becomes

$$|j m j_1 j_2\rangle = \sum_{\substack{m'_1, m'_2 \\ j'_1, j'_2}} |m'_1 m'_2 j'_1 j'_2\rangle \langle m'_1 m'_2 j'_1 j'_2 | j m j_1 j_2\rangle$$

First notice that we must have $j'_1 = j_1$ and $j'_2 = j_2$ to give

$$|j m j_1 j_2\rangle = \sum_{m'_1, m'_2} |m'_1 m'_2 j_1 j_2\rangle \langle m'_1 m'_2 j_1 j_2 | j m j_1 j_2\rangle$$

Finally, we require $m = m_1 + m_2$ since $\hat{M} = \hat{M}_1 + \hat{M}_2$ to get

$$|jmj_1j_2\rangle = \sum_{\substack{m_1, m_2 \text{ such that} \\ m = m_1 + m_2}} |m_1m_2j_1j_2\rangle \langle m_1m_2j_1j_2 | jmj_1j_2\rangle \quad (5.169)$$

where note the change in the summation index. The transformation coefficients $\langle m_1m_2j_1j_2 | jmj_1j_2\rangle$ are called the Clebsch–Gordon coefficients. The next section demonstrates how to find these coefficients.

Suppose we start with two particles in the state $|jmj_1j_2\rangle$ and make a measurement of the uncoupled observables $\{\hat{J}_1^2, \hat{J}_2^2, \hat{M}_1, \hat{M}_2\}$. The probability of finding the particles in the state $|m_1m_2j_1j_2\rangle$ must be related to the Clebsch–Gordon coefficients by

$$|\langle m_1m_2j_1j_2 | jmj_1j_2\rangle|^2 \quad (5.170)$$

Example 5.15

Show $[\hat{J}^2, \hat{J}_1^2] = 0$

SOLUTION

$$\begin{aligned} N[\hat{J}^2, \hat{J}_1^2] &= [\hat{J}_1^2 + \hat{J}_2^2 + 2(\hat{J}_{1x}\hat{J}_{2x} + \hat{J}_{1y}\hat{J}_{2y} + \hat{J}_{1z}\hat{J}_{2z}), \hat{J}_1^2] = 2[\hat{J}_{1x}\hat{J}_{2x} + \hat{J}_{1y}\hat{J}_{2y} + \hat{J}_{1z}\hat{J}_{2z}, \hat{J}_1^2] \\ &= 2[\hat{J}_{1x}, \hat{J}_1^2]\hat{J}_{2x} + [\hat{J}_{1y}, \hat{J}_1^2]\hat{J}_{2y} + [\hat{J}_{1z}, \hat{J}_1^2]\hat{J}_{2z} = 0 \end{aligned}$$

Example 5.16

Two particles are in the state $|j=2, m=2, j_1=1, j_2=1\rangle$ what must be the state given by $|m_1m_2j_1j_2\rangle$?

SOLUTION

First note that $2 = m = m_1 + m_2$. However $j_1=1, j_2=1$ means that neither m_1 nor m_2 can be larger than 1. Therefore we require $m_1 = m_2 = 1$. Therefore $|j=2, m=2, j_1=1, j_2=1\rangle$ and $|m_1=1, m_2=1, j_1=1, j_2=1\rangle$ must be the “same” basis state.

Example 5.17

For $|j=2, m=2, j_1=1, j_2=1\rangle$, what is the probability of finding the particle in the state $|m_1=1, m_2=1, j_1=1, j_2=1\rangle$?

SOLUTION

The probability is given by Equation 5.170

$$P = |\langle m_1m_2j_1j_2 | jmj_1j_1\rangle|^2 = |\langle m_1=1, m_2=1, j_1=1, j_2=1 | j=2, m=2, j_1=1, j_2=1\rangle|^2 = 1$$

5.7.2 CLEBSCH–GORDON COEFFICIENTS

Consider two particles with given j_1 and j_2 .

Particle 1

Angular momentum: J_1^2 operator

z -Component: M_1 operator

Hilbert space: $\{|j_1 m_1\rangle: -j_1 \leq m_1 \leq j_1\}$

Particle 2

Angular momentum: J_2^2 operator

z -Component: M_2 operator

Hilbert space: $\{|j_2 m_2\rangle: -j_2 \leq m_2 \leq j_2\}$

Put both particles together in one box. We therefore consider the direct product space for the combined system. For right now, consider the particles to be uncoupled.

$$B_u = \{|j_1 j_2 m_1 m_2\rangle = |j_1 m_1\rangle |j_2 m_2\rangle\}$$

There is an alternate basis set that spans the combined space

$$B_c = \{|j_1 j_2 j m\rangle: |j_1 - j_2| \leq j \leq j_1 + j_2, -j \leq m \leq j\}$$

which is the “coupled set.” We know $j \leq j_1 + j_2$ since $m \leq j$ and $m = m_1 + m_2$ and $m_1 \leq j_1$ and $m_2 \leq j_2$. The combined space considers the total angular momentum $\vec{J} = \vec{J}_1 + \vec{J}_2$ to be a constant of the motion

$$J_1 |j_1 j_2 j m\rangle = j_1(j_1 + 1) |j_1 j_2 j m\rangle$$

$$J_2 |j_1 j_2 j m\rangle = j_2(j_2 + 1) |j_1 j_2 j m\rangle$$

Note that the factor of $\hbar^2$ has been suppressed for convenience. The sets B_u and B_c span the same space. J^2 and M are not new constants. An arbitrary vector in the space can be expanded in either B_u or B_c . A basis vector in the alternate basis set B_c can be written as a sum over the original basis set B_u by using the completeness relation for the original set $\sum_{m_1, m_2} |j_1 j_2 m_1 m_2\rangle \langle j_1 j_2 m_1 m_2| = 1$

$$|j_1 j_2 j m\rangle = \left[\sum_{m_1, m_2} |j_1 j_2 m_1 m_2\rangle \langle j_1 j_2 m_1 m_2| \right] |j_1 j_2 j m\rangle = \sum_{m_1, m_2} |j_1 j_2 m_1 m_2\rangle \langle j_1 j_2 m_1 m_2 | j_1 j_2 j m\rangle$$

This transformation allows one to change between basis sets as might be appropriate for a given problem. The inner products $\langle j_1 j_2 m_1 m_2 | j_1 j_2 j m\rangle$ are called the Clebsch–Gordon coefficients which we need to find. Notice that j_1 and j_2 occur in all the kets and bras and, for convenience of notation, they are sometimes omitted from the equation.

$$|j m\rangle = \sum_{m_1, m_2} |m_1 m_2\rangle \langle m_1 m_2 | j m\rangle$$

However, one must keep the restraints in mind

$$m = m_1 + m_2 \quad -j \leq m \leq +j \quad -j_1 \leq m_1 \leq +j_1 \quad -j_2 \leq m_2 \leq +j_2 \quad |j_1 - j_2| \leq j \leq j_1 + j_2$$

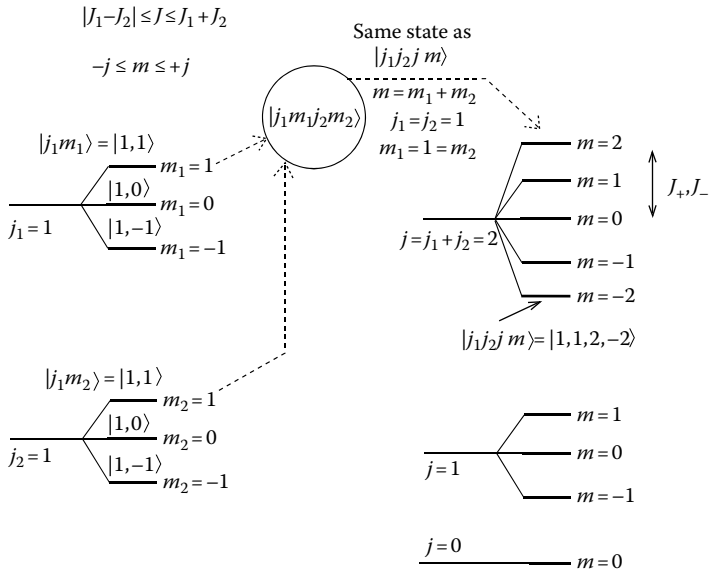


FIGURE 5.37 Nine product states on the left combine to form nine states on the right. Dotted lines indicate two left-hand states form the single right-hand state.

We now show how to find the Clebsch–Gordon coefficients (i.e., expansion coefficients) for the specific example of $j_1 = j_2 = 1$.

1. Draw a diagram similar to the left-hand side of Figure 5.37. Horizontal lines represent possible states. There are three possible values of m_1 ranging in $-j_1 \leq m_1 \leq +j_1$. Same for m_2 .
2. Conceptually take the direct product to provide nine product states of the form

$$|j_1 j_2 m_1 m_2\rangle$$

which is used later.

3. Draw the right-hand side of Figure 5.37 according to the following steps.
 - a. j can take values in the range $|j_1 - j_2| \leq j \leq j_1 + j_2$. The given example for $j_1 = j_2 = 1$ requires $j = 0, 1, 2$. On the other hand, if $j_1 = 1/2$ and $j_2 = 1$ then $j = 1/2, 1, 3/2$.
 - b. For each j , the values of m are $-j \leq m \leq j$. For example, when $j = 2$ then $m = 0, \pm 1, \pm 2$.
 - c. Determine the values of m for every j .
 - d. Note that the total number of states on the right-hand side of Figure 5.37 must be the same as the total number of direct product states on the left-hand side.
4. Evaluate the Clebsch–Gordon coefficients starting with the largest j .
 - a. For example, $j = 2$ is the largest j .
 - b. Look at the largest m for the largest j . For us, $m = 2$ is the largest.
 - c. Realize that a state on the left-hand side of Figure 5.37 is the same as one on the right-hand side, namely

$$|j_1 = 1, j_2 = 1, m_1 = 1, m_2 = 1\rangle = |j_1 = 1, j_2 = 1, j = 2, m = 2\rangle$$

$\leftarrow$ Left side of the figure $\rightarrow$
 $\leftarrow$ Right side of the figure $\rightarrow$

This works since j_1, j_2 retain their values on both sides of the figure. For the top lines (see the dotted lines in Figure 5.37) in the example $j = j_1 + j_2 = 2$ and $m = m_1 + m_2 = 2$.

d. Dropping j_1, j_2 for convenience,

$$|jm\rangle = \sum_{m_1 m_2} |m_1 m_2\rangle \langle m_1 m_2 | jm\rangle$$

The previous step just found all inner products $\langle m_1 m_2 | jm\rangle$ for the state $|j=2, m=2\rangle$. There is only one.

$$|j=2, m=2\rangle = |m_1=1, m_2=1\rangle \langle m_1=1, m_2=1 | j=2, m=2\rangle \quad (5.171)$$

However, step 4c shows that $|j=2, m=2\rangle$ and $|m_1=1, m_2=1\rangle$ are the same states. The Clebsch–Gordon coefficient must be $\langle m_1=1, m_2=1 | j=2, m=2\rangle = 1$.

e. Find the remaining $j=2$ states by applying the lowering operator to both sides of Equation 5.171. The lowering operator is

$$J^- = J_1^- + J_2^-$$

where J^- operates on $|j, m\rangle$, J_1^- operates on m_1 in $|m_1 m_2\rangle$, and J_2^- operates on the m_2 part of $|m_1 m_2\rangle$. Operating on the state given in step 4c yields

$$J^- |j=2, m=2\rangle = (J_1^- + J_2^-) |j_1 j_2 m_1 m_2\rangle \quad (5.172)$$

The previous section provides the equations for the effects of the raising and lowering operators. For notational convenience, we suppress the factors of $\hbar$.

$$J^+ |jm\rangle = [j(j+1) - m(m+1)]^{1/2} |j, m+1\rangle$$

$$J^- |jm\rangle = [j(j+1) - (m-1)m]^{1/2} |j, m-1\rangle$$

Similar equations hold for $J_1^\pm, J_2^\pm$. Equation 5.172 provides

$$\sqrt{2 \cdot 3 - 2} |j=2, m=1\rangle = \sqrt{2} |m_1=0\rangle |m_2=1\rangle + |m_1=1\rangle \sqrt{2} |m_2=0\rangle$$

So there are only two Clebsch–Gordon coefficients in

$$|j=2, m=1\rangle = \sum_{\substack{m_1 m_2 \\ \text{such that} \\ m=m_1+m_2}} |m_1 m_2\rangle \quad \leftarrow \text{Clebsch–Gordon coefficient} \rightarrow$$

namely

$$\langle m_1=0, m_2=1 | j=2, m=2\rangle = 1/\sqrt{2}$$

$$\langle m_1=1, m_2=0 | j=2, m=2\rangle = 1/\sqrt{2}$$

Notice that $\sqrt{2 \cdot 3 - 2} = 2$ was moved to the right-hand side to isolate $|j=2, m=1\rangle$ on the left-hand side so as to be able to identify the Clebsch–Gordon coefficient on the right-hand side.

Continue using J^- for all of the $J=2$ states.

5. Find the C–G coefficients for the next largest j : $j_{\text{largest}} - 1$. The procedure turns out to be the same for all of the other j . For this step, the largest $m=1$ state can come from two states

$$|j=1, m=1\rangle = c_1 |m_1=1, m_2=0\rangle + c_2 |m_1=0, m_2=1\rangle \quad (5.173)$$

where the m values add up as required by $m = m_1 + m_2$.

a. Find c_1 and c_2

a.1. Operate on Equation 5.173 with the “raising” operator $J^+ = J_1^+ + J_2^+$ to get

$$0 = J^+ |j = 1, m = 1\rangle = c_1 |m_1 = 1, m_2 = 1\rangle + c_2 |m_1 = 1, m_2 = 1\rangle \quad (5.174)$$

The zero occurs because the state $|j = 1, m = 2\rangle = J^+ |j = 1, m = 1\rangle$ does not exist. Therefore require

$$0 = c_1 + c_2 \quad (5.175)$$

a.2. Need another condition on c_1 and c_2

Using the results of the last substep $c_1 + c_2 = 0$, Equation 5.173 becomes

$$|j = 1, m = 1\rangle = c_1 |m_1 = 1, m_2 = 0\rangle - c_1 |m_1 = 0, m_2 = 1\rangle$$

Use the normalization of $|j = 1, m = 1\rangle$ to find

$$1 = \langle j = 1, m = 1 | j = 1, m = 1 \rangle = c_1^2 \langle 10 | 10 \rangle + c_1^2 \langle 01 | 01 \rangle$$

so that $c_1 = 1/\sqrt{2}$ ignoring any phase factor. Now we have

$$|j = 1, m = 1\rangle = \frac{1}{\sqrt{2}} |m_1 = 1, m_2 = 0\rangle - \frac{1}{\sqrt{2}} |m_1 = 0, m_2 = 1\rangle$$

There are two Clebsch–Gordon coefficients in this last equation (any others are zero).

b. Find the remaining Clebsch–Gordon coefficients for $j = 1$ by repeated application of

$$J^- = J_1^- + J_2^-$$

6. Continue the process for all j .

7. The final state is $|j_1 = 1, j_2 = 1, j = 0, m = 0\rangle = |j_1 = 1, j_2 = 1, m_1 = 0, m_2 = 0\rangle$.

Example 5.18

Find the basis vectors $|jm\rangle$ for two spin $1/2$ particles.

SOLUTION

First use $j = j_1 + j_2 = 1$. As before, we have the same state for the coupled $j = 1, m = 1$ and uncoupled $m_1 = 1/2, m_2 = 1/2$ basis vectors.

$$|j = 1, m = 1, j_1 = 1/2, j_2 = 1/2\rangle = |m_1 = 1/2, m_2 = 1/2, j_1 = 1/2, j_2 = 1/2\rangle$$

Apply the lowering operator to each side $\hat{J}_- = \hat{J}_-^{(1)} + \hat{J}_-^{(2)}$ where

$$\hat{J}_\pm |jm\rangle = \sqrt{j(j+1) - m(m\pm 1)} |jm\rangle$$

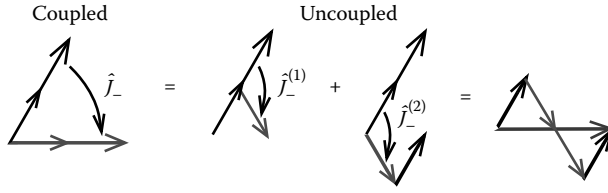


FIGURE 5.38 Effects of the various lowering operators.

with similar results for $\hat{J}_{\pm}^{(1)}, \hat{J}_{\pm}^{(2)}$. Again note the suppression of $\hbar^2$ for convenience. We find

$$\begin{aligned}\hat{J}_- |j=1, m=1, j_1=1/2, j_2=1/2\rangle &= \hat{J}_-^{(1)} |m_1=1/2, m_2=1/2, j_1=1/2, j_2=1/2\rangle \\ &\quad + \hat{J}_-^{(2)} |m_1=1/2, m_2=1/2, j_1=1/2, j_2=1/2\rangle \\ \sqrt{2} |j=1, m=0, j_1=1/2, j_2=1/2\rangle &= |m_1=-1/2, m_2=1/2, j_1=1/2, j_2=1/2\rangle \\ &\quad + |m_1=1/2, m_2=-1/2, j_1=1/2, j_2=1/2\rangle\end{aligned}$$

So

$$|j=1, m=0\rangle = \frac{1}{\sqrt{2}} |m_1=-1/2, m_2=1/2\rangle + \frac{1}{\sqrt{2}} |m_1=1/2, m_2=-1/2\rangle$$

Next use $j=|j_1-j_2|=0$ which requires $m=0$ and therefore

$$|j=0, m=0\rangle = c_1 |m_1=1/2, m_2=-1/2\rangle + c_2 |m_1=-1/2, m_2=1/2\rangle$$

This must be orthogonal to the previous results so use

$$|j=0, m=0\rangle = \frac{1}{\sqrt{2}} |m_1=-1/2, m_2=1/2\rangle - \frac{1}{\sqrt{2}} |m_1=1/2, m_2=-1/2\rangle$$

The cartoon in Figure 5.38 illustrates the effect of the lowering operators. The left portion shows that the coupled lowering operator rotates the entire initial state to give $m=0$. The middle portion shows how the lowering operator for each spin rotates only a portion to change the corresponding z -component from up to down. The right-hand side portion shows how to add the two uncoupled diagrams to get the same result as for the coupled case.

5.8 QUANTUM MECHANICAL REPRESENTATIONS

A representation (or picture) in quantum theory refers to the manner in which the theory models the time evolution of fundamental dynamical quantities. One might be interested in the time dependence of energy, momentum, position, or angular momentum for example. In classical mechanics, these dynamical variables evolve in time according to an equation of motion such as Newton's equation. The values of the dynamical variables describe the state of the classical system. Usually the word "dynamics" refers to the motion of an object (as related to impelling forces and supported by a mathematical framework). In quantum theory, the dynamical variables correspond to Hermitian operators. However, the quantum theory handles the time dependence in at least three ways. The Heisenberg picture assigns the time dependence to the Hermitian operators; this representation most closely mimics the classical approach. The Schrödinger picture assigns the time dependence to the wave functions; this representation most resembles that for classical wave motion. The interaction picture combines the best of both representations—the wave functions

move in Hilbert space only due to driving forces not included in the Hamiltonian. Because of the importance of the time evolution as related to dynamics, the quantum representations form one of the conceptual cornerstones for the theory.

This section discusses three representations used in quantum theory in the first section. The remainder of the section explores the mathematical descriptions of the particle as a result of using a particular representation.

5.8.1 DISCUSSION OF THE SCHRÖDINGER, HEISENBERG, AND INTERACTION REPRESENTATIONS

Quantum theory generally employs the Schrödinger, Heisenberg, and interaction representations. For the “Schrödinger representation,” the wave functions (i.e., vectors in Hilbert space) carry the dynamics of the particle or system (not the basis states though). The states depend on time according to

$$|\Psi(t)\rangle = \sum_E \beta_E(t) |\phi_E\rangle = \sum_E \beta_E(t) |E\rangle$$

Again, note that the basis states do not depend on time. The wave function resides in a Hilbert space defined by the basis vectors. This wave function moves in the space and its components therefore change with time. The wave functions in optics (i.e., the electric field) most closely resemble those for the quantum mechanical Schrödinger representation. In optics we know the energy density and power flow (etc.) once we know the motion of the electric field. As part of the definition of the Schrödinger picture, we require the operators (especially those corresponding to observables) to be explicitly independent of time. For example, in the coordinate-representation of the Schrödinger picture, we know the momentum to be given by

$$\hat{P} = \frac{\hbar}{i} \nabla$$

It does not depend on time. In fact, we have surmised this form of the momentum by working with the time-dependent wave function $e^{ikx - i\omega t}$ (a wave function in the Schrödinger picture, see Section 5.2). The top portion of Figure 5.39 attempts to describe the situation. We model the motion (i.e., dynamics) of a physical system (denoted “universe”) by the motion of the wave function in Hilbert space. The figure shows that the detection equipment—eyes in this case—do not change with time so that “the manner of making an observation” does not depend on time. The “act of observing” a particular quantity does not depend on when we make the observation. This point of view seems very natural since we assume that any change in a physical quantity must be due to changes in the physical systems and not our detection apparatus (eyes). The point of view is reminiscent of imagining oneself as an observer stationary in space and about whom the system moves—a very typical point of view (egocentric).

The “Heisenberg representation” assigns all of the time dependence to the operators and none to the wave functions. This representation resembles classical mechanics where the dynamical variables, such as momentum, depend on time. The wave functions in this representation do not depend on time. The wave functions in the Heisenberg representation consist of the superposition of the basis vectors of the form

$$|\Psi\rangle = \sum_E \beta_E |\phi_E\rangle = \sum_E \beta_E |E\rangle$$

where the β_E do not depend on time. In one sense, the wave functions (and especially the basis) form the lattice-work of a “stage” (as if for a theater) that defines the specific system that can be observed. The operators contain all the dynamics but they need to have the wave functions to give

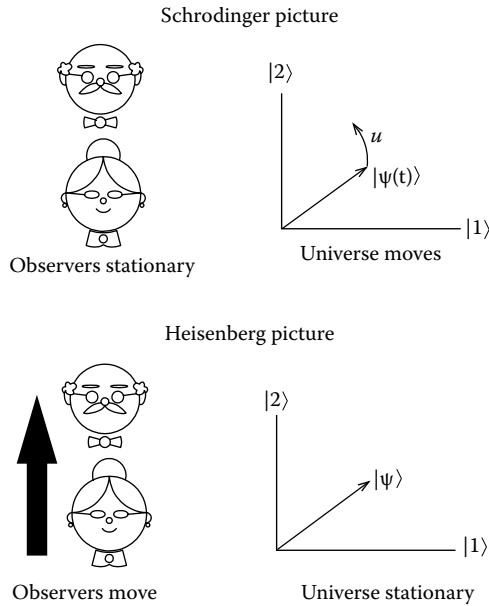


FIGURE 5.39 Cartoon representation of the Schrödinger and Heisenberg pictures.

information on the specific system. The wave functions essentially form the “lattice work of the stage.” The bottom portion of Figure 5.39 attempts to illustrate this paradigm by having the observers move rather than the system. However, the observations made in the two portions of the figure must agree.

This brings out another point for comparing and contrasting operators and vectors in the quantum theory. Regardless of the representation, an operator must contain all possible outcomes to an observation or operation. We can understand this point of view using the basis vector expansion of an operator found in Chapter 3 and Section 5.1. For example using the energy basis set, the Hamiltonian

$$\hat{H} = \sum_{\text{All } E} E|\phi_E\rangle \langle\phi_E| = \sum_{\text{All } E} E|E\rangle \langle E|$$

consists of all possible results of the observation because of the sum over all the energy eigenvalues E . However, the “wave functions” are written as a specific sum over the basis set; only a certain combination of basis vectors appears in the sum for the wave functions. For example, the wave function

$$|\Psi\rangle = \sum_{\text{Some } E} \beta_E |\phi_E\rangle = \sum_{\text{Some } E} \beta_E |E\rangle$$

contains information on only specific eigenvalues E . Even if it contains all eigenvalues, the sum refers to only one certain mixture (i.e., one vector in the Hilbert space) because of the specific values of β chosen. In summary, operators contain all possible results of a measurement while vectors represent specific instances of the system in question.

The “interaction representation” assigns some time dependence to the operators and some to the wave functions. We will find this representation especially suited for an “open” system. First, consider a “closed system” for which the number of particles and the total energy contained within the system remain constant. Basically one assumes that the Schrödinger’s equation has been solved

for this closed system. The time evolution of the system trivially involves only factors of the form $e^{-iEt/\hbar}$ as related to the evolution operator. We assign this trivial time dependence to the “operators.” With only the simple closed system present, the wave functions remain stationary in the vector space as defined by the time-independent basis set. Essentially this much corresponds to the Heisenberg representation. Now, if we include extra forces (above and beyond those included for the trivial solution) then any additional motion induced in the system appears in the wave function. For example, we might have a chunk of semiconductor material for which we can find the solution to Schrödinger’s equation for the holes and electrons. For the Heisenberg representation, we remove the time dependence from the wave function and assign it to the operators. The system consists of the chunk of material. Now second, consider the open system consisting of the semiconductor absorbing light. The original Hamiltonian for the closed system does not include this matter–light interaction so that the absorbed light will cause effects not taken into account by the original Hamiltonian. We assign this additional time dependence to the wave function. Of course we also work with the new Hamiltonian but it and all other operators are assigned the trivial time dependence. In this way, the wave functions move in Hilbert space only due to the additional forces not accounted by the original closed system.

5.8.2 SCHRÖDINGER REPRESENTATION

Previous sections show how the time-dependent wave function satisfies Schrödinger’s equation:

$$\hat{\mathcal{H}}|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle \quad (5.176)$$

The wave function moves in Hilbert space as shown in Figure 5.40. The components β depend on time but the basis vectors do not depend on time. The unitary evolution operator moves the initial wave function forward in time according to

$$\hat{u}(t, t_0) |\psi(t_0)\rangle = |\psi(t)\rangle \quad (5.177)$$

without changing the normalization of the function. The evolution operator actually depends on the difference in time and can be written as $\hat{u}(t, t_0) = \hat{u}(t - t_0)$.

For either open or closed systems, we define the evolution operator

$$\hat{\mathcal{H}} \hat{u}(t, t_0) |\psi(t_0)\rangle = i\hbar \frac{\partial}{\partial t} \hat{u}(t, t_0) |\psi(t_0)\rangle \quad \text{or} \quad \hat{\mathcal{H}} \hat{u}(t, t_0) = i\hbar \frac{\partial}{\partial t} \hat{u}(t, t_0) \quad (5.178)$$

by substituting Equation 5.177 into Equation 5.176. Equation 5.177 gives the initial condition of $\hat{u}(t_0, t_0) = 1$.

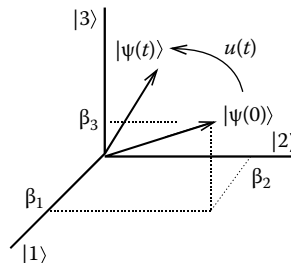


FIGURE 5.40 The wave function moves through Hilbert space in the Schrödinger picture.

Consider a closed system. For simplicity, set the initial time to zero $t_0 = 0$. Schrödinger's equation can be formally integrated when the Hamiltonian does not depend on time (i.e., a close system). Rearranging Equation 5.176 provides

$$\frac{\partial}{\partial t} |\psi(t)\rangle = \frac{\hat{\mathcal{H}}}{i\hbar} |\psi(t)\rangle$$

Consider the Hamiltonian operator to be similar to a constant and solve the simple differential equation to obtain

$$|\psi(t)\rangle = \exp\left(\frac{\hat{\mathcal{H}} t}{i\hbar}\right) |\psi(0)\rangle = \hat{u}(t) |\psi(0)\rangle \quad (5.179)$$

As discussed in Chapter 3, the operator $\hat{u}(t)$ is unitary (i.e., $\hat{u}^{-1} = \hat{u}^+$) since the Hamiltonian $\hat{\mathcal{H}}$ is Hermitian. For the energy basis set $\{\phi_n(x)\}$, the time dependence of the wave function must be

$$\begin{aligned} \Psi(x, t) &= \exp\left(\frac{\hat{\mathcal{H}} t}{i\hbar}\right) \Psi(x, 0) = \exp\left(\frac{\hat{\mathcal{H}} t}{i\hbar}\right) \sum_n \beta_n(0) \phi_n(x) \\ &= \sum_n \beta_n(0) \exp\left(\frac{\hat{\mathcal{H}} t}{i\hbar}\right) \phi_n(x) = \sum_n \beta_n(0) \exp\left(\frac{E_n t}{i\hbar}\right) \phi_n(x) \end{aligned}$$

The evolution operator will play a pivotal role for the Heisenberg representation. Of particular note is the phase factor $\exp(E_n t/i\hbar)$. This phase factor occurs whenever one finds the basis vectors for the entire systems (i.e., closed Hamiltonian) regardless of how complicated the system.

5.8.3 RATE OF CHANGE OF THE AVERAGE OF AN OPERATOR IN THE SCHRÖDINGER PICTURE

In this section, we discuss how an observed value (not the operator!) evolves in time for the Schrödinger representation. The next section on Ehrenfest's theorem then shows how Schrödinger's quantum mechanics reproduces results for classical mechanics. We expect the classical analog of a quantum mechanical system to involve an average over the quantum mechanical microscopic quantities. We expect to recover Newton's second law by calculating the rate of change of the expectation value of the quantum mechanical momentum operator. We therefore start the discussion by considering the rate of change of the expectation value of an operator using the Schrödinger picture.

Let $\hat{A} = \hat{A}(\vec{r}, t)$ be an operator in the Schrödinger picture where usually the operator does not explicitly depend on time. Suppose further that the wave vector $|\psi(t)\rangle$ is a solution to Schrödinger's equation. The time rate of change of the expectation value of the operator can be calculated

$$\frac{d}{dt} \langle \hat{A} \rangle = \frac{d}{dt} \langle \psi | \hat{A} | \psi \rangle = \left\langle \frac{\partial \psi}{\partial t} \right| \hat{A} | \psi \rangle + \langle \psi | \frac{\partial \hat{A}}{\partial t} | \psi \rangle + \langle \psi | \hat{A} \left| \frac{\partial \psi}{\partial t} \right\rangle$$

The derivative moves into the bra vector since the bra vector symbolizes an integral with respect to spatial coordinates. Now use Schrödinger's equation for the time derivatives of the wave functions to obtain

$$\frac{d}{dt} \langle \hat{A} \rangle = \left\langle \frac{\hat{\mathcal{H}}}{i\hbar} \psi \right| \hat{A} | \psi \rangle + \langle \psi | \frac{\partial \hat{A}}{\partial t} | \psi \rangle + \langle \psi | \hat{A} \left| \frac{\hat{\mathcal{H}}}{i\hbar} \psi \right\rangle$$

Evaluating the left-most inner product by using the definition of the adjoint

$$\left\langle \frac{\hat{\mathcal{H}}}{i\hbar} \psi \right| = \left[\left| \frac{\hat{\mathcal{H}}}{i\hbar} \psi \right\rangle \right]^+ = \left[\frac{\hat{\mathcal{H}}}{i\hbar} |\psi\rangle \right]^+ = \langle \psi | \frac{\hat{\mathcal{H}}^+}{(-i)\hbar} = \langle \psi | \frac{\hat{\mathcal{H}}}{(-i)\hbar}$$

The rate of change of the expectation value of the operator can now be rewritten as

$$\frac{d}{dt} \langle \hat{A} \rangle = \langle \psi | \frac{\hat{\mathcal{H}}}{(-i)\hbar} \hat{A} | \psi \rangle + \langle \psi | \frac{\partial \hat{A}}{\partial t} | \psi \rangle + \langle \psi | \hat{A} \frac{\hat{\mathcal{H}}}{i\hbar} | \psi \rangle$$

Collecting terms provides

$$\frac{d}{dt} \langle \hat{A} \rangle = \frac{i}{\hbar} \langle [\hat{\mathcal{H}}, \hat{A}] \rangle + \left\langle \frac{\partial \hat{A}}{\partial t} \right\rangle \quad (5.180)$$

Usually the expectation value of the time derivatives of the operator (last term) is zero for the Schrödinger picture.

Example 5.19

For the infinitely deep potential well, calculate the rate of change of the momentum for an electron.

SOLUTION

The operator in Equation 5.180 becomes $\hat{A} = \hat{p}$. The Hamiltonian is given by $\hat{\mathcal{H}} = \hat{p}^2/2m$. It is easy to calculate that $[\hat{\mathcal{H}}, \hat{p}] = 0$. We assume $\partial \hat{p}/\partial t = 0$ as usual for the Schrödinger representation. Therefore, the rate of change of the expected value of momentum must be $d\langle \hat{p} \rangle/dt = 0$.

5.8.4 EHRENFEST'S THEOREM FOR THE SCHRÖDINGER REPRESENTATION

Now we discuss Ehrenfest's theorem showing that Schrödinger's quantum mechanics leads to Newton's second law. We first show that because a quantum particle can be considered as smeared-out over a volume of space (at least in the sense of statistics), the classical dynamical variable corresponds to the quantum mechanical average of the operator.

Consider an example for the force exerted on a body to see why quantum mechanics averages an operator in the Schrödinger representation. Figure 5.41 shows the probability density for the location of a quantum mechanical particle. We might imagine the particle of mass m as “smeared out” over the region. Suppose $\vec{f}$ represents the force per unit mass. The total classical force must be $\vec{F} = \sum_i \vec{f}_i \Delta m_i$ where the mass $m = \sum_i \Delta m_i$ might not be uniformly distributed across the region of space. The figure shows more mass near the center and less at the “boundaries.” The amount of

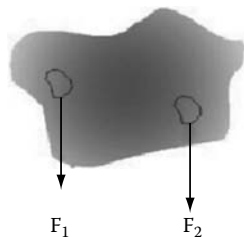


FIGURE 5.41 A quantum mechanical object described by a wave function.

mass in a given region must be proportional to the probability density p_r of finding the mass in a small region Δx . The small mass could be an electron in the present case. For the 1-D case, we write $\Delta m_i \sim p_r dx \sim \psi^* \psi \Delta x$. We can therefore write the total force as

$$\vec{F} = \sum_i \vec{F}_i \Delta m_i \sim \sum_i \psi^*(x_i) \vec{F}_i(x_i) \psi(x_i) \Delta x \rightarrow \int \psi^*(x) \vec{F}(x) \psi(x) dx = \langle \mathcal{F} \rangle$$

Therefore, because the quantum mechanical particle effectively occupies a large volume of space, classical quantities like force and interaction energy do not occur at one specific point; instead they occur over the region of space. We expect the quantum mechanical operator to be averaged over a region of space to produce the corresponding classical quantity. Furthermore, this shows that the time-dependence of the wave function translates to a time dependence of the classical quantity through the averaging procedure.

We now show “Ehrenfest’s theorem,” which relates the classical force to the rate of change of the expected value of momentum for a single particle

$$\vec{F}_{\text{class}} = \frac{d\langle \hat{p} \rangle}{dt} \quad \text{with} \quad \frac{\partial \hat{p}}{\partial t} = 0$$

for the Schrödinger picture. The time rate of change of the expected value of an operator is obtained from Equation 5.180.

$$\frac{d\langle \hat{p} \rangle}{dt} = \frac{i}{\hbar} \langle [\hat{\mathcal{H}}, \hat{p}] \rangle = \frac{i}{\hbar} \left\langle \left[\frac{\hat{p}^2}{2m} + V(\vec{r}), \hat{p} \right] \right\rangle = \frac{i}{\hbar} \left\langle \left[\frac{\hat{p}^2}{2m}, \hat{p} \right] \right\rangle + \frac{i}{\hbar} \langle [V(\vec{r}), \hat{p}] \rangle$$

where we have used the commutator identity $[\hat{A} + \hat{B}, \hat{C}] = [\hat{A}, \hat{C}] + [\hat{B}, \hat{C}]$. Then since $[\hat{p}^2/2m, \hat{p}] = 0$ we must have $\langle [\hat{p}^2/2m, \hat{p}] \rangle = 0$. Finally, we need to evaluate the commutator between the potential energy and the momentum.

$$[V(\vec{r}), \hat{p}] f = \left[V(\vec{r}), \frac{\hbar}{i} \nabla \right] f = V \frac{\hbar}{i} \nabla f - \frac{\hbar}{i} \nabla (Vf) = V \frac{\hbar}{i} \nabla f - \frac{\hbar}{i} V \nabla f - \left(\frac{\hbar}{i} \nabla V \right) f = i\hbar (\nabla V) f$$

where we use an arbitrary function f because the commutator is an operator. As a result, we can conclude the operator relation $[V(\vec{r}), \hat{p}] = i\hbar \nabla V$. Putting all the steps together we arrive at Ehrenfest’s theorem:

$$\frac{d\langle \hat{p} \rangle}{dt} = \frac{i}{\hbar} \langle [\hat{\mathcal{H}}, \hat{p}] \rangle = \frac{i}{\hbar} \langle i\hbar \nabla V \rangle = \langle -\nabla V \rangle = \langle F \rangle$$

By way of illustration, consider Figure 5.42 showing a wave packet traveling to the right with speed v . The wave function clearly depends on time because it moves. The expectation value of the position operator $\hat{x}$ gives the position of the center of the wave packet. Now because the wave packet moves, the expectation value of the position operator must depend on time $\langle \hat{x} \rangle = \bar{x}(t)$. We therefore find that the average of an operator depends on time even though the operator itself remains independent of time.

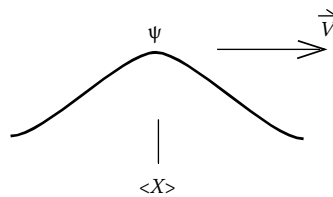


FIGURE 5.42 If a wave function depends on time then averages using that wave function must also depend on time.

5.8.5 HEISENBERG REPRESENTATION

The Heisenberg representation assigns the dynamics to the operators. None of the wave functions depend on time so that none of the dynamics appears in the wave functions. We can find the time-dependent operators from those in the Schrödinger picture. The simplest procedure requires all expectation values to be invariant across representations; that is the averages have the same values when computed in any of the representations. The observed classical value (represented by the quantum average) should not depend on the particular quantum picture used to calculate the average values. Suppose we represent the state of the system by the ket $|\psi_s(t)\rangle$ in the Schrödinger picture (where “s” denotes Schrödinger). The expectation value of an operator $\hat{O}_s$ can be written as

$$\langle \psi_s(t) | \hat{O}_s | \psi_s(t) \rangle = \langle \psi_h | \hat{u}^\dagger \hat{O}_s \hat{u} | \psi_h \rangle \quad (5.181)$$

where $\hat{u}$ represents a unitary operator. For convenience, we set the origin of time to $t=0$ rather than an arbitrary time t_0 . That is, we define the Heisenberg wave function to be $|\psi_h\rangle = |\psi_s(0)\rangle$. Therefore, in order for the expectation value to be independent of picture, we define the time-dependent Heisenberg operator to be

$$\hat{O}_h(t) = \hat{u}^\dagger \hat{O}_s \hat{u} \quad (5.182)$$

We found the unitary evolution operator $\hat{u}$ for “closed” systems in the Schrödinger picture; a perturbation approach can be used for open systems. Essentially, a-closed system is one for which energy and mass does not enter or leave the system such as through light or heat. The total energy remains constant. With the Hamiltonian thought of as constant, Section 5.1 showed the evolution operator had the following form.

$$\hat{u}(t) = \exp\left(\frac{\hat{\mathcal{H}}_s t}{i\hbar}\right) \quad (5.183)$$

where $\hat{\mathcal{H}}_s$ denotes the Schrödinger Hamiltonian. An open system would involve integral over time (see the following section on time dependent perturbation theory) since energy enters or leaves the system and one does not imagine the Hamiltonian as constant. We do not need to subscript the Hamiltonian with an “s” in this case because, as will be seen in the second example below, it has the same form in either the Schrödinger or Heisenberg representation. We can show that commutator expressions in the Schrödinger picture produce similar results in the Heisenberg picture (refer to the chapter exercises).

Example 5.20

Find the Heisenberg representation of the momentum operator $\hat{p}$ for the infinitely deep square well without an external interaction.

SOLUTION

The Heisenberg momentum operator must be given by

$$\hat{p}_h = \hat{u}^\dagger \hat{p} \hat{u} = \exp\left(-\frac{\hat{\mathcal{H}}}{i\hbar} t\right) \hat{p} \exp\left(\frac{\hat{\mathcal{H}}}{i\hbar} t\right)$$

Inside the well, the Schrödinger Hamiltonian has the form $\hat{\mathcal{H}} = \hat{p}^2/2m$. Now, since the momentum operator commutes with the Hamiltonian $[\hat{p}, \hat{\mathcal{H}}] = [\hat{p}, \hat{p}^2/2m] = 0$ then any function of the Hamiltonian must also commute with momentum

$$\left[\hat{p}, \exp\left(\frac{\hat{\mathcal{H}} t}{i\hbar}\right) \right] = 0$$

as can be easily verified by Taylor expanding the exponential. Therefore, the Heisenberg representation of the momentum operator can be written as

$$\hat{p}_h = \hat{u}^+ \hat{p} \hat{u} = \exp\left(-\frac{\hat{\mathcal{H}}}{i\hbar} t\right) \exp\left(\frac{\hat{\mathcal{H}}}{i\hbar} t\right) \hat{p} = \hat{p}$$

In the simple case of an infinitely deep well, we see that the Heisenberg and Schrödinger representations are the same for the momentum operator. Especially notice that the unitary operator $\hat{u}$ is written in terms of Schrödinger quantities.

Example 5.21

What is the Heisenberg representation of the Schrödinger Hamiltonian without an external interaction?

SOLUTION

The Schrödinger and Heisenberg representations have identical Hamiltonians since $[\hat{u}, \hat{\mathcal{H}}_s] = 0$

$$\hat{\mathcal{H}}_h = \hat{u}^+ \hat{\mathcal{H}}_s \hat{u} = \exp\left(-\frac{\hat{\mathcal{H}}_s}{i\hbar} t\right) \hat{\mathcal{H}}_s \exp\left(\frac{\hat{\mathcal{H}}_s}{i\hbar} t\right) = \hat{\mathcal{H}}_s$$

5.8.6 HEISENBERG EQUATION

Next, we show the principal method of calculating the time evolution of the Heisenberg operators. As demonstrated in the present section, the dynamics of the Heisenberg operators can be found using the Heisenberg equation given by

$$\frac{d\hat{O}_h}{dt} = \frac{i}{\hbar} [\hat{\mathcal{H}}_h, \hat{O}_h] + \left(\frac{\partial}{\partial t} \hat{O}_s\right)_h \quad (5.184)$$

Often the last term is defined as $(\partial\hat{O}_s/\partial t)_h \equiv \partial\hat{O}_h/\partial t$. The Hamiltonian generates displacements in time. The commutator for the operators takes the place of the Schrödinger equation for the wave functions. This last equation has a form somewhat similar to that for the Schrödinger picture in Equation 5.180. For the Heisenberg representation, we do not need to calculate an expectation value. We will see how the operators in the Heisenberg representation obey equations of motion very similar to the dynamical variables in classical mechanics.

Equation 5.184 holds for either an open or closed system as we now show. Starting with Equation 5.182, $\hat{O}_h(t) = \hat{u}^+ \hat{O}_s \hat{u}$, we find

$$\frac{d\hat{O}_h}{dt} = \frac{d}{dt} \hat{u}^+ \hat{O}_s \hat{u} = \left(\frac{d}{dt} \hat{u}^+\right) \hat{O}_s \hat{u} + \hat{u}^+ \left(\frac{d\hat{O}_s}{dt}\right) \hat{u} + \hat{u}^+ \hat{O}_s \left(\frac{d}{dt} \hat{u}\right) \quad (5.185)$$

Equation 5.178, $\hat{\mathcal{H}}_s \hat{u}(t) = i\hbar \partial_t \hat{u}(t)$ for $t_0 = 0$, provides $\partial_t \hat{u}^+ = i\hat{u}^+ \hat{\mathcal{H}}_s / \hbar$ by taking the adjoint of both sides. Therefore, Equation 5.185 becomes

$$\frac{d\hat{O}_h}{dt} = \left(\frac{i}{\hbar} \hat{u}^+ \hat{\mathcal{H}}_s\right) \hat{O}_s \hat{u} + \hat{u}^+ \left(\frac{d\hat{O}_s}{dt}\right) \hat{u} + \hat{u}^+ \hat{O}_s \left(-\frac{i}{\hbar} \hat{\mathcal{H}}_s \hat{u}\right)$$

Finally, substituting $\hat{u}\hat{u}^+ = 1$ between the Hamiltonian and the operator $\hat{O}$ provides

$$\frac{d\hat{O}_h}{dt} = \left(\frac{i}{\hbar}\hat{u}^+\hat{\mathcal{H}}_s\hat{u}\right)(\hat{u}^+\hat{O}_s\hat{u}) + \hat{u}^+\left(\frac{d\hat{O}_s}{dt}\right)\hat{u} + \hat{u}^+\hat{O}_s\hat{u}\left(-\frac{i}{\hbar}\hat{u}^+\hat{\mathcal{H}}_s\hat{u}\right) = \frac{i}{\hbar}[\hat{\mathcal{H}}_h, \hat{O}_h] + \left(\frac{d\hat{O}_s}{dt}\right)_h$$

as required.

Example 5.22

Show Equation 5.184 for the closed system using Equation 5.183 where $\hat{\mathcal{H}}_s = \hat{\mathcal{H}} = \hat{\mathcal{H}}_h$.

$$\begin{aligned}\frac{d\hat{O}_h}{dt} &= \frac{d}{dt}(\hat{u}^+\hat{O}_s\hat{u}) = \frac{d}{dt}\left\{\exp\left(-\frac{\hat{\mathcal{H}}}{i\hbar}t\right)\hat{O}_s\exp\left(+\frac{\hat{\mathcal{H}}}{i\hbar}t\right)\right\} \\ &= \left\{-\frac{\hat{\mathcal{H}}}{i\hbar}\exp\left(-\frac{\hat{\mathcal{H}}}{i\hbar}t\right)\right\}\hat{O}_s\exp\left(+\frac{\hat{\mathcal{H}}}{i\hbar}t\right) + \exp\left(-\frac{\hat{\mathcal{H}}}{i\hbar}t\right)\left\{\frac{\partial}{\partial t}\hat{O}_s\right\}\exp\left(+\frac{\hat{\mathcal{H}}}{i\hbar}t\right) \\ &\quad + \exp\left(-\frac{\hat{\mathcal{H}}}{i\hbar}t\right)\hat{O}_s\left\{\frac{\hat{\mathcal{H}}}{i\hbar}\exp\left(+\frac{\hat{\mathcal{H}}}{i\hbar}t\right)\right\}\end{aligned}$$

Using the definition of a Heisenberg operator (Equation 5.185) and combining terms produces

$$\frac{d\hat{O}_h}{dt} = -\frac{\hat{\mathcal{H}}}{i\hbar}\hat{O}_h + \hat{O}_h\frac{\hat{\mathcal{H}}}{i\hbar} + \left(\frac{\partial}{\partial t}\hat{O}_s\right)_h$$

where the time derivative of the Schrödinger operator is usually 0. Forming the commutator provides the required results in Equation 5.184.

5.8.7 NEWTON'S SECOND LAW FROM THE HEISENBERG REPRESENTATION

We can easily recover Newton second law of motion from the Heisenberg representation starting with the 1-D Schrödinger Hamiltonian, for example,

$$\hat{\mathcal{H}}_s = \frac{\hat{p}^2}{2m} + V(x)$$

This Hamiltonian represents a closed system so that the demonstration can follow either of two routes. We use the general definition of the evolution operator in Equations 5.177 and 5.178, and leave the corresponding demonstration for the closed evolution operator to the chapter exercises. Let $\hat{p}_h = \hat{p}_h(t)$ be the Heisenberg momentum operator. We wish to calculate its rate of change using Equation 5.184.

$$\frac{d\hat{p}_h}{dt} = \frac{i}{\hbar}[\hat{\mathcal{H}}_h, \hat{p}_h] = \frac{i}{\hbar}[\hat{u}^+\hat{\mathcal{H}}_s\hat{u}, \hat{u}^+\hat{p}\hat{u}] = \frac{i}{\hbar}\hat{u}^+[\hat{\mathcal{H}}_s, \hat{p}]\hat{u} \quad (5.186)$$

since $\hat{u}^+\hat{u} = 1 = \hat{u}\hat{u}^+$. Substituting for the Hamiltonian we find

$$\begin{aligned}\frac{d\hat{p}_h}{dt} &= \frac{i}{\hbar}\hat{u}^+\left[\frac{\hat{p}^2}{2m} + V(x), \hat{p}\right]\hat{u} = \frac{i}{\hbar}\hat{u}^+[V(x), \hat{p}]\hat{u} = \frac{i}{\hbar}\hat{u}^+\left[V(x), \frac{\hbar}{i}\frac{\partial}{\partial x}\right]\hat{u} \\ &= \hat{u}^+\left(-\frac{\partial V}{\partial x}\right)\hat{u} = \hat{u}^+F\hat{u} = F_h\end{aligned}$$

This last result is Newton's second law! We see that the Heisenberg operators most naturally take the place of the classical dynamical variables.

5.8.8 INTERACTION REPRESENTATION

As previously mentioned, the interaction representation combines portions of the Schrödinger and Heisenberg representations. Both the operators and wave functions depend on time. We identify the dynamics embedded in the wave function as due to the interaction between the system and an external agent. Therefore, the wave functions move in Hilbert space in response to the “extra” potentials V imposed on the system. The operators carry the dynamics of the closed system.

Suppose the Hamiltonian for the system has the form

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{V} \quad (5.187)$$

where the closed-system Hamiltonian $\hat{\mathcal{H}}_0$ must be independent of time. Consider Schrödinger’s equation in operator form

$$\hat{\mathcal{H}}|\Psi_s(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi_s(t)\rangle \quad \text{or} \quad (\hat{\mathcal{H}}_0 + \hat{V})|\Psi_s(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi_s(t)\rangle \quad (5.188)$$

Define the interaction wave function through the relation

$$|\Psi_s(t)\rangle = \hat{u} |\Psi_I(t)\rangle \quad (5.189)$$

using the unitary evolution operator previously defined for the closed system

$$\hat{u}(t) = \exp\left(\frac{\hat{\mathcal{H}}_0}{i\hbar} t\right) \quad (5.190)$$

The subscripts “s” and “I” stand for Schrödinger and Interaction, respectively. The inverse unitary operator $\hat{u}^+$ essentially removes the time dependence from the wave function $|\Psi_s(t)\rangle$ attributable to the Hamiltonian $\hat{\mathcal{H}}_0$. However, the wave function retains some time dependence due to the added potential $\hat{V}$ occurring in the full Hamiltonian $\hat{\mathcal{H}}$.

We can write the Schrödinger equation using the interaction representation. Substituting Equation 5.189 into Equation 5.188 produces

$$(\hat{\mathcal{H}}_0 + \hat{V})\hat{u}(t)|\Psi_I(t)\rangle = i\hbar \frac{\partial}{\partial t} \hat{u}(t)|\Psi_I(t)\rangle \quad (5.191)$$

Now, differentiate both terms on the right-hand side of Equation 5.191

$$\begin{aligned} (\hat{\mathcal{H}}_0 + \hat{V})\hat{u}(t)|\Psi_I(t)\rangle &= i\hbar \frac{\partial}{\partial t} \exp\left(\frac{\hat{\mathcal{H}}_0}{i\hbar} t\right) |\Psi_I(t)\rangle \\ &= i\hbar \left\{ \frac{\hat{\mathcal{H}}_0}{i\hbar} \exp\left(\frac{\hat{\mathcal{H}}_0}{i\hbar} t\right) |\Psi_I(t)\rangle + \exp\left(\frac{\hat{\mathcal{H}}_0}{i\hbar} t\right) \frac{\partial}{\partial t} |\Psi_I(t)\rangle \right\} \\ &= \hat{\mathcal{H}}_0 \hat{u} |\Psi_I(t)\rangle + i\hbar \hat{u} \frac{\partial}{\partial t} |\Psi_I(t)\rangle \end{aligned}$$

Canceling the terms involving $\hat{\mathcal{H}}_0$ from both sides produces

$$\hat{V} \hat{u}(t) |\Psi_I(t)\rangle = i\hbar \hat{u} \frac{\partial}{\partial t} |\Psi_I(t)\rangle$$

Operating on both sides with the adjoint of the evolution operator and defining the interaction potential as $\hat{V}_I = \hat{u}^\dagger \hat{V} \hat{u}$ yields

$$\hat{u}^\dagger \hat{V} \hat{u} |\Psi_I(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi_I(t)\rangle \quad \text{or} \quad \hat{V}_I |\Psi_I(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi_I(t)\rangle \quad (5.192)$$

As a result, the wave function satisfies a Schrödinger-like equation with the interaction potential $\hat{V}_I$ in the interaction representation taking the place of the Hamiltonian. The last equation makes it clear that, for the interaction representation, the wave-function carries the dynamics induced by the added potential V .

The section on time-dependent perturbation theory will demonstrate a unitary evolution operator $\hat{U}(t)$ that moves the interaction wave function through Hilbert space according to $|\Psi_I(t)\rangle = \hat{U} |\Psi_I(0)\rangle$. The operator $\hat{U}(t)$ should not be confused with the operator $\hat{u}$ that changes the Schrödinger wave function into the interaction one $|\psi_s\rangle = \hat{u} |\psi_h\rangle$. The operator $\hat{U}(t)$ has the form $\hat{U} = \hat{T} e^{\frac{i}{\hbar} \int_{t_0}^t \hat{V}_I(t_1) dt_1}$ for an interaction that starts at t_0 . The operator $\hat{T}$ denotes the time-ordered product.

5.9 TIME-INDEPENDENT PERTURBATION THEORY

Perturbation theory provides a technique for finding approximate solutions to a partial differential equation. Exact wave functions and energies can be conveniently found for a limited number of potentials including the constant potentials (and step potentials that produce plane waves), linear potentials (Airy functions), and the quadratic potentials (Hermite polynomials with multiplicative exponentials). These provide starting points to find the approximate wave functions and energies for other systems with similar potentials. The perturbation technique assumes that two systems with similar potentials must have similar energy eigenfunctions and eigenvalues. Knowing the eigenfunctions and eigenvalues for one system makes it possible to deduce the eigenfunctions and eigenvalues for the other by using matrix elements of the perturbing potential.

5.9.1 INITIAL DISCUSSION OF PERTURBATIONS

The meaning of “perturbation” can be ascertained by considering a quantum system, such as a particle in a well, for which we can find an exact “solution” to Schrödinger’s wave equation. Now suppose that we make some slight change in the system (i.e., we perturb the system). For example, we might “slightly” change the depth or shape of the well. We already know the energy basis functions and eigenvalues for the original unperturbed case. It seems reasonable that the original solution should not be much different than the solution for the perturbed one so long as the changes remain small. Perturbation theory mimics this physical intuition by first exactly solving Schrödinger’s equation for the unperturbed system and then finding new basis vectors for the perturbed system by expressing these new basis vectors in terms of the original basis set.

Generally quantum mechanics uses both time-independent and time-dependent perturbation theories. Both techniques find approximate solutions to the Schrödinger equation. The time-independent perturbation theory applies to Schrödinger equations that have time-independent potential energy terms. For example, maybe we change the simple square well into a more complicated well having a nonconstant potential at the bottom. However, for the potential energy to be independent of time, “changing the potential” really means that we have a system with a potential energy function that differs in small ways compared with a simpler system for which we exactly know the energy eigenfunctions and eigenvalues. We really do not “change” the simple system in the sense of a time-dependent process.

As previously discussed, the time-independent Schrödinger wave equation (SWE) represents a Sturm-Liouville system that consists of boundary conditions and an eigenfunction equation to determine the energy basis set and allowed energy levels. Changing the potential energy term in

Schrödinger's equation necessarily changes the Sturm-Liouville system and also therefore, the basis functions and energy eigenvalues. The perturbation theory allows us to find the new basis states and allowed energy values for the new Sturm-Liouville system based on the original basis states and energy values from the original Sturm-Liouville system. Of course we know that the basis states and energy values are important when measuring the energy since a superposed wave function collapses to one of the basis states and the result of the measurement must be one of the allowed energy values.

Time-dependent perturbation theory applies to systems where the potential energy explicitly depends on time. As an example, suppose we apply an RF field (i.e., light wave) to an electron in an atom or well. With time-dependent perturbation theory, we assume the perturbation does not affect the basis vectors (on average) and so we do not need to calculate new basis vectors. Instead, we assume that the particle makes a transition from one basis state to another, either by absorbing a quantum of energy or by emitting one. The reader will recognize this as the process of optical emission and absorption such as for lasers and light emitting diodes.

5.9.2 NONDEGENERATE PERTURBATION THEORY

Time-independent perturbation theory can be used to find new energy basis vectors $|v_m\rangle$ from the old basis $|u_m\rangle$ when the Hamiltonian has been altered by a small potential

$$\hat{\mathcal{H}}_0 \rightarrow \hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{V}(x)$$

That is, for a single-particle system, the altered Hamiltonian will have the form $\hat{\mathcal{H}} = \frac{\hat{p}^2}{2m} + \hat{V}_0 + \hat{V}$ so that $\hat{\mathcal{H}}_0 = \frac{\hat{p}^2}{2m} + \hat{V}_0$. For perturbation theory, we normally separate out the “easy to solve” terms in the Hamiltonian $\hat{\mathcal{H}}$ described by Schrödinger's equation

$$\hat{\mathcal{H}}_0|\Psi\rangle = i\hbar \frac{\partial}{\partial t}|\Psi\rangle$$

Using separation of variables, we find a Sturm-Liouville problem (i.e., the time-independent Schrödinger equation)

$$\hat{\mathcal{H}}_0|u_m\rangle = E_m|u_m\rangle$$

for the basis of vectors $\{|u_m\rangle\}$ and energy eigenvalues E_m where the subscript m is not to be confused with the mass of the particle. In this section, we assume nondegenerate eigenvalues (i.e., for each eigenvalue there exists only one eigenvector). Question: How do the eigenvectors and eigenvalues alter when the Hamiltonian slightly alters?

Example 5.23

Consider the infinitely deep well shown in Figure 5.43 with the potential energy $V=0$ at the bottom of the well. Earlier in the chapter, the energy eigenfunctions were shown to be

$$\langle x|E_m\rangle = u_m(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{m\pi x}{L}\right)$$

where E_m are the energy eigenvalues. Now suppose that we construct a new system identical to the first except the bottom of the well slopes downward to the right so that

$$V(x) = -ax \quad a > 0$$

The electron tends to migrate to the regions of lowest potential; the electron has the greatest chance of being found on the right-hand side of the well. Although the problem can be exactly “solved” using Airy functions, we will solve this problem using the perturbation theory in later examples.

Now let us state the problem for the new system. The new time-independent Schrödinger equation has the form

$$\hat{\mathcal{H}}|v_m\rangle = W_m|v_m\rangle$$

where $\{|v_m\rangle\}$ is the energy basis set for the modified system with Hamiltonian

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}$$

where W_m are the modified energy eigenvalues, and $\hat{\mathcal{V}}$ represents the “small” additional potential. We expect the modified basis set to be very similar to the unmodified basis set. For example, Figure 5.43 shows that small modifications of the well produce new modes that approximate sinusoids. The new basis vectors reside in the same Hilbert space as the old ones. We can actually find a unitary transformation that changes one basis set into another as will be seen in the next section. Figure 5.44 shows that the new basis vector $|v_m\rangle$ must be related to the old $|u_m\rangle$ by a small “difference” vector

$$\delta|u_m\rangle = |v_m\rangle - |u_m\rangle$$

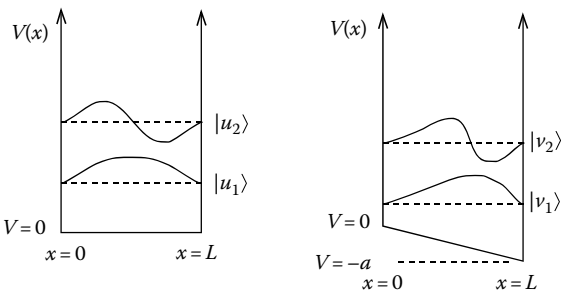


FIGURE 5.43 Applying the ramp perturbs the infinitely deep well.

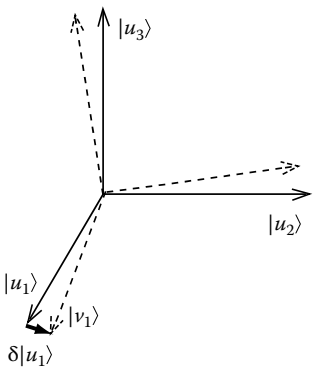


FIGURE 5.44 The new basis vector is related to the old by a rotation.

The small difference vector $\delta|u_m\rangle$ can also be written as a “sum over the old basis vectors” since the new basis vectors $|v_m\rangle$ can be written as a linear combination of the old ones $|u_m\rangle$.

Although a number of derivations for the time-independent perturbation formulas are possible, we follow the more traditional approach that explicitly keeps track of the “order” of approximation by using a parameter λ (not to be confused with a wavelength). The new (i.e., modified) Hamiltonian can be written as

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \lambda \hat{V}$$

where $\lambda=0$ produces the old (i.e., unmodified) Hamiltonian and $\lambda=1$ provides the new Hamiltonian. At the end of the procedure, we will set the parameter λ equal to 1.

Similar to a Taylor expansion, one can write the new basis vectors as

$$|v_m\rangle \cong |v_m^{(0)}\rangle + \lambda^1 |v_m^{(1)}\rangle + \lambda^2 |v_m^{(2)}\rangle + \cdots = \sum_{a=0}^{\infty} \lambda^a |v_m^{(a)}\rangle \quad (5.193a)$$

Note the use of a to indicate the order of the approximation. To see the similarity with a Taylor expansion, consider the function $v(x, \lambda)$ and expand in the parameter λ to find

$$v_m(x, \lambda) = v_m(x, 0) + \frac{1}{1!} \frac{\partial v_m(x, \lambda)}{\partial \lambda} \bigg|_{\lambda=0} \lambda^1 + \frac{1}{2!} \frac{\partial^2 v_m(x, \lambda)}{\partial \lambda^2} \bigg|_{\lambda=0} \lambda^2 + \cdots \quad (5.193b)$$

The correction terms in Equation 5.193a then take the form

$$\langle x | v_m^{(a)} \rangle = \frac{1}{a!} \frac{\partial^a}{\partial \lambda^a} v_m(x, \lambda = 0)$$

We will not need this last explicit form for the approximation. The powers of λ indicate the “order of the approximation” similar to the numbers a appearing as superscripts in the kets. The reader should realize that Equation 5.193a is “not” an orthonormal expansion! However, each little piece of the new basis vector, namely each $|v_m^{(a)}\rangle$, can be written as an orthonormal expansion in the original basis set $|u_m\rangle$. We further assume that the new energy eigenvalues can be written as an approximation

$$W_m = \sum_a \lambda^a W_m^{(a)} \quad (5.194)$$

The Schrödinger equation for the new system is

$$\hat{\mathcal{H}}|v_m\rangle = (\hat{\mathcal{H}}_0 + \lambda \hat{V})|v_m\rangle = W_m|v_m\rangle$$

or substituting the approximations for the new basis vectors and eigenvalues

$$(\hat{\mathcal{H}}_0 + \lambda \hat{V})(|v_m^{(0)}\rangle + \lambda^1 |v_m^{(1)}\rangle + \cdots) = (W_m^{(0)} + \lambda W_m^{(1)} + \cdots)(|v_m^{(0)}\rangle + \lambda^1 |v_m^{(1)}\rangle + \cdots) \quad (5.195)$$

Multiply the terms in Equation 5.195, separate powers of λ , and then equate those terms with the same power of λ . So we consider several cases for $\lambda^{(a)}$.

Case λ^0 : $\hat{\mathcal{H}}_0 |v_m^{(0)}\rangle = W_m^{(0)} |v_m^{(0)}\rangle$

However, we already know the eigenvectors and eigenvalues for the original Hamiltonian so that

$$|v_m^{(0)}\rangle = |u_m\rangle \quad W_m^{(0)} = E_m$$

Case λ^1 : $\hat{\mathcal{H}}_0 |v_m^{(1)}\rangle + \hat{V} |v_m^{(0)}\rangle = W_m^{(0)} |v_m^{(1)}\rangle + W_m^{(1)} |v_m^{(0)}\rangle$

Substituting the results from the first case we find

$$\hat{\mathcal{H}}_0 |v_m^{(1)}\rangle + \hat{V} |u_m\rangle = E_m |v_m^{(1)}\rangle + W_m^{(1)} |u_m\rangle \quad (5.196)$$

Next make an “orthonormal expansion” of the first-order wave function as

$$|v_m^{(1)}\rangle = \sum_n a_{nm}^{(1)} |u_n\rangle \quad (5.197)$$

where $a_{nm}^{(i)}$ represent the expansion coefficients. Notice the order of the subscripts on the constants “a” and also notice that these constants carry the order of approximation as the superscript. The last expression says that each little piece of the new basis vector $|v_m\rangle$ is part of the old vector space and can be written as the summation over the old basis set. Substitute the orthonormal expansion into Equation 5.196 to get

$$\hat{\mathcal{H}}_0 \sum_n a_{nm}^{(1)} |u_n\rangle + \hat{V} |u_m\rangle = E_m \sum_n a_{nm}^{(1)} |u_n\rangle + W_m^{(1)} |u_m\rangle$$

Moving the original Hamiltonian under the summation and allowing it to operate on its eigenvector gives us

$$\sum_n a_{nm}^{(1)} E_n |u_n\rangle + \hat{V} |u_m\rangle = \sum_n a_{nm}^{(1)} E_m |u_n\rangle + W_m^{(1)} |u_m\rangle$$

The orthonormal expansion allows us to use our basic projection operators $\langle u_k |$ to isolate specific terms

$$\sum_n a_{nm}^{(1)} E_n \langle u_k | u_n \rangle + \langle u_k | \hat{V} | u_m \rangle = \sum_n a_{nm}^{(1)} E_m \langle u_k | u_n \rangle + W_m^{(1)} \langle u_k | u_m \rangle$$

The inner products between the discrete basis vectors produce Kronecker delta functions $\langle u_k | u_n \rangle = \delta_{kn}$. We find

$$a_{km}^{(1)} E_k + \mathcal{V}_{km} = a_{km}^{(1)} E_m + W_m^{(1)} \delta_{km} \quad (5.198)$$

Notice the matrix elements $\mathcal{V}_{km} = \langle u_k | \hat{V} | u_m \rangle$ for the small added potential in Equation 5.198. For functions, the matrix elements have the form

$$\mathcal{V}_{km} = \langle k | \hat{V} | m \rangle = \int dx u_k^* \hat{V} u_m$$

Two pieces of information are available from the Equation 5.198.

$$\begin{aligned} a_{km}^{(1)} &= -\frac{\mathcal{V}_{km}}{E_k - E_m} & k \neq m \\ W_m^{(1)} &= \mathcal{V}_{mm} & k = m \end{aligned} \quad (5.199a)$$

Therefore knowing the matrix of the potential $\hat{\mathcal{V}}$ evaluated in the original basis set u_m provides the first-order correction to the energy. Using Equation 5.194, the energy correct to first-order becomes

$$W_m = E_m + \lambda \hat{\mathcal{V}}_{mm} \Big|_{\lambda \rightarrow 1} = E_m + \hat{\mathcal{V}}_{mm} \quad (5.199b)$$

Similarly, using Equation 5.193a, the new basis vectors (not normalized) correct to first order can be written as

$$|v_m\rangle \cong |v_m^{(0)}\rangle + \lambda |v_m^{(1)}\rangle = |u_m\rangle + \lambda \sum_n a_{nm}^{(1)} |u_n\rangle \Big|_{\lambda=1} = |u_m\rangle + \sum_n a_{nm}^{(1)} |u_n\rangle \quad (5.199c)$$

Notice that setting $\lambda = 1$ in Equation 5.199c indicates that the perturbation is fully active (but not time dependent). We will need to normalize $|v_m\rangle$ to find the basis vector correct to first order. Notice that Equation 5.199a does not specify the coefficient $a_{mm}^{(1)}$. We will find these to be zero during the normalization procedure.

The new basis vectors to first-order approximation:

As just mentioned, Equations 5.193a and 5.197 produce the new basis vectors

$$|v_m\rangle \cong |v_m^{(0)}\rangle + \lambda |v_m^{(1)}\rangle = |u_m\rangle + \lambda \sum_n a_{nm}^{(1)} |u_n\rangle = |u_m\rangle + \sum_n a_{nm}^{(1)} |u_n\rangle$$

The new basis vector will be formulated once the coefficients $a_{nm}^{(1)}$ are known. The “off diagonal” elements ($m \neq n$) can be calculated from Equation 5.199a. However, the denominator in that equation does not allow one to calculate the “on diagonal” elements ($m = n$). By normalizing the new (first-order) basis vectors, we can find the diagonal elements $a_{mm}^{(1)}$

$$1 = \langle v_m | v_m \rangle \cong \langle u_m | u_m \rangle + \sum_a (a_{am}^{(1)})^* \langle u_a | u_m \rangle + \sum_b a_{bm}^{(1)} \langle u_m | u_b \rangle + \sum_{ab} (a_{am}^{(1)})^* a_{bm}^{(1)} \langle u_a | u_b \rangle$$

Replacing the inner products by the Kronecker delta functions, we obtained

$$0 \cong (a_{mm}^{(1)})^* + a_{mm}^{(1)} + \sum_a |a_{am}^{(1)}|^2 \quad (5.200)$$

Neglecting the second-order terms in the summation, we find

$$0 \cong (a_{mm}^{(1)})^* + a_{mm}^{(1)} = 2\text{Re}[a_{mm}^{(1)}]$$

so that

$$\text{Re}[a_{mm}^{(1)}] = 0 \quad \text{for all } m$$

The normalization procedure does not determine the imaginary part of $a_{mm}^{(1)}$. However, if one assumes that the difference vector $v_m - u_m$ is perpendicular to u_m then one finds the imaginary part of $a_{mm}^{(1)}$ must likewise be zero. Finally, Equation 5.199 becomes

$$|v_m\rangle \cong |u_m\rangle - \sum_{n \neq m} \frac{\mathcal{V}'_{nm}}{E_n - E_m} |u_n\rangle \quad (5.201)$$

correct to first order.

Equation 5.201 is related to a unitary rotation operator. We can see this by multiplying $\langle u_m |$ “on the right” of both sides of Equation 5.201 and then summing over both sides with respect to the index m .

$$\hat{S} = \sum_m |v_m\rangle \langle u_m| \cong \sum_m |u_m\rangle \langle u_m| - \sum_m \sum_{n \neq m} \frac{\mathcal{V}'_{nm}}{E_n - E_m} |u_n\rangle \langle u_m| \quad (5.202)$$

Using the completeness relation

$$\sum_m |u_m\rangle \langle u_m| = 1$$

Equation 5.202 becomes

$$\hat{S} \cong \hat{1} - \sum_m \sum_{n \neq m} \frac{\mathcal{V}'_{nm}}{E_n - E_m} |u_n\rangle \langle u_m|$$

We will derive this relation by another method in the next section.

The new energy levels to first-order approximation:

The new energy eigenvalues to first-order approximation are obtained from Equations 5.194 and 5.199 with $\lambda = 1$

$$W_m = E_m + \mathcal{V}'_{mm} \quad (5.203)$$

Example 5.24

For the infinitely deep well with $\mathcal{V}' = -ax$, calculate the new energy eigenvector $|v_1\rangle$ and eigenvalue W_1 correct to within first-order approximation.

SOLUTION

The perturbed well has potential energy given by $\mathcal{V}' = -ax$. To find the new basis functions and energy eigenvalues, we need the unperturbed energy eigenvalues

$$E_m = \frac{m^2 \pi^2 \hbar^2}{2m_e} \quad m = 1, 2, \dots$$

(where m_e denotes the mass of the electron) and the unperturbed eigenfunctions

$$u_m(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{m\pi x}{L}\right)$$

First using Equation 5.203, determine the new energy eigenvalues

$$\begin{aligned}
 W_1 &= E_1 + \mathcal{V}_{11} \\
 \mathcal{V}_{11} &= \langle u_1 | \hat{\mathcal{V}} | u_1 \rangle = \int_0^L dx \, u_1^*(x) (-ax) u_1(x) \\
 &= \int_0^L dx \, \frac{2}{L} (-ax) \sin^2\left(\frac{\pi x}{L}\right) = -\frac{aL}{2}
 \end{aligned}$$

Therefore, the energy of the perturbed first energy level is

$$W_1 = E_1 - \frac{aL}{2}$$

where we have been assuming that $a > 0$. Figure 5.45 indicates that the energy of the state decreases as the perturbation coefficient a increases. Notice that the second term in W_1 represents an average over the width of the well since

$$\langle \mathcal{V} \rangle = [\mathcal{V}(L) - \mathcal{V}(0)]/2 = [(-aL) - (-a0)]/2 = -aL/2$$

So the energy of the eigenstate changes by the average of the perturbing energy.

Continuing with the solution to the example, we next calculate the first-order correction to the first basis vector using Equation 5.201.

$$|v_1\rangle \cong |u_1\rangle - \sum_{n>1} \frac{\mathcal{V}_{n1}}{E_n - E_1} |u_n\rangle \quad (5.204)$$

where

$$E_n = \frac{n^2 \pi^2 \hbar^2}{2m_e} \quad n = 2, \dots$$

For the sake of illustration, we keep only the $n = 2$ term since

$$\frac{\mathcal{V}_{n1}}{E_n - E_1} = \frac{2m_e \mathcal{V}_{n1}}{(n^2 - 1)\pi^2 \hbar^2}$$

in Equation 5.204 decreases with increasing n . The corrected wave function to lowest-order approximation is therefore

$$|v_1\rangle \cong |u_1\rangle - \frac{\mathcal{V}_{21}}{E_2 - E_1} |u_2\rangle$$

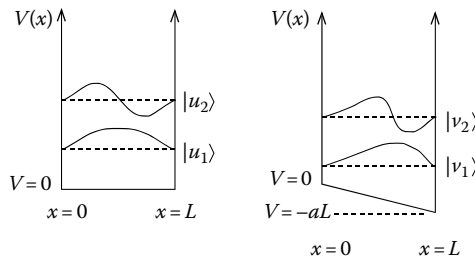


FIGURE 5.45 The infinitely deep square well and the perturbed well.

calculate the matrix elements of the perturbation as follows (notice that the matrix elements is calculated using the unperturbed basis vectors)

$$\mathcal{V}_{21} = \int_0^L dx \frac{2}{L} (-ax) \sin\left(\frac{2\pi x}{L}\right) \sin\left(\frac{\pi x}{L}\right) = 0.282aL$$

Therefore, the corrected wave function for the first energy level is

$$v_1(x) \cong |u_1\rangle - \frac{\mathcal{V}_{21}}{E_2 - E_1} |u_2\rangle \cong \sqrt{\frac{2}{L}} \sin\left(\frac{\pi x}{L}\right) - \frac{0.282aL}{E_2 - E_1} \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi x}{L}\right)$$

or substituting for the original energy level values

$$v_1(x) \cong |u_1\rangle - \frac{\mathcal{V}_{21}}{E_2 - E_1} |u_2\rangle \cong \sqrt{\frac{2}{L}} \sin\left(\frac{\pi x}{L}\right) - \frac{0.564aLm_e}{3\pi^2\hbar^2} \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi x}{L}\right)$$

5.9.3 UNITARY OPERATOR FOR TIME-INDEPENDENT PERTURBATION THEORY

In the previous section, we started with the energy basis vectors $\{|u_m\rangle\}$ for an unperturbed Hamiltonian $\hat{\mathcal{H}}_0$ such that $\hat{\mathcal{H}}_0|u_m\rangle = E_m|u_m\rangle$ where E_m represents the unperturbed energy eigenvalues. A new Hamiltonian $\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}$ with small time-independent perturbation potential $\hat{\mathcal{V}}$ produces a new energy basis set $\{|v_m\rangle\}$ and allowed energy levels W_m such that $\hat{\mathcal{H}}|v_m\rangle = W_m|v_m\rangle$.

In this section, we describe the same situation but develop a unitary “change of basis operator.” We look for an approximate expression for the unitary operator $\hat{S}$ that maps the original basis vectors into the new ones according to

$$|v_m\rangle = \hat{S}|u_m\rangle \quad (5.205)$$

As the reader knows from Chapter 3, the operator $\hat{S}$ can also be written as

$$\hat{S} = \sum_m |v_m\rangle\langle u_m| = \sum_{nm} S_{nm} |u_n\rangle\langle u_m| \quad (5.206)$$

Without a perturbing energy (i.e., $V=0$), the rotation operator $\hat{S}$ reduces to the unit operator (i.e., $\hat{S} = \hat{1}$) since then $|v_m\rangle = |u_m\rangle$. With a perturbation, we find the original basis vectors rotate into new ones. We therefore realize that the unitary rotation operator $\hat{S}$ must be a function of the perturbing energy $\hat{\mathcal{V}}$. We can make the same statement for the matrix of the operators. The matrix elements of $\hat{S}$, namely $S_{nm} = \langle u_n|\hat{S}|u_m\rangle \equiv \langle n|\hat{S}|m\rangle$, must be functions of the “small” matrix elements of $\hat{\mathcal{V}}$, namely $V_{nm} = \langle u_n|\hat{\mathcal{V}}|u_m\rangle \equiv \langle n|\hat{\mathcal{V}}|m\rangle$. We can write the functional dependence as $S_{nm} = S_{nm}(\mathcal{V}_{nm})$.

In some sense, the perturbation $\mathcal{V}_{nm}$ is similar to the rotation angle for the unitary operator S_{nm} . Obviously, once we know $\hat{S}$, then we also know the new basis set just as in the previous section. To find the approximation for $\hat{S}$, we follow a procedure that essentially duplicates that in the previous section (Figure 5.46).

We want an approximation for $\hat{S}$. The first step consists of Taylor expanding the matrix S_{nm} as

$$S_{nm}(\mathcal{V}_{nm}) = S_{nm}(0) + \frac{1}{1!} \frac{\partial S_{nm}(0)}{\partial \mathcal{V}_{nm}} \mathcal{V}_{nm} + \cdots = \sum_{nm} \sum_{i=0}^{\infty} \frac{1}{i!} \frac{\partial^i S_{nm}(0)}{\partial (\mathcal{V}_{nm})^i} (\mathcal{V}_{nm})^i \quad (5.207)$$

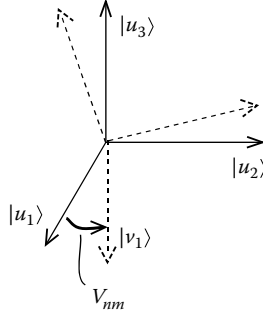


FIGURE 5.46 The operator S rotates the basis set.

Keep in mind that i gives the order of approximation and that $\mathcal{V}_{nm}$ is small. Substituting Equation 5.207 into Equation 5.206 provides

$$\hat{S} = \sum_{nm} S_{nm} |u_n\rangle \langle u_m| = \sum_{nm} \sum_{i=0}^{\infty} \frac{1}{i!} \frac{\partial^i S_{nm}(0)}{\partial \mathcal{V}_{nm}^i} (\mathcal{V}_{nm})^i |u_n\rangle \langle u_m| \quad (5.208)$$

To make the notation a little more compact, define

$$S_{nm}^{(i)} = \frac{1}{i!} \frac{\partial^i S_{nm}(0)}{\partial (\mathcal{V}_{nm})^i}$$

where i in “ (i) ” denotes the order of the approximation whereas i on a term like $(\mathcal{V}_{nm})^i$ indicates the power of i (i.e., a multiplication). Equation 5.208 becomes

$$\hat{S} = \sum_{nm} S_{nm} |u_n\rangle \langle u_m| = \sum_{nm} \sum_{i=0}^{\infty} (S_{nm}^{(i)} (\mathcal{V}_{nm})^i |u_n\rangle \langle u_m|$$

Next, substitute the rotation matrix $\hat{S}$ into Schrödinger’s equation

$$\begin{aligned} (\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) |v_m\rangle &= W_m |v_m\rangle \\ (\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) \hat{S} |u_m\rangle &= W_m \hat{S} |u_m\rangle \end{aligned}$$

to get

$$(\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) \sum_{ab} \sum_{i=0}^{\infty} (S_{ab}^{(i)} (\mathcal{V}_{ab})^i |u_a\rangle \langle u_b| u_m\rangle = W_m \sum_{ab} \sum_{i=0}^{\infty} (S_{ab}^{(i)} (\mathcal{V}_{ab})^i |u_a\rangle \langle u_b| u_m\rangle$$

using the orthonormality of the original basis vectors $\langle u_b | u_m \rangle = \delta_{bm}$ we find

$$(\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) \sum_a \sum_{i=0}^{\infty} (S_{am}^{(i)} (\mathcal{V}_{am})^i |u_a\rangle = W_m \sum_a \sum_{i=0}^{\infty} (S_{am}^{(i)} (\mathcal{V}_{am})^i |u_a\rangle \quad (5.209)$$

We need an expansion for W_m (which also depends on the matrix elements $\mathcal{V}_{nm}$)

$$W_m = \sum_{i=0}^{\infty} W_m^{(i)} \quad (5.210)$$

Substituting Equation 5.210 into Equation 5.209 yields

$$(\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) \sum_a \sum_{i=0}^{\infty} (S_{am}^{(i)})(\mathcal{V}_{am})^i |u_a\rangle = \sum_{j=0}^{\infty} W_m^{(j)} \sum_a \sum_{i=0}^{\infty} (S_{am}^{(i)})(\mathcal{V}_{am})^i |u_a\rangle$$

The $\mathcal{V}_{nm}$ are considered independent terms with the power i being the order of “smallness.” For example, $\mathcal{V}_{nm}\mathcal{V}_{am}$ is a second-order correction. Next, move the operators inside the summation on the left-hand side and then operate on both sides with $\langle k|$ to get

$$\sum_a \sum_{i=0}^{\infty} (S_{am}^{(i)})(\mathcal{V}_{am})^i \{ E_a \langle k|a\rangle + \langle k|\hat{\mathcal{V}}|a\rangle \} = \sum_{j=0}^{\infty} W_m^{(j)} \sum_a \sum_{i=0}^{\infty} (S_{am}^{(i)})(\mathcal{V}_{am})^i \langle k|a\rangle$$

Set $\langle k|\hat{\mathcal{V}}|a\rangle = V_{ka}$ and use the orthonormality relation $\langle k|a\rangle = \delta_{ka}$

$$\sum_{i=0}^{\infty} \left\{ (S_{km}^{(i)})(\mathcal{V}_{km})^i E_k + \sum_a (S_{am}^{(i)})(\mathcal{V}_{am})^i \mathcal{V}_{ka} \right\} = \sum_{i,j=0}^{\infty} W_m^{(j)} (S_{km}^{(i)})(\mathcal{V}_{km})^i \quad (5.211)$$

Equating corresponding orders of approximation in $\mathcal{V}_{nm}$ we get the following cases:

Case $i=0$: zeroth order

$$S_{km}^{(0)} E_k = W_m^{(0)} S_{km}^{(0)}$$

or, as we have found previously

$$W_m^{(0)} = E_m$$

We found previously that

$$S_{km}^{(0)} = \delta_{km}$$

Case $i=1$: first order

We keep only those terms in Equation 5.211 that give either $\mathcal{V}$ powers of 1 or $W^{(1)}$. Notice that a term such as

$$W_m^{(1)} S_{km}^{(0)} (\mathcal{V}_{km})^0$$

is first order because of $W^{(1)}$ even though V has a power of 0.

$$S_{km}^{(1)} \mathcal{V}_{km} E_k + \sum_a S_{am}^{(0)} (\mathcal{V}_{am})^0 \mathcal{V}_{ka} = W_m^{(0)} S_{km}^{(1)} (\mathcal{V}_{km})^1 + W_m^{(1)} S_{km}^{(0)} (\mathcal{V}_{km})^0$$

Substituting known quantities of

$$S_{km}^{(0)} = \delta_{km} W_m^{(0)} = E_m$$

we get

$$S_{km}^{(1)} \mathcal{V}_{km} E_k + \sum_a \delta_{am} \mathcal{V}_{ka} = E_m S_{km}^{(1)} \mathcal{V}_{km} + W_m^{(1)} \delta_{km}$$

or, removing the summation,

$$S_{km}^{(1)} \mathcal{V}_{km} E_k + \mathcal{V}_{km} = E_m S_{km}^{(1)} \mathcal{V}_{km} + W_m^{(1)} \delta_{km}$$

for $k \neq m$

$$S_{km}^{(1)} = \frac{1}{E_m - E_k}$$

and for $k = m$

$$W_m^{(1)} = \mathcal{V}_{mm}$$

same as in the previous section.

Rotation operator to first order:

The operator

$$\hat{S} = \sum_{nm} S_{nm} |u_n\rangle \langle u_m| = \sum_{nm} \sum_{i=0}^{\infty} (S_{nm}^{(i)}) (\mathcal{V}_{nm})^i |u_n\rangle \langle u_m|$$

can be manipulated to provide

$$\hat{S} = \sum_{nm} \{ S_{nm}^{(0)} |u_n\rangle \langle u_m| + (S_{nm}^{(1)}) (\mathcal{V}_{nm})^1 |u_n\rangle \langle u_m| + \dots \}$$

or

$$\hat{S} = \sum_{nm} \{ \delta_{nm} |u_n\rangle \langle u_m| + \frac{\mathcal{V}_{nm}}{E_m - E_n} |u_n\rangle \langle u_m| + \dots \}$$

The completeness relation can be used on the first term to get

$$\hat{S} = \hat{1} + \sum_{\substack{n \\ m \neq n}} \frac{\mathcal{V}_{nm}}{E_m - E_n} |u_n\rangle \langle u_m|$$

with $W_m^{(1)} = \mathcal{V}_{mm}$. These are the same results as obtained in the previous section. For example, the new basis vector $|v_m\rangle$ corresponding to the old basis vector $|u_m\rangle$ is

$$\begin{aligned} |v_m\rangle &= \hat{S}|m\rangle \cong \left\{ \hat{1} + \sum_{\substack{a \\ b \neq a}} \frac{\mathcal{V}_{ab}}{E_b - E_a} |a\rangle \langle b| \right\} |m\rangle \\ &= |m\rangle + \sum_{\substack{a \\ b \neq a}} \frac{\mathcal{V}_{am}}{E_m - E_a} |a\rangle \end{aligned}$$

5.10 TIME-DEPENDENT PERTURBATION THEORY

Interactions between particles or systems can produce energy transitions. For optoelectronics, one of the primary transition processes uses the interaction of electromagnetic energy with an atomic

system. A Hamiltonian $\hat{\mathcal{H}}_0$ describes the atomic system and provides the energy basis states and the energy levels. The interaction potential $\hat{\mathcal{V}}(t)$ (i.e., the perturbation) depends on time. The theory assumes that the perturbation does not change the basis states or the energy levels, but rather induces transitions between these fixed levels. The perturbation rotates the particle wave function (electron or hole) through Hilbert space so that the probability of the particle occupying one energy level or another changes with time. Therefore, the goal of the time-dependent perturbation theory consists of finding the time dependence of the wave function components.

Typically studies of optoelectronics apply the time-dependent perturbation theory to an electromagnetic wave interacting with an atom or an ensemble of atoms. Fermi's golden rule describes the matter–light interaction in this semiclassical approach, which uses the nonoperator form of the EM field. The same theory applies to other systems such as phonons.

5.10.1 PHYSICAL CONCEPT

Suppose the Hamiltonian

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}(t) \quad (5.212)$$

describes an atomic system subjected to a perturbation. The Hamiltonian $\hat{\mathcal{H}}_0$ refers to the atom and determines the energy basis states $\{|n\rangle = |E_n\rangle\}$ so that $\hat{\mathcal{H}}_0|n\rangle = E_n|n\rangle$. The interaction potential $\hat{\mathcal{V}}(t)$ describes the interaction of an external agent with the atomic system.

Consider an electromagnetic field incident on the atomic system as indicated in Figure 5.47 for the initial time $t = 0$. Assume that the atomic system consists of a quantum well with an electron in the first level as indicated by the dot in the figure. The atomic system can absorb a photon from the field and promote the electron from the first to the second level (subject to transition rules). The right-hand portion of Figure 5.47 shows the same information as the electron transitions from energy basis vector $|E_1\rangle$ to the basis vector $|E_2\rangle$ when the atom absorbs a quantum of energy. This transition of the electron from one basis vector to another should remind the reader of the effect of the ladder operators.

The transition of the electron from one state to another requires the electron occupation probability to change with time. Suppose the wave function for the electron has the form

$$|\psi(t)\rangle = \sum_n \beta_n(t)|n\rangle \quad (5.213)$$

In the case without any perturbation, the wave function evolves according to

$$|\psi(t)\rangle = e^{\hat{\mathcal{H}}_0 t / (i\hbar)} \sum_n \beta_n(0)|n\rangle = \sum_n \beta_n(0) e^{E_n t / (i\hbar)} |n\rangle \quad (\text{no perturbation}) \quad (5.214)$$

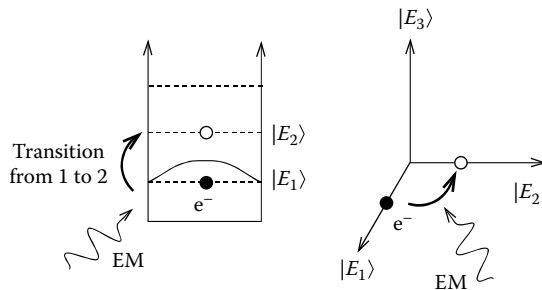


FIGURE 5.47 An electron absorbs a photon and makes a transition from the lowest level to next highest one.

where $\beta_n(t) = \beta_n(0) e^{E_n t / (\hbar)}$. In this “no perturbation” case, the probability of finding the electron in a particular state n at time t , denoted by $P(n, t)$, does not change from its initial value at $t=0$, denoted by $P(n, t=0)$, since

$$P(n, t) = |\beta_n(t)|^2 = |\beta_n(0) e^{E_n t / (\hbar)}|^2 = |\beta_n(0)|^2 = P(n, t=0) \quad (\text{no perturbation}) \quad (5.215)$$

This behavior occurs because the Hamiltonian describes a “closed system” that does not interact with external agents. The eigenvectors are exact “solutions” to Schrödinger’s equation using the Hamiltonian $\hat{H}_0$ in this case. The exact Hamiltonian introduces only the trivial factor $e^{E_n t / (\hbar)}$ into the motion of the wave function through Hilbert space.

What about the case of an atomic system interacting with the external agent? Now we see that Equation 5.214 cannot accurately describe this external-agent case because Equation 5.215 shows $P(n, t)$ does not change. The perturbation $\hat{V}(t)$ must produce an expansion coefficient with more than just the trivial factor. We will see below that the wave function must have the form

$$|\psi(t)\rangle = \sum_n a_n(t) e^{E_n t / (\hbar)} |n\rangle \quad (5.216)$$

in the Schrödinger picture where the trivial factor $e^{E_n t / (\hbar)}$ comes from $\hat{H}_0$ and the time-dependent term $a_n(t)$ comes from the perturbation $\hat{V}(t)$. Essentially, working in the Schrödinger picture produces the trivial factor $e^{E_n t / (\hbar)}$ in the wave function (without a perturbative driving force). Incorporating the interaction produces the nontrivial time dependence in the wave function. If the electron starts in state $|i\rangle$ at time $t=0$ (the i in the ket stands for initial) then the probability of finding it in state n after a time t must be

$$P(n, t) = |a_n(t) e^{E_n t / (\hbar)}|^2 = |a_n(t)|^2 \quad (5.217)$$

At time $t=0$, all of the a ’s must be zero except a_i because the electron starts in the initial state i . Also then, $a_i(0) = 1$ because the probabilities sum to 1. For later times t , any increase in a_n for $n \neq i$ must be attributed to increasing probability of finding the particle in state n . So, if the particle starts in state $|i\rangle$ then $a_n(t)$ gives the probability amplitude of a transition from state $|i\rangle$ to state $|n\rangle$ after a time t .

An example helps illustrates how the motion of the wave function in Hilbert space correlates with the transition probability. Consider the three vector diagrams in Figure 5.48. At time $t=0$, the wave function $|\psi(t)\rangle$ coincides with the $|1\rangle$ axis. The probability amplitude at $t=0$ must be $\beta_n(0) = a_n(0) = \delta_{ni}$ and therefore the probability values must be $\text{Prob}(n=1, t=0) = 1$ and $\text{Prob}(n \neq 1, t=0) = 0$. Therefore the particle definitely occupies the first energy eigenstate at $t=0$. The second plot in Figure 5.48 at $t=2$, shows the electron partly occupies both the first and second eigenstates. There exists a nonzero probability of finding it in either basis state. According to the figure,

$$\text{Prob}(n=1, t=2) = \text{Prob}(n=2, t=2) = 0.5$$

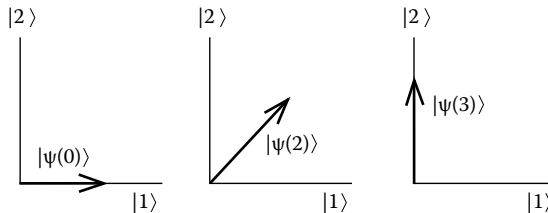


FIGURE 5.48 The probability of the electron occupying the second state increases with time.

The third plot in Figure 5.48 at time $t = 3$ shows that the electron must be in state $|2\rangle$ alone since the wave function $|\psi(3)\rangle$ coincides with basis vector $|2\rangle$. At $t = 3$, the probability of finding the electron in state $|2\rangle$ must be

$$\text{Prob}(n = 2, t = 3) = |\beta_2|^2 = 1$$

Notice how the probability of finding the particle in state $|1\rangle$ decreases with time, while the probability of finding the particle in state $|2\rangle$ increases.

Unlike the unperturbed system, multiple measurements of the energy of the electron do not always return the same value. The reason concerns the fact that the eigenstates of $\hat{\mathcal{H}}_0$ do not describe the full system. In particular, it does not describe the external agent (light field) nor the interaction between the light field and the atomic system. The external agent, the electromagnetic field, disturbs the state of the particle between successive measurements. The basis function for the atomic system alone does not include one for the optical field. However, given the basis set for the full Hamiltonian $\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}} + \hat{\mathcal{H}}_{\text{Other}}$ (where $\hat{\mathcal{H}}_{\text{Other}}$ is the environment and $\hat{\mathcal{V}}$ the interaction between the atomic and environmental systems) and then a measurement of $\hat{\mathcal{H}}$ must cause the full wave function to collapse to one of the full basis vectors from which it does not move (we have not included the case of degenerate eigenstates).

Several points should be kept in mind while reading through the next section that shows the calculation of the time-dependent probability. First, the procedure uses the Schrödinger representation but does not replace β_n with $a_n e^{E_n t / (i\hbar)}$ (see Problem 5.82 for this alternate procedure). Instead, the procedure directly finds β_n , which then turns out to have the form $a_n e^{E_n t / (i\hbar)}$. Second, these components β_n have exact expressions until we make an approximation of the form $\beta(t) = \beta^{(0)}(t) + \beta^{(1)}(t) + \dots$ (similar to the Taylor expansion). Third, assume that the particle starts in state $|i\rangle$ so that $\beta_n(0) = \beta_n^{(0)}(0) = \delta_{ni}$ and $\beta_n^{(j)}(0) = 0$ for $j \geq 1$. Fourth, the transition matrix elements $\mathcal{V}_{fi} = \langle f | \mathcal{V} | i \rangle$ determine the final states f that can be reached from the initial states i . That is, if $\mathcal{V}_{fi} = \langle f | \mathcal{V} | i \rangle = 0$ then a transition cannot take place. Stated equivalently, these selection rules determine the allowed transitions.

5.10.2 TIME-DEPENDENT PERTURBATION THEORY FORMALISM IN THE SCHRÖDINGER PICTURE

The perturbed Hamiltonian $\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}(x, t)$ consists of the Hamiltonian $\hat{\mathcal{H}}_0$ for the closed system and the perturbation $\hat{\mathcal{V}}(t)$. Schrödinger's equation becomes

$$\hat{\mathcal{H}}|\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \rightarrow (\hat{\mathcal{H}}_0 + \hat{\mathcal{V}})|\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad (5.218)$$

The unperturbed Hamiltonian $\hat{\mathcal{H}}_0$ produces the energy basis set $\{u_n = |n\rangle\}$ so that

$$\hat{\mathcal{H}}_0 |n\rangle = E_n |n\rangle$$

We assume that the Hamiltonian $\hat{\mathcal{H}}$ has the same basis set $\{u_n = |n\rangle\}$ as $\hat{\mathcal{H}}_0$. The boundary conditions on the system determine the basis set and the eigenvalues. This step relegates the perturbation to causing transitions between the basis vectors.

As usual, we write the solution to the Schrödinger wave equation (SWE)

$$\hat{\mathcal{H}}|\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad (5.219)$$

and

$$|\Psi(t)\rangle = \sum_n \beta_n(t) |n\rangle \quad (5.220)$$

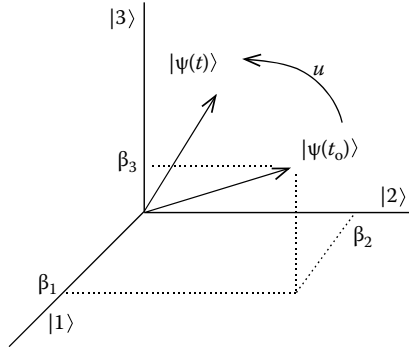


FIGURE 5.49 The Hamiltonian causes the wave functions to move in Hilbert space.

Recall that the wave vector $|\Psi(t)\rangle$ moves in Hilbert space in response to the Hamiltonian $\hat{\mathcal{H}}$ (via the evolution operator) as indicated in Figure 5.49. The components $\beta_n(t)$ must be related to the probability of finding the electron in the state $|n\rangle$. As an important point, we assume that the particle starts in state $|i\rangle$ at time $t=0$ (where $i=1, 2, \dots$ and should not be confused with the complex number $i = \sqrt{-1}$). To find the components $\beta_n(t)$, start by substituting $|\Psi(t)\rangle$ (Equation 5.220) into Schrödinger's equation (Equation 5.219)

$$(\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad \rightarrow \quad (\hat{\mathcal{H}}_0 + \hat{\mathcal{V}}) \sum_n \beta_n(t) |n\rangle = i\hbar \frac{\partial}{\partial t} \sum_n \beta_n(t) |n\rangle$$

Move the unperturbed Hamiltonian and the potential inside the summation to find

$$\sum_n \beta_n(t) (E_n + \hat{\mathcal{V}}) |n\rangle = i\hbar \sum_n \dot{\beta}_n(t) |n\rangle$$

where the dot over the symbol β indicates the time derivative. Operate on both sides of the equation with $\langle m|$ to find

$$\sum_n \beta_n(t) (E_n \langle m|n\rangle + \langle m|\hat{\mathcal{V}}|n\rangle) = i\hbar \sum_n \dot{\beta}_n(t) \langle m|n\rangle$$

The orthonormality of the basis vectors $\langle m|n\rangle = \delta_{mn}$ transforms the previous equation to

$$E_m \beta_m(t) + \sum_n \beta_n(t) \langle m|\hat{\mathcal{V}}(x, t)|n\rangle = i\hbar \dot{\beta}_m(t)$$

which can be rewritten as

$$\dot{\beta}_m(t) - \frac{E_m}{i\hbar} \beta_m(t) = \frac{1}{i\hbar} \sum_n \beta_n(t) \mathcal{V}_{mn}(t) \quad (5.221)$$

where the matrix elements can be written as

$$\mathcal{V}_{mn}(t) = \langle m|\hat{\mathcal{V}}(x, t)|n\rangle = \int dx u_m^* \hat{\mathcal{V}}(x, t) u_n$$

for the basis set consisting of functions of x .

We must solve Equation 5.221 for the components $\beta_n(t)$; this can most easily be handled by using an integrating factor $\mu_m(t)$. Rather than actually solve for the integrating factor, we will just state the results (see Appendix E)

$$\mu_m(t) = \exp\left(-\frac{E_m}{i\hbar}t\right) \quad (5.222)$$

Multiplying the integrating factor on both sides of Equation 5.221, we can write

$$\mu_m \dot{\beta}_m - \frac{E_m}{i\hbar} \mu_m \beta_m = \frac{1}{i\hbar} \sum_n \mu_m \beta_n(t) \mathcal{V}_{mn} \quad (5.223)$$

Noting that

$$\frac{d}{dt}(\mu_m \beta_m) = \dot{\mu}_m \beta_m + \mu_m \dot{\beta}_m \quad \text{and} \quad \dot{\mu}_m = -\frac{E_m}{i\hbar} \exp\left(-\frac{E_m}{i\hbar}t\right) = -\frac{E_m}{i\hbar} \mu_m$$

Equation 5.223 becomes

$$\frac{d}{dt}[\mu_m(t)\beta_m(t)] = \frac{1}{i\hbar} \mu_m(t) \sum_n \beta_n(t) \mathcal{V}_{mn}(t) \quad (5.224)$$

We need to solve this last equation for the components $\beta_n(t)$ in the first and last terms. Assume that the perturbation starts at $t=0$ and integrate both sides with respect to time.

$$\mu_m(t)\beta_m(t) = \mu_m(0)\beta_m(0) + \frac{1}{i\hbar} \int_0^t d\tau \mu_m(\tau) \sum_n \beta_n(\tau) \mathcal{V}_{mn}(\tau) \quad (5.225)$$

Substituting for $\mu_m(t)$, noting from Equation 5.222 that $\mu_m(0) = 1$, and using the fact that the particle starts in state $|i\rangle$ so that

$$\beta_n(0) = \delta_{ni} \quad (5.226)$$

we find

$$\beta_m(t) = \mu_m^{-1}(t)\delta_{mi} + \frac{\mu_m^{-1}(t)}{i\hbar} \sum_n \int_0^t d\tau \mu_m(\tau) \beta_n(\tau) \mathcal{V}_{mn}(\tau) \quad (5.227)$$

To this point, the solution is exact.

Now we make the approximation by writing the components $\beta_n(t)$ as a summation

$$\beta_n(t) = \beta_n^{(0)}(t) + \beta_n^{(1)}(t) + \dots$$

where the superscripts provide the order of the approximation. Substituting the approximation for the components $\beta_n(t)$ into Equation 5.227 provides

$$\beta_m^{(0)}(t) + \beta_m^{(1)}(t) + \dots = \mu_m^{-1}(t)\delta_{mi} + \frac{\mu_m^{-1}(t)}{i\hbar} \sum_n \int_0^t d\tau \mu_m(\tau) [\beta_n^{(0)}(\tau) + \beta_n^{(1)}(\tau) + \dots] \mathcal{V}_{mn}(\tau)$$

Note that the approximation term $\beta_n^{(0)} \mathcal{V}_{mn}$ has order “(1)” even though $\beta_n^{(0)}$ has order “(0)” since we consider the interaction potential $\mathcal{V}_{mn}$ to be small (i.e., it has order “(1)”). Equating corresponding orders of approximation in the previous equation provides

$$\beta_m^{(0)}(t) = \mu_m^{-1}(t) \delta_{mi} \quad (5.228)$$

$$\beta_m^{(1)}(t) = \frac{\mu_m^{-1}(t)}{i\hbar} \sum_n \int_0^t d\tau \mu_m(\tau) \beta_n^{(0)}(\tau) \mathcal{V}_{mn}(\tau) \quad (5.229)$$

and so on. Notice how Equation 5.229 invokes Equation 5.228 in the integral. So once we solve for the zeroth-order approximation for the component, we can immediately find the first-order approximation. Higher-order terms work the same way. This last equation gives the lowest order correction to the probability amplitude.

The Kronecker delta function in Equation 5.228 suggests considering two separate cases when finding the probability amplitude correction $\beta_m^{(1)}(t)$. The first case for $m = i$ corresponds to finding the probability amplitude for the particle remaining in the initial state. The second case $m \neq i$ produces the probability amplitude for the particle making a transition to state m .

Case $m = i$. We calculate the probability amplitude $\beta_i(t)$ for the particle to remain in the initial state. The lowest order approximation gives (using Equations 5.228 and 5.222)

$$\beta_n^{(0)}(t) = \delta_{ni} \mu_n^{-1}(t) = \delta_{ni} \exp\left(\frac{E_n}{i\hbar} t\right) \quad (5.230)$$

Substituting Equation 5.230 into Equation 5.229 with $m = i$, we find

$$\beta_i^{(1)}(t) = \frac{\mu_i^{-1}(t)}{i\hbar} \sum_n \int_0^t d\tau \mu_i(\tau) \beta_n^{(0)}(\tau) \mathcal{V}_{in}(\tau) = \frac{\mu_i^{-1}(t)}{i\hbar} \int_0^t d\tau \mu_i(\tau) \exp\left(\frac{E_i}{i\hbar} \tau\right) \mathcal{V}_{ii}(\tau)$$

Substituting Equation 5.222 for the remaining integrating factors in the previous equation we find

$$\beta_i^{(1)}(t) = \frac{1}{i\hbar} \exp\left(\frac{E_i}{i\hbar} t\right) \int_0^t d\tau \mathcal{V}_{ii}(\tau)$$

So therefore the approximate value for $\beta_i(t)$ must be

$$\beta_i(t) = \beta_i^{(0)}(t) + \beta_i^{(1)}(t) + \dots = \exp\left(\frac{E_i}{i\hbar} t\right) + \frac{1}{i\hbar} \exp\left(\frac{E_i}{i\hbar} t\right) \int_0^t d\tau \mathcal{V}_{ii}(\tau) + \dots \quad (5.231)$$

Case $m \neq i$: We find the component $\beta_m(t)$ corresponding to a final state $|m\rangle$ different from the initial state $|i\rangle$. The lowest order approximation $\beta_m^{(0)}$ for $m \neq i$ must be

$$\beta_m^{(0)}(t) = 0$$

The procedure finds the probability amplitude for a particle to make a transition from the initial state $|i\rangle$ to a different final state $|m\rangle$.

We start with Equation 5.229

$$\beta_m^{(1)}(t) = \frac{\mu_m^{-1}(t)}{i\hbar} \sum_n \int_0^t d\tau \mu_m(\tau) \beta_n^{(0)}(\tau) \mathcal{V}_{mn}(\tau) = \frac{\mu_m^{-1}(t)}{i\hbar} \sum_n \int_0^t d\tau \mu_m(\tau) \delta_{ni} \mu_i^{-1}(\tau) \mathcal{V}_{mi}(\tau)$$

Substitute Equation 5.222 for the integrating factors to find

$$\beta_m^{(1)}(t) = \frac{1}{i\hbar} \exp\left(\frac{E_m}{i\hbar} t\right) \int_0^t d\tau \exp\left(-\frac{E_m - E_i}{i\hbar} \tau\right) \mathcal{V}_{mi}(\tau)$$

We often write the difference in energy as $E_m - E_i = E_{mi}$ and also

$$\omega_{mi} = \omega_m - \omega_i = \frac{E_m - E_i}{\hbar} = \frac{E_{mi}}{\hbar} \quad (5.232)$$

The reader must keep track of the distinction between matrix elements and this new notation for differences between quantities—matrix elements refer to operators. Using this notation

$$\beta_m^{(1)}(t) = \frac{1}{i\hbar} \exp\left(\frac{E_m}{i\hbar} t\right) \int_0^t d\tau \exp\left(-\frac{E_{mi}}{i\hbar} \tau\right) \mathcal{V}_{mi}(\tau) \quad (5.233)$$

Therefore, the components $\beta_m(t)$ for $m \neq i$ are approximately given by

$$\beta_m(t) = \beta_m^{(0)}(t) + \beta_m^{(1)}(t) + \cdots = 0 + \frac{1}{i\hbar} \exp\left(\frac{E_m}{i\hbar} t\right) \int_0^t d\tau \exp\left(-\frac{E_{mi}}{i\hbar} \tau\right) \mathcal{V}_{mi}(\tau) + \cdots \quad (5.234)$$

In summary, the expansion coefficients in

$$|\Psi(t)\rangle = \sum_n \beta_n(t) |n\rangle \quad (5.235a)$$

are given by Equations 5.234 and 5.232

$$\beta_m(t) = \delta_{mi} \exp\left(\frac{E_i}{i\hbar} t\right) + \frac{1}{i\hbar} \exp\left(\frac{E_m}{i\hbar} t\right) \int_0^t d\tau \exp\left(-\frac{E_{mi}}{i\hbar} \tau\right) \mathcal{V}_{mi}(\tau) + \cdots \quad (5.235b)$$

5.10.3 EXAMPLE FOR FURTHER THOUGHT AND QUESTIONS

Up to this point, we have discussed both the time-independent and the time-dependent perturbation theories. For time-independent perturbation theory, a small change in the Hamiltonian of the system produces a small change in the energy basis set and energy eigenvalues. A particle in the modified system can occupy one of the new basis states. Time-dependent perturbations $\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \mathcal{V}(x, t)$, on the other hand, induce a particle to make transitions between basis states. The unperturbed Hamiltonian $\hat{\mathcal{H}}_0$ produces the energy basis vectors. We now discuss a system for which the particle rides along with the shifting energy levels.

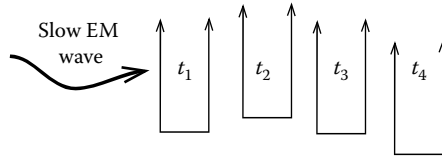


FIGURE 5.50 EM wave applied to infinitely deep well.

The problem can be restated. For time-independent perturbations, we might imagine that a particle starts in state $|u_i\rangle$ of the original (unperturbed) system. Now, we slowly change the physical system and keep track of the particle. These slow adiabatic changes take place on a time scale much longer than any time constant associated with the system. For this example, we find that (to first order) the particle stays in the same eigenstate but the eigenstate changes $|u_i\rangle \rightarrow |v_i\rangle$ (notice that the subscript i stays the same but different basis vectors). The following discussion compares the results from the time-dependent and time-independent cases.

Case 1: Time-independent perturbation theory

Consider the infinitely deep well from two points of view—both of which give similar results. First consider time-independent perturbation theory. Suppose we apply a very slowly oscillating electric field to an infinitely deep well as shown in Figure 5.50; the change might take years for example. Suppose initially, the bottom of the well has the potential $\mathcal{V}=0$ at time $t=0$. We can consider the time t to be a parameter that, in effect, gives the perturbed potential at the bottom of the well. We assume the potential at the bottom of the well has the form

$$\mathcal{V}(t) = c \sin[\omega_0(t - t')]$$

where

ω_0 is a very small angular frequency

t' just sets the phase

Let E_n be the unperturbed energy of the state $|u_n\rangle$. We found in the previous section that the time-independent perturbation theory (to first order) provided the formula for the energy of the perturbed eigenstates $|v_n\rangle$ as

$$W_n = E_n + \langle u_n | \hat{\mathcal{V}} | u_n \rangle$$

Here we consider the time t to be a parameter and c must be small. The expectation value becomes

$$\langle u_n | \hat{\mathcal{V}} | u_n \rangle = \mathcal{V}(t) \langle u_n | u_n \rangle = \mathcal{V}(t)$$

since the inner product involves an integral over x but not t . Therefore the modified energy eigenvalues must be

$$W_n(t) = E_n + \mathcal{V}(t) \quad (5.236)$$

Using the new basis set, a general wave function has the form

$$|\Psi(t)\rangle = \sum_n \beta_n(0) \exp\left(\frac{W_n t}{i\hbar}\right) |v_n\rangle \quad (5.237)$$

Working through the time-independent perturbation formulas for the basis vectors we find

$$|v_n\rangle \cong |u_n\rangle - \sum_{m \neq n} \frac{\mathcal{V}_{mn}}{E_m - E_n} |u_m\rangle = |u_n\rangle \quad (5.238)$$

since

$$\langle u_m | \mathcal{V}(t) | u_n \rangle = \mathcal{V}(t) \langle u_m | u_n \rangle = 0 \quad m \neq n$$

Equation 5.238 shows that the shape of the wave function does not change. Equation 5.236 shows that the energy-separation between levels

$$W_{n+1} - W_n = E_{n+1} - E_n$$

remains unchanged. Figure 5.50 shows the well moving higher and lower in energy. By substituting Equations 5.236 and 5.238 into Equation 5.237, a general wave function can be written as

$$|\Psi(t)\rangle = \sum_n \beta_n(0) \exp\left(\frac{E_n t}{i\hbar}\right) |u_n\rangle \exp\left(\frac{ct \sin[\omega_0(t - t')]}{i\hbar}\right)$$

Therefore the probability of the electron occupying the state n (to low order of approximation) can be written as

$$\text{Prob}_{\text{new}}(n) = \left| \beta_n(0) \exp\left(\frac{E_n t}{i\hbar}\right) \exp\left(\frac{c \sin(\omega_0 t) t}{i\hbar}\right) \right|^2 = |\beta_n(0)|^2 = \text{Prob}_{\text{old}}(n)$$

which shows that the slow perturbation does not change the probability of occupying any given level.

Case 2: Time-dependent perturbation theory

Next, consider the same situation using time-dependent perturbation theory. Actually, we use the same procedure as for the time-dependent perturbation theory without making the approximations. The Hamiltonian is given by

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}(t)$$

where we assume that the energy eigenstates for both Hamiltonians $\hat{\mathcal{H}}$, $\hat{\mathcal{H}}_0$ are $|u_n\rangle$. Schrödinger's equation reads

$$(\hat{\mathcal{H}}_0 + \mathcal{V}(t)) |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle$$

Substitute the expansion

$$|\Psi(t)\rangle = \sum_n \beta_n(t) |u_n\rangle$$

to get

$$\sum_n \beta_n(t) [E_n + \mathcal{V}(t)] |u_n\rangle = i\hbar \sum_n \dot{\beta}_n(t) |u_n\rangle$$

where the dot above β_n in the right-hand term indicates a derivative with respect to time. Operating with $\langle u_m |$ on both sides and using $\langle u_m | u_n \rangle = \delta_{mn}$ gives

$$\beta_m(t)[E_m + V(t)] = i\hbar\dot{\beta}_m(t) \quad (5.239)$$

There are not any transitions between energy levels due to the selection rule embodied in $\langle u_m | \hat{V} | u_n \rangle = 0$ for $m \neq n$. The small size of $\hbar\omega_0$ compared with the energy-separation between allowed energy levels provides another reason that there are not any transitions. Equation 5.239 can be rewritten as

$$\frac{d\beta_m}{\beta_m} = dt \frac{E_m + V(t)}{i\hbar}$$

which has the solution

$$\beta_m(t) = \beta_m(0) \exp\left(\frac{E_m t}{i\hbar}\right) \exp\left(\frac{1}{i\hbar} \int_0^t d\tau V(\tau)\right)$$

Assume $V(t) = c \sin [\omega_0(t - t')]$ with ω_0 very small and using t' to set the phase. Assume the observation time extends from t_0 to time t such that $t - t_0$ is very small compared with $1/\omega_0$. The integral can be replaced by

$$\int_{t_0=0}^t d\tau V(\tau) \cong c \sin [\omega_0(t - t')] \int_{t_0=0}^t d\tau = c \sin [\omega_0(t - t')] t$$

since $V(t) = c \sin [\omega_0(t - t')]$ is approximately constant over the region of integration. The general wave function has the form

$$|\Psi(t)\rangle = \sum_n \beta_n(0) \exp\left(\frac{E_n t}{i\hbar}\right) |u_n\rangle \exp\left(\frac{ct \sin [\omega_0(t - t')]}{i\hbar}\right)$$

the same as the time-independent perturbation theory. The probability of the particle remaining in a given state must be the same as for the time-independent case.

5.10.4 TIME-DEPENDENT PERTURBATION THEORY IN THE INTERACTION REPRESENTATION

The interaction representation for quantum mechanics is especially suited for time-dependent perturbation theory. Once again, the Hamiltonian $\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{V}(x, t)$ consists of the atomic Hamiltonian $\hat{\mathcal{H}}_0$ and the interaction potential $\hat{V}(x, t)$ due to an external agent. The atomic Hamiltonian produces the basis set $\{|n\rangle\}$ satisfying $\hat{\mathcal{H}}_0|n\rangle = E_n|n\rangle$. Both the operators and the wave functions depend on time in the interaction representation. The wave functions move through Hilbert space only in response to the interaction potential $\hat{V}(x, t)$. A unitary operator $\hat{u} = \exp[\hat{\mathcal{H}}_0 t / (i\hbar)]$ removes the trivial motion from the wave function and places it in the operators; consequently, the operators depend on time. Without any potential $\hat{V}(x, t)$, the wave functions remain stationary and the operators remain trivially time dependent; that is, the interaction picture reduces to the Heisenberg picture. The motion of the wave function in Hilbert space reflects the dynamics embedded in the interaction potential.

The evolution operator removes the trivial time dependence from the wave function

$$\hat{u}(t) = \exp\left(\frac{\hat{\mathcal{H}}_0}{i\hbar}t\right) \quad \text{with } \hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}(x, t) \quad (5.240)$$

The interaction potential in the interaction picture has the form $\hat{\mathcal{V}}_I = \hat{u}^\dagger \hat{\mathcal{V}} \hat{u}$ and produces the interaction wave function $|\Psi_I\rangle$ given by

$$|\Psi_s\rangle = \hat{u} |\Psi_I\rangle \quad (5.241)$$

The wave function $|\Psi_s\rangle$ is the usual Schrödinger wave function embodying the dynamics of the full Hamiltonian $\hat{\mathcal{H}}$. The equation of motion for the interaction wave function can be written as (Section 5.8)

$$\hat{\mathcal{V}}_I |\Psi_I(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi_I(t)\rangle \quad \text{or} \quad \frac{\partial}{\partial t} |\Psi_I(t)\rangle = \frac{1}{i\hbar} \hat{\mathcal{V}}_I |\Psi_I(t)\rangle \quad (5.242)$$

We wish to find an expression for the wave function in the interaction representation. First, formally integrate Equation 5.242

$$|\Psi_I(t)\rangle = |\Psi_I(0)\rangle + \frac{1}{i\hbar} \int_0^t d\tau \hat{\mathcal{V}}_I(\tau) |\Psi_I(\tau)\rangle \quad (5.243)$$

where we have assumed that the interaction starts at $t=0$. We can write another equation (see below) by substituting Equation 5.243 into itself, which assumes that the interaction wave functions only slightly move in Hilbert space for small interaction potentials.

Zeroth-order approximation:

The lowest order approximation can be found by noting small interaction potentials $\hat{\mathcal{V}}(x, t)$ lead to small changes in the wave function with time. Neglecting the small integral term in Equation 5.243 produces the lowest order approximation

$$|\Psi_I(t)\rangle \cong |\Psi_I(0)\rangle = |\Psi_s(0)\rangle \quad (5.244)$$

where the second equality comes from the fact that $\hat{u}(0) = \hat{1}$ in Equation 5.240. This last equation says that to lowest order, the interaction-picture wave function remains stationary in Hilbert space. Therefore to lowest order, the probabilities calculated by projecting the wave function $|\Psi_I(t)\rangle$ onto the basis vectors remain independent of time. The trivial terms $e^{-iE\tau/\hbar}$ that occur in changing back from the interaction to Schrödinger picture do not have any effect on the probability of finding a particle in a given basis state.

Higher-order approximation:

We obtain subsequent approximations by substituting the wave functions into the integral. The total first-order approximation can be found by substituting Equation 5.243 into Equation 5.242

$$|\Psi_I(t)\rangle = |\Psi_I(0)\rangle + \frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_I(t_1) |\Psi_I(0)\rangle \quad (5.245)$$

The total second-order approximation can be found by substituting Equation 5.244 into Equation 5.242 to obtain

$$|\Psi_1(t)\rangle = \left\{ 1 + \frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_1(t_1) + \left(\frac{1}{i\hbar} \right)^2 \int_0^t dt_1 \int_0^{t_1} dt_2 \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) \right\} |\Psi_1(0)\rangle \quad (5.246)$$

We can continue this process to find any order of approximation.

5.10.5 EVOLUTION OPERATOR IN THE INTERACTION REPRESENTATION

We can find a unitary operator that moves the interaction wave function forward in time. Equation 5.245 essentially gives the evolution operator $\hat{U}$ defined by

$$|\psi_1(t)\rangle = \hat{U}(t) |\psi_1(0)\rangle \quad (5.247)$$

Note the use of capital U so as not to confuse $\hat{U}$ with the operator $\hat{u}$ that maps between the Schrödinger and interaction pictures. Equation 5.245 approximates $\hat{U}$ by

$$\hat{U} = \left\{ 1 + \frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_1(t_1) + \left(\frac{1}{i\hbar} \right)^2 \int_0^t dt_1 \int_0^{t_1} dt_2 \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) \right\} \quad (5.248)$$

which is somewhat reminiscent of writing the operator as an exponential. For example, if the interaction potential were independent of time (but it is not) then the operator would reduce to

$$\hat{U} \approx 1 + \frac{\hat{\mathcal{V}}_1 t}{i\hbar} + \left(\frac{\hat{\mathcal{V}}_1 t}{i\hbar} \right)^2 + \dots = \exp \left(\frac{\hat{\mathcal{V}}_1 t}{i\hbar} \right)$$

In order to see how this operator can be related to an exponential, we must digress and discuss the time-ordered product.

We define the time-ordered product $\hat{T}$ as follows:

$$\hat{T} \left\{ \hat{\mathcal{V}}(t_1) \hat{\mathcal{V}}(t_2) \hat{\mathcal{V}}(t_3) \right\} = \hat{\mathcal{V}}(t_1) \hat{\mathcal{V}}(t_3) \hat{\mathcal{V}}(t_2) \quad \text{when } t_1 > t_3 > t_2 \quad (5.249)$$

The time-ordered product can also be defined in terms of a step function.

$$\Theta(t) = \begin{cases} 1 & t > 0 \\ 1/2 & t = 0 \\ 0 & t < 0 \end{cases} \quad (5.250)$$

Note the $1/2$ for $t=0$. The third term in Equation 5.247 has two operators and notice that the integration limits require $t_1 > t_2$. We will want to change the limits on both integrals to cover the interval $(0, t)$. Therefore we must keep track of the time ordering. The time-ordered product of two operators can be written in terms of the step function as

$$\hat{T} \hat{\mathcal{V}}(t_1) \hat{\mathcal{V}}(t_2) = \Theta(t_1 - t_2) \hat{\mathcal{V}}(t_1) \hat{\mathcal{V}}(t_2) + \Theta(t_2 - t_1) \hat{\mathcal{V}}(t_2) \hat{\mathcal{V}}(t_1) \quad (5.251a)$$

Consider the following integral

$$\begin{aligned} \frac{1}{2!} \int_0^t dt_1 \int_0^t dt_2 \hat{T} \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) &= \frac{1}{2} \int_0^t dt_1 \int_0^t dt_2 \Theta(t_1 - t_2) \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) \\ &+ \frac{1}{2} \int_0^t dt_1 \int_0^t dt_2 \Theta(t_2 - t_1) \hat{\mathcal{V}}_1(t_2) \hat{\mathcal{V}}_1(t_1) \end{aligned} \quad (5.251b)$$

Interchanging the dummy variables t_1, t_2 in the last integral shows that it is the same as the middle integral. Therefore, by the properties of the step function we find

$$\frac{1}{2!} \int_0^t dt_1 \int_0^t dt_2 \hat{T} \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) = \int_0^t dt_1 \int_0^{t_1} dt_2 \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) \quad (5.252)$$

which agrees with the second integral in Equation 5.247.

We are now in a position to write an operator that evolves the wave function for the interaction representation. Substituting Equation 5.248 into Equation 5.247 yields

$$|\Psi_I(t)\rangle = \hat{T} \left\{ \hat{1} + \frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_1(t_1) + \left(\frac{1}{i\hbar}\right)^2 \frac{1}{2} \int_0^t dt_1 \int_0^t dt_2 \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) + \cdots \right\} |\Psi_I(0)\rangle \quad (5.253)$$

The term in brackets can be written as an exponential

$$\hat{T} \left\{ \hat{1} + \frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_1(t_1) + \left(\frac{1}{i\hbar}\right)^2 \frac{1}{2!} \int_0^t dt_1 \int_0^t dt_2 \hat{\mathcal{V}}_1(t_1) \hat{\mathcal{V}}_1(t_2) + \cdots \right\} = \hat{T} e^{\frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_1(t_1)}$$

Therefore, as a result, the evolution operator in the interaction representation has the form

$$\hat{U} = \hat{T} e^{\frac{1}{i\hbar} \int_0^t dt_1 \hat{\mathcal{V}}_1(t_1)}$$

5.11 INTRODUCTION TO OPTICAL TRANSITIONS

The previous sections in this chapter have been primarily concerned with the mathematics, structure, and introduction to quantum mechanics. For optoelectronics, the optical transition provides one of the most important applications for the time-dependent perturbation theory. For completeness, the present section provides an introduction to the “semiclassical” theory (i.e., light represented by waves rather than a photon field) for optically induced transitions. Further discussion is best found in the book *Physics of Optoelectronics*. The discussion culminates in Fermi’s golden rule as found in the subsequent section.

5.11.1 EM INTERACTION POTENTIAL

Suppose an electron occupies an energy eigenstate in a single atom and that an electromagnetic wave washes over that atom. What is the probability that the electron will make an upward or a

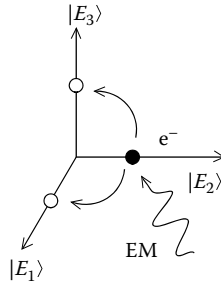


FIGURE 5.51 The EM wave can induce upward and downward transitions.

downward transition to a higher or lower energy level, respectively? Interestingly, the frequency of the electromagnetic waves necessary to induce a transition does not necessarily need to be in the optical range; it all depends on the type of “atom.” Figure 5.51 shows an electron occupying the second energy level along with an incident electromagnetic wave. If the atom (i.e., electron) absorbs energy from the wave (stimulated absorption) then the electron makes an upward transition. If the wave induces a downward transition (stimulated emission) then the atom releases energy to the bathing field. “Semiclassical” theory describes the effects of a classical electromagnetic traveling wave. With this form of interaction, one ignores the particle properties (i.e., discrete energy properties) of the electromagnetic wave. The coherent states found in the quantum theory of light most closely describe classical electromagnetic waves (refer to the references covering Quantum Optics or to Quantum Electrodynamics QED). More in-depth discussion of the absorption and emission of light can be found there as well as in the book *Physics of Optoelectronics*. For now, let us see how perturbation theory can be applied to the problem of optically induced transitions.

Classically a material can produce or absorb light when an electromagnetic field interacts with dipoles within the material. The classical expression for the dipole interaction energy can be written as proportional to $\vec{p} \cdot \vec{E}$ where $\vec{p}$, $\vec{E}$ denote the dipole moment and electric field, respectively. Quantum mechanically, we represent the interaction energy by operators. We might write $\hat{V} = \hat{\mu}E$ for the quantum mechanical 1-D case. The “dipole moment” operator $\hat{\mu}$, which is Hermitian, describes the strength of the interaction between the oscillating electric field and the atom. Sometimes people write the interaction energy in the explicitly Hermitian form $\hat{V} = \frac{1}{2} \{ \hat{\mu}E + (\hat{\mu}E)^+ \}$ just in case they use the complex form of the field. We will use the complex form of the field, explicitly $E = E_0 e^{-i\omega t}$.

Assume that the unperturbed Hamiltonian $\hat{H}_0$ describes an “atom” (located at the origin of the coordinate system) without any incident EM wave. The Hermitian interaction energy can be written as

$$\hat{V}(x, t) = \begin{cases} \hat{\mu}(x) \frac{E_0}{2} e^{-i\omega t} + [\hat{\mu}(x) \frac{E_0}{2} e^{-i\omega t}]^+ & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (5.254)$$

which provides a small perturbation for the full Hamiltonian $\hat{H} = \hat{H}_0 + \hat{V}(x, t)$. Some books set $\hat{H}' = \hat{\mu} \frac{E_0}{2}$. We assume that the angular frequency of the incident electromagnetic field is always positive $\omega > 0$. The matrix elements of the dipole operator $\hat{\mu}$ will be real constants of proportionality. Also assume a real amplitude E_0 for the oscillating electric field. Equation 5.254 shows that the interaction potential must be Hermitian $\hat{V} = \hat{V}^+$ and therefore it must be an observable. Equation 5.254 can be rewritten as

$$\hat{V}(x, t) = \hat{\mu}(x) \frac{E_0}{2} (e^{-i\omega t} + e^{+i\omega t}) = \hat{\mu}(x) E_0 \cos(\omega t) \quad (5.255)$$

We can see that the interaction potential must be Hermitian from this last expression by noting the dipole moment operator $\hat{\mu}$ must be Hermitian. The reader should realize that a phase factor could be added to the exponential term in the interaction energy to obtain a sine wave rather than the cosine wave. As is appropriate for time-dependent perturbation theory, assume the set $\{|u_n\rangle = |n\rangle\}$ contains the energy eigenvectors for the unperturbed Hamiltonian.

5.11.2 INTEGRAL FOR THE PROBABILITY AMPLITUDE

In Section 5.10.2, we show the wave function

$$|\Psi(t)\rangle = \sum_n \beta_n(t) |n\rangle \quad (5.256)$$

satisfies Schrödinger's equation

$$\hat{\mathcal{H}}|\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad (5.257)$$

provided

$$\beta_n(t) = \begin{cases} e^{-i\omega_i t} + \frac{1}{i\hbar} e^{-i\omega_i t} \int_0^t d\tau \mathcal{V}'_{ii}(\tau) + \cdots & n = i \\ \frac{1}{i\hbar} e^{-i\omega_n t} \int_0^t d\tau e^{i\omega_n \tau} \mathcal{V}'_{ni}(\tau) + \cdots & n \neq i \end{cases} \quad (5.258)$$

where $\omega_{ni} = \frac{E_n}{\hbar} - \frac{E_i}{\hbar}$ and the electron is assumed to start in state $|i\rangle$. For example, $|i\rangle = |2\rangle$ for Figure 5.51. Recall that the component $\beta_n(t)$ of the vector describes the probability amplitude of finding the electron in state $|n\rangle$ after a time t ; consequently, the probability must be given by $\beta_n^* \beta_n$. Obviously therefore, the component $\beta_n(t)$ must be related to the probability (and the transition rate) of the electron making a transition from state $|i\rangle$ to state $|n\rangle$ since the electron started in state $|i\rangle$.

$$\text{Prob}(i \rightarrow n) = |\beta_n(t)|^2 \quad (5.259)$$

We can take the case of either $n = i$ or $n \neq i$. If we take the case of $n = i$ then we are calculating the probability that the particle will not make a transition. Although that is interesting in itself, we are more interested in the case of $n \neq i$. We can find the rate of transition by taking the time derivative of the probability

$$R_{i \rightarrow n} = \frac{d}{dt} \text{Prob}(i \rightarrow n) \quad (5.260)$$

To find the probability and rate of transition (to first-order approximation) for the case of $n \neq i$, we must calculate the integral in

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \int_0^t d\tau e^{i\omega_n \tau} V_{ni}(x, \tau) = \frac{1}{i\hbar} e^{-i\omega_n t} \int_0^t d\tau e^{i\omega_n \tau} \langle n | \hat{\mathcal{V}}(x, \tau) | i \rangle \quad (5.261)$$

from Equation 5.258. Notice in the matrix element $\langle n | \hat{\mathcal{V}} | i \rangle$ how the perturbation induces a transition from right to left. The reader should keep in mind that ω represents the angular frequency of the

electromagnetic wave whereas ω_{ni} denotes the angular frequency corresponding to the difference in energy. Many times people say that the atom requires the light to have angular frequency $\omega_{ni} = (E_n - E_i)/\hbar$ in order for the atom to participate in stimulated absorption or emission. However, this section indicates there is some slight probability for a transition when $\omega \neq \omega_{ni}$ for small times. The integral in Equation 5.261 can be evaluated by substituting Equation 5.254 to get

$$\begin{aligned}\beta_n(t) &= \frac{1}{i\hbar} e^{-i\omega_n t} \int_0^t d\tau e^{i\omega_{ni}\tau} \langle n | \hat{V}(x, \tau) | i \rangle \\ &= \frac{1}{i\hbar} e^{-i\omega_n t} \int_0^t d\tau e^{i\omega_{ni}\tau} \langle n | \left\{ \hat{\mu}(x) \frac{E_o}{2} e^{-i\omega\tau} + \left[\hat{\mu}(x) \frac{E_o}{2} e^{-i\omega\tau} \right]^+ \right\} | i \rangle\end{aligned}$$

Now calculate the adjoint, distribute the projection operator and the ket through the braces, and use the definition

$$\langle n | \hat{\mu} | i \rangle = \mu_{ni} \quad (5.262)$$

to find

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \int_0^t d\tau e^{i\omega_{ni}\tau} \left\{ \mu_{ni} \frac{E_o}{2} e^{-i\omega\tau} + \mu_{ni} \frac{E_o}{2} e^{+i\omega\tau} \right\}$$

Keep in mind that the matrix element μ_{ni} is just a constant of proportionality that describes the strength of the interaction between the impressed electromagnetic field and the atom. It is this induced dipole matrix element μ_{ni} that gives the “transition selection rules.” The induced dipole matrix element is a nontrivial factor and should be explored in greater detail depending on the transitions which could also involve phonons. Factoring out the constant values from the integral, produces

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_o}{2} \int_0^t d\tau e^{i\omega_{ni}\tau} [e^{-i\omega\tau} + e^{+i\omega\tau}]$$

or

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_o}{2} \int_0^t d\tau [e^{i(\omega_{ni}-\omega)\tau} + e^{i(\omega_{ni}+\omega)\tau}] \quad (5.263)$$

Performing the integration provides

$$\beta_n(t) = \frac{-1}{\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_o}{2} \left[\frac{e^{i(\omega_{ni}-\omega)t} - 1}{\omega_{ni} - \omega} + \frac{e^{i(\omega_{ni}+\omega)t} - 1}{(\omega_{ni} + \omega)} \right] \quad (5.264)$$

Equation 5.264 contains terms for both absorption and emission of light. Notice from the denominators that the first term dominates when $\omega \cong \omega_{ni}$ and the second term dominates when $\omega \cong -\omega_{ni}$. Recalling the definition

$$\omega_{ni} = \frac{E_n}{\hbar} - \frac{E_i}{\hbar} \quad (5.265)$$

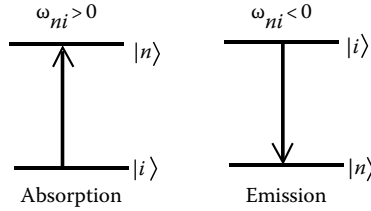


FIGURE 5.52 The sign of ω_{ni} indicates absorption or emission.

and the fact that the angular frequency of the incident light must always be positive $\omega > 0$, we see that the first term in Equation 5.264 corresponds to the absorption of light since

$$0 < \omega \cong \omega_{ni} = \frac{E_n}{\hbar} - \frac{E_i}{\hbar} \rightarrow E_n \geq E_i \quad (5.266)$$

so that the energy of the final state must be larger than the energy of the initial state. The second term in Equation 5.264 corresponds to emission since

$$0 < \omega \cong -\omega_{ni} = \frac{E_i}{\hbar} - \frac{E_n}{\hbar} \rightarrow E_i \geq E_n \quad (5.267)$$

so that the initial state, in this case, has a larger energy than the final state which can only happen when the atom emits a photon. Figure 5.52 shows a type of two-level atom. Although we used the denominators of Equation 5.264 to determine which term corresponds to absorption and emission, another method consists of looking at the arguments of the exponential functions in Equation 5.263. We come back to the problem of calculating the probability of absorption and emission after a brief interlude for the monumentally important subject of the rotating wave approximation.

5.11.3 ROTATING WAVE APPROXIMATION

We wish to evaluate integrals such as

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_0}{2} \int_0^t d\tau [e^{i(\omega_{ni}-\omega)\tau} + e^{i(\omega_{ni}+\omega)\tau}] \quad (5.268)$$

The exponentials have arguments that correspond to very high frequencies or very low frequencies. For example, when $\omega \cong \omega_{ni}$, we see that the first exponential has approximately constant value while the second one has frequency $\omega + \omega_{ni} \cong 2\omega_{ni}$.

There are two methods to evaluate integrals with “slow” and “fast” functions. The previous section showed one method consists of evaluating the integral in Equation 5.268

$$\beta_n(t) = \frac{-1}{\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_0}{2} \left[\frac{e^{i(\omega_{ni}-\omega)t} - 1}{\omega_{ni} - \omega} + \frac{e^{i(\omega_{ni}+\omega)t} - 1}{(\omega_{ni} + \omega)} \right] \quad (5.269)$$

and neglecting terms based on the size of the denominator. When the angular frequency of the wave ω approximately matches the atomic resonant frequency $\omega \cong \omega_{ni}$ then the first term in Equation 5.269 dominates the second term. Of course, we could also have $\omega \cong -\omega_{ni}$, in which case the second term dominates by virtue of the denominator.

The second method (refer to the book *Physics of Optoelectronics*), the rotating wave approximation, averages a sinusoidal wave over many cycles and finds a result of zero. This method applies to integrals of the form

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_o}{2} \int_0^t d\tau [e^{i(\omega_{ni}-\omega)\tau} + e^{i(\omega_{ni}+\omega)\tau}] \quad (5.270)$$

Notice this method applies to the integral prior to integrating rather than using Equation 5.259. Using, for example, $\omega \cong \omega_{ni}$ means that $\exp \{i(\omega_{ni} - \omega)t\}$ must be approximately constant while $\exp \{i(\omega_{ni} + \omega)t\}$ must be a high-frequency sinusoidally varying wave. The integral looks very similar to an average from calculus given by

$$\langle f \rangle = \frac{1}{t} \int_0^t dt' f(t') \quad (5.271)$$

If over the interval $(0, t)$, the first integrand in Equation 5.220 does not change much, then its integral will be nonzero. On the other hand, the second term runs through many oscillations (rotating wave) and the average over the interval (Equation 5.271) yields zero.

5.11.4 ABSORPTION

Now return to the calculation for the probability of a transition. First consider the case for absorption where $\omega \cong \omega_{ni}$. We found Equation 5.264

$$\beta_n(t) = \frac{-1}{\hbar} e^{-i\omega_n t} \mu_{ni} \left[\frac{E_o}{2} \frac{e^{i(\omega_{ni}-\omega)t} - 1}{\omega_{ni} - \omega} + \frac{e^{i(\omega_{ni}+\omega)t} - 1}{(\omega_{ni} + \omega)} \right]$$

from Equation 5.263

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_o}{2} \int_0^t d\tau [e^{i(\omega_{ni}-\omega)\tau} + e^{i(\omega_{ni}+\omega)\tau}]$$

The rotating wave approximation allows us to drop the second term in both of the above two equations. Therefore, absorption produces the time-dependent probability amplitude

$$\beta_n(t) = \frac{-1}{\hbar} e^{-i\omega_n t} \mu_{ni} \left[\frac{E_o}{2} \frac{e^{i(\omega_{ni}-\omega)t} - 1}{\omega_{ni} - \omega} \right] \quad (5.272)$$

Recall that β_n is the component of the wave function parallel to the $|n\rangle$ axis. The component β_n depends on time in a nontrivial manner and causes the wave function to move away from the i th axis and move closer to the n th axis.

We can find the transition probability for absorption

$$\begin{aligned} \text{Prob}(i \rightarrow n) &= |\beta_n|^2 = \left(\frac{\mu_{ni} E_o}{2\hbar} \right)^2 \frac{[e^{i(\omega_{ni}-\omega)t} - 1][e^{i(\omega_{ni}-\omega)t} - 1]^*}{(\omega_{ni} - \omega)^2} \\ &= \left(\frac{\mu_{ni} E_o}{2\hbar} \right)^2 \frac{2 - e^{i(\omega_{ni}-\omega)t} - e^{-i(\omega_{ni}-\omega)t}}{(\omega_{ni} - \omega)^2} \end{aligned} \quad (5.273)$$

Using the trigonometric identities,

$$e^{i(\omega_{ni}-\omega)t} + e^{-i(\omega_{ni}-\omega)t} = 2 \cos [(\omega_{ni} - \omega)t] \quad \text{and} \quad \cos(2\theta) = \cos^2(\theta) - \sin^2(\theta) = 1 - 2 \sin^2(\theta)$$

where $\theta = (\omega_{ni} - \omega)t/2$, the probability for an upward transition can be written as

$$\text{Prob}_{\text{abs}}(i \rightarrow n) = |\beta_n|^2 = \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{\sin^2 \left[\frac{1}{2} (\omega_{ni} - \omega)t \right]}{(\omega_{ni} - \omega)^2} \quad (5.274)$$

Before discussing this last result, consider the case for stimulated emission.

5.11.5 EMISSION

The case for emission is obtained when $\omega \cong -\omega_{ni} > 0$. Equation 5.263, specifically

$$\beta_n(t) = \frac{1}{i\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_0}{2} \int_0^t d\tau [e^{i(\omega_{ni}-\omega)\tau} + e^{i(\omega_{ni}+\omega)\tau}]$$

gives Equation 5.264, here repeated

$$\beta_n(t) = \frac{-1}{\hbar} e^{-i\omega_n t} \mu_{ni} \frac{E_0}{2} \left[\frac{e^{i(\omega_{ni}-\omega)t} - 1}{\omega_{ni} - \omega} + \frac{e^{i(\omega_{ni}+\omega)t} - 1}{(\omega_{ni} + \omega)} \right]$$

The rotating wave approximation allows us to drop the first term in both of the above two equations. Therefore, for emission, the component of the wave function parallel to the $|n\rangle$ axis (the probability amplitude) must be

$$\beta_n(t) = \frac{-1}{\hbar} e^{-i\omega_n t} \mu_{ni} \left[\frac{E_0}{2} \frac{e^{i(\omega_{ni}+\omega)t} - 1}{\omega_{ni} + \omega} \right]$$

Following the same procedure as for absorption, we find

$$\text{Prob}_{\text{emis}}(i \rightarrow n) = |\beta_n|^2 = \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{\sin^2 \left[\frac{1}{2} (\omega_{ni} + \omega)t \right]}{(\omega_{ni} + \omega)^2} \quad (5.275)$$

The reader might be surprised to find the probability for absorption to be numerically the same as the probability for emission. This is easy to see from the last equation by setting

$$\omega_{ni} = \frac{E_n}{\hbar} - \frac{E_i}{\hbar} = - \left(\frac{E_i}{\hbar} - \frac{E_n}{\hbar} \right) = -\omega_{in}$$

to get

$$\text{Prob}_{\text{emis}}(i \rightarrow n) = |\beta_n|^2 = \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{\sin^2 \left[\frac{1}{2} (-\omega_{ni} + \omega)t \right]}{(-\omega_{ni} + \omega)^2} = \text{Prob}_{\text{abs}}(i \rightarrow n) \quad (5.276a)$$

from Equation 5.274. Because the probabilities are equal, we leave off the subscript for absorption and emission and write

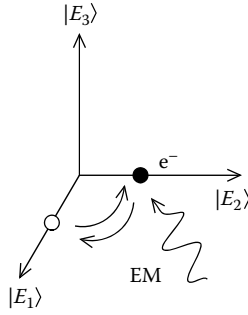


FIGURE 5.53 Absorption and emission of a quantum of energy between two states.

$$\text{Prob}(i \rightarrow n) = |\beta_n|^2 = \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{\sin^2 \left[\frac{1}{2} (\omega_{ni} - \omega) t \right]}{(\omega_{ni} - \omega)^2} \quad (5.276b)$$

Notice however, an atom in its ground state cannot emit a photon and so the probability for that emission event must be zero.

We should make a few comments. First Equation 5.276a shows that absorption and emission have the same probability

$$\text{Prob}_{\text{emis}}(i \rightarrow n) = \text{Prob}_{\text{abs}}(n \rightarrow i)$$

The transition must occur between the same two states (as shown in Figure 5.53). The dipole matrix element has the same value for either transition $i \rightarrow n$ or $n \rightarrow i$ since it is Hermitian (and assumed real) and therefore $\mu_{in} = \mu_{ni}$. We cannot expect the relation to hold in the case of transitions involving three levels for example when $2 \rightarrow 1$ and $2 \rightarrow 3$. In this case, the dipole matrix element is not necessarily the same for the two transitions.

5.11.6 DISCUSSION OF THE RESULTS

Figure 5.54 shows a plot of the probability as a function of angular frequency for two different times t_1 and t_2 ($t_2 > t_1$). Notice that the probability becomes narrower for larger times.

Let us discuss the case of stimulated emission with the proviso that the same considerations hold for the case of absorption. The highest probability for emission occurs when $\omega = \omega_{ni}$ as shown in the figure. We can find the peak probability from Equation 5.276b by Taylor expanding the sine term (assuming the argument is small) to get

$$\text{Peak Prob} = |\beta_n|^2 = \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{t^2}{4} \omega = \omega_{ni} \quad (5.277)$$

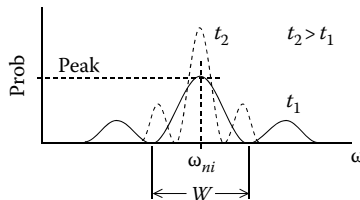


FIGURE 5.54 Plot of probability versus driving frequency and parameterized by time.

which occurs when the frequency of the electromagnetic wave ω exactly matches the natural resonant frequency of the atom ω_{ni} . The width of the probability curve can be estimated by finding the point where it touches the horizontal axis. Setting the sine term in Equation 5.276b to zero

$$\sin^2 \left[\frac{1}{2} (\omega_{ni} - \omega) t \right] = 0$$

which occurs at $(\omega_{ni} - \omega)t/2 = \pm\pi/2$, we find that the width is

$$W = \frac{2\pi}{t}$$

The width of the curve narrows with time. According to the figure, a frequency off-resonance can induce a transition.

Equations 5.274 and 5.276 show that stronger electric fields increase the rate of transition. Equation 5.276 shows that for small times (as is appropriate for the approximation of the probability amplitudes β_n), that the transition probability increases linearly with time. This might lead someone to anticipate that the transition requires some average time. If we know the probability as a function of time $P(t)$ then we can calculate an average time as

$$\bar{t} = \langle t \rangle = \int_0^{\infty} t P(t) dt$$

We can similarly calculate the variance for the time for emission as

$$\sigma_t^2 = E(t - \bar{t})^2 = \langle t^2 \rangle - \langle t \rangle^2$$

We then see that the exact difference in energy E between the initial and final level is not exactly known by the Heisenberg uncertainty relation

$$\sigma_E \sigma_t \geq \frac{\hbar}{2}$$

5.12 FERMİ'S GOLDEN RULE

Fermi's golden rule gives the “rate” of transition from a single state to a set of states, which can be described by the “density-of-state” function. In extensions of the rule, the single initial state is expanded into a range as well.

5.12.1 INTRODUCTORY CONCEPTS ON PROBABILITY

Fermi's golden rule provides a computational tool incorporating the time-dependent perturbation theory (in particular, the interaction potential) to determine transition rates from one state to another. It can be applied to cases of particles scattering from a localized potential or to optoelectronic cases involving phonon or photon absorption, emission, or scattering. For context, we focus on the absorption of a photon by a system having an initial electron state $|i\rangle$ and a range of possible “final” states $\{|n\rangle\}$.

As shown in Figure 5.55, an electron makes a transition from an initial state $|i\rangle$ to one of the many final states $\{|n\rangle\}$. The probability of transition must be given by

$$P_V = \sum_n P(i \rightarrow n) \quad (5.278)$$

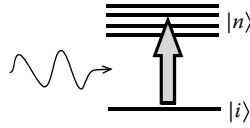


FIGURE 5.55 Schematic illustration of an electromagnetically induced transition from an initial state i to one of the final states n .

This last equation represents the total probability of transition from a single initial state to one of many possible final states. The subscript V occurs since later, the units of volume will be included. For a semiconductor, the final states closely approximate a continuum. In such a case, the probability $P(i \rightarrow n)$ should be interpreted as the probability of transition “per final state” and the summation should be changed to an integral over the final states.

The total probability in Equation 5.278 requires a sum over the integers n corresponding to the final states $\{|n\rangle\}$. Apparently, we imagine that the electron lodges itself in one of the final energy basis states somewhat similar to the manner in which a rolling marble might lodge itself in an indentation in the floor. However, we know that the particle as a quantum object has a wave function that might be a linear combination of the energy basis states $|n\rangle$. In such a case, the electron simultaneously exists in two or more states $|n\rangle$ (consider two for simplicity) and cannot really be considered as in any one final state. According to classical probability theory, it would appear that we should subtract the probability that the electron can be in both states at the same time from Equation 5.278 to find

$$\text{Prob}(A \text{ or } B) \sim \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \text{ and } B)$$

However, we assume that a measurement of the energy of the electron has taken place, the wave function has collapsed, and that the electron resides in one of the energy basis states. Therefore the $\text{Prob}(A \text{ or } B)$ reduces to the sum of probabilities as in Equation 5.278

$$\text{Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B)$$

since upon observation, the particle can only be found in a single state. Fermi’s golden rule therefore integrates over the range of final states to find the number of transitions occurring per unit time.

This section also shows how Fermi’s golden rule can be used to demonstrate the semiconductor gain. A detailed treatment must wait for discussions on the density operator, the Bloch wave function and the reduced density of states.

5.12.2 DEFINITION OF THE DENSITY OF STATES

Later chapters will discuss the density of states in greater detail; however, we give an introduction now in order to discuss the transition probability provided by Fermi’s golden rule. The localized states provide the simplest starting point because we do not need the added complexity of determining the allowed wave vectors.

The energy density-of-states (DOS) function measures the number of energy states in each unit energy interval in each unit volume of the crystal

$$g(E) = \frac{\text{\#States}}{\text{Energy} \cdot \text{XalVol}} \quad (5.279)$$

We need to explore the reasons for dividing by the energy and the crystal volume.

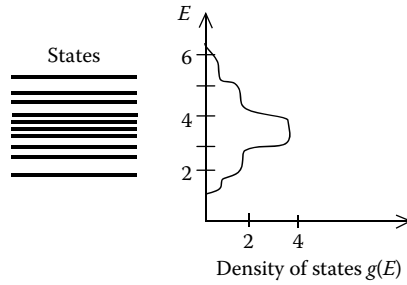


FIGURE 5.56 The density of states for the discrete levels shown on the left-hand side. The plot assumes the system has unit volume (1 cm^3) and the levels have energy measured in eV.

First consider the conceptual method of calculating the DOS and then the reason for including the energy in the denominator to form “per unit energy.” Suppose we have a system with the energy levels shown on the left-hand side of Figure 5.56. Assume for now that the states occur in a unit volume of material (say 1 cm^3). The figure shows four energy states in the energy interval between 3 and 4 eV. The density of states at $E = 3.5$ must be

$$g(3.5) = \frac{\text{\#States}}{\text{Energy} \times \text{Vol}} = \frac{4}{1 \text{ eV} \times 1 \text{ cm}^3} = 4$$

Similarly, between 4 and 5 eV, we find two states and the density-of-states function has the value $g(4.5) = 2$, and so on. Essentially, we just add up the number of states with a given energy. The graph shows the number of states versus energy; for illustration, the graph has been flipped over on its side. Generally we use finer energy scales and the material has larger numbers of states (10^{17}) so that the graph generally appears much smoother than the one in Figure 5.56 since the energy levels essentially form a continuum.

The “per unit energy” characterizes the type of state and the type of material. For transitions though, a large number of states at a particular energy (see subsequent sections below) can be expected to increase the transition rate. For a marble example, if a marble rolls across a floor, the larger the number of indentations increases the likelihood of the marble lodging itself into one of these indentations (or scatters from them). The “per unit energy” part would be somewhat similar to a marble rolling uphill with the indentations near the end of the trajectory (assuming the marble is free from obstructions or other states on its way up). The greater the number of closely spaced indentations near the top (i.e., higher number “per unit energy” because vertical position equates to energy) the more likely the marble will be captured by the indentations. However, unlike the marble example, without final states available to the quantum particle, the quantum particle will never be found anywhere but in the initial state. You see, the marble (without reference to the electron for the moment) has other states available to it all the way up the hill although they are not indentations. The marble states are characterized by position and speed. If those marble states are eliminated, then the marble would not leave its initial state.

The definition of density of states uses “per unit crystal volume” in order to remove geometrical considerations from the measure of the type of state. Obviously, if each unit volume has N_v states (electron traps for example) given by

$$N_v = \int dE g(E) = \int d(\text{energy}) \frac{\text{\#States}}{\text{Energy} * \text{Vol}} = \frac{\text{\#States}}{\text{Vol}} \quad (5.280)$$

then the volume V must have $N = N_v V$ states. Changing the volume changes the total number. To obtain a measure of the “type of state,” we need to remove the trivial dependence on crystal volume.

Generally a person would find a total number 10^{17} states very significant for a 1 cm^3 semiconductor than for a cube with 10 km on a side. Such a cube would have less than one state in each cm^3 ! Making a device out of this 1 cm^3 piece would produce nearly perfect devices if the states were related to imperfections. Lower numbers of transition states in a material translate to fewer transitions. So, it is important to know the number of states in a “standard” volume to know the quality of the material or, in the case that the states perform a useful function (e.g., optical transitions), the suitability of the material for the desired function.

What are the states? The states can be those in an atom. The states can also be traps that an electron momentarily occupies until being released back into the conduction band. The states might be recombination centers that electrons enter where they recombine with holes. Traps and recombination centers can be produced by defects in the crystal. Surface states occur on the surface of semiconductors as an inevitable consequence of the interrupted crystal structure. The density of defects can be low within the interior of the semiconductor and high near the surface; as a result, the density of states can depend on position. Later we discuss the “extended” states in a semiconductor.

Let us consider several examples for the density of states. First, suppose a crystal has two discrete states (i.e., single states) in each unit volume of crystal. Figure 5.57 shows the two states on the left-hand side of the graph. The density-of-state function consists of two Dirac delta functions of the form

$$g(E) = \delta(E - E_1) + \delta(E - E_2)$$

Integrating over energy gives the number of states in each unit volume

$$N_v = \int_0^\infty dE g(E) = \int_0^\infty dE [\delta(E - E_1) + \delta(E - E_2)] = 2$$

If the crystal has the size $1 \times 4 \text{ cm}^3$ then the total number of states in the entire crystal must given by

$$N = \int_0^4 dV N_v = 8$$

as illustrated in Figure 5.58. Although this example shows a uniform distribution of states within the volume V , the number of states per unit volume N_v can depend on the position within the crystal. For example, the growth conditions of the crystal can vary or perhaps the surface becomes damaged after growth.

As a second example, consider localized states near the conduction band of a semiconductor as might occur for amorphous silicon. Figure 5.59 shows a sequence of graphs. The first graph shows

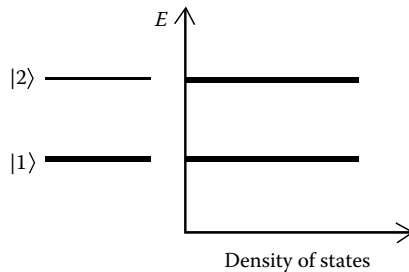


FIGURE 5.57 The density of states for two discrete states shown on the left side.

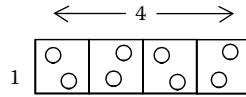


FIGURE 5.58 Each unit volume has two states and the full volume has eight.

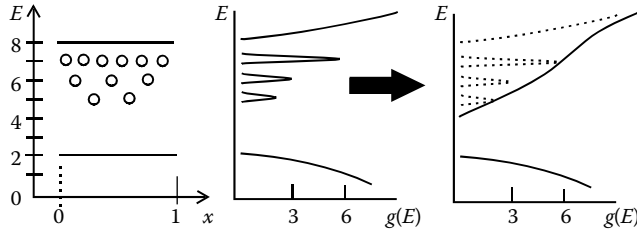


FIGURE 5.59 Transition from discrete localized states to the continuum.

the distribution of states versus the position x within the semiconductor. Notice that the states come closer together (in energy) near the conduction band edge. As a note, amorphous materials have mobility edges rather than band edges. The second graph shows the density-of-states function versus energy. A sharp Gaussian spike represents the number of states at each energy. At 7 eV, the material has six states (traps) per unit length in the semiconductor as shown in the first graph. The second graph shows a spike at 7 eV. Actual amorphous silicon has very large numbers of traps near the upper mobility edge and they form a continuum as represented in the third graph. This example shows how the density of states depends on position and how closely space discrete levels form a continuum.

5.12.3 EQUATIONS FOR FERMI'S GOLDEN RULE

The previous section shows that the probability of a transition from an initial state $|i\rangle$ to a final state $|n\rangle$ can be written as

$$\text{Prob}(i \rightarrow n) = |\beta_n|^2 = \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{\sin^2 \left[\frac{1}{2} (\omega_{ni} - \omega) t \right]}{(\omega_{ni} - \omega)^2} \quad (5.281)$$

with an applied electric field of

$$\vec{E}(x, t) = E_0 \cos(\omega t) \quad (5.282)$$

which leads to the perturbing interaction energy

$$\hat{V}(x, t) = \hat{\mu}(x) \frac{E_0}{2} (e^{-i\omega t} + e^{+i\omega t}) = \hat{\mu}(x) E_0 \cos(\omega t) \quad (5.283)$$

The dipole moment operator $\hat{\mu}$ provides the matrix elements μ_{ni} that describe the interaction strength between the field and the atom. The dipole matrix element μ_{ni} can be zero for certain final states $|n\rangle$ and Equation 5.281 then shows that the transition from the initial to the proposed final state cannot occur. As in Section 5.8, the symbol ω_{ni} represents the difference in energy between the final state $|n\rangle$ and initial state $|i\rangle$

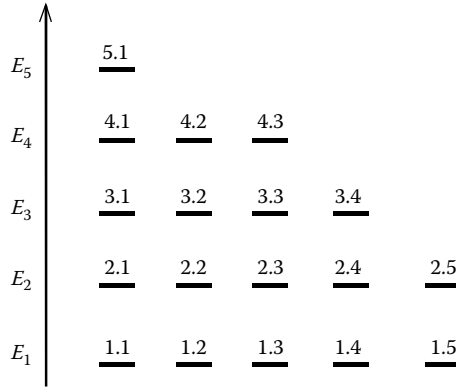


FIGURE 5.60 Example collection of states to receive the transiting particle. Note that all but level E_5 have degenerate levels.

$$\omega_{ni} = \frac{E_n - E_i}{\hbar}$$

where ω_{ni} gives the angular frequency of emitted/absorbed light when the system makes a transition from state $|i\rangle$ to state $|n\rangle$. The incident electromagnetic field has angular frequency ω . Equation 5.281 gives the probability of transition for each “final” state $|n\rangle$ and each “initial” state $|i\rangle$. In this section, we are interested in the density of final states but not in the density of initial states. We therefore take the units for Equation 5.281 as the “probability per final state.”

Equation 5.278 shows that the total probability of the electron leaving an initial state i must be related to the probability that it makes a transition into any number of final states. How can we change the formula if the final states have the same energy? What is the transition probability if some of the final states have energy E_1 , some have energy E_2 , and so on. Figure 5.60 shows the situation for a collection of final states. For conceptual purposes, the states are indexed by $n.a$ where n represents the level number in E_n and a represents the state with energy E_n . So for example, the five states 1.1 to 1.5 all have energy E_1 . Let N_n be the total number of states with the same energy E_n . For example, level E_2 has $N_2 = 5$. This notation requires a slight change to that used for Equation 5.278 since the integer n does not describe all of the states. We must include the index a as follows:

$$P_V = [P(i \rightarrow 1.1) + P(i \rightarrow 1.2) + \cdots + P(i \rightarrow N_1)] \\ + [P(i \rightarrow 2.1) + P(i \rightarrow 2.2) + \cdots + P(i \rightarrow N_2)] + \cdots \quad (5.284)$$

Transitions to final states all having the same energy must have equal probability as can be seen from Equation 5.281 (the same ω_{ni})

$$P(i \rightarrow n.a) = P(i \rightarrow n.b) \quad (5.285)$$

As a reminder, each probability on the right-hand side is the probability per final state (and per initial state). Equation 5.284 can be rewritten using Equation 5.285 to find

$$P_V = N_1 P(i \rightarrow 1) + N_2 P(i \rightarrow 2) + \cdots = \sum_n N_n P(i \rightarrow n)$$

Because the index n really refers to the energy level E_n , we can change the dummy index to E_n or to E to obtain

$$P_V = \sum_E N_E P(i \rightarrow E) \quad (5.286)$$

where again $P(i \rightarrow E)$ is the probability per single final state with energy E . Now include the small energy interval ΔE centered on the energy E . The value of ΔE is small enough to include only the states at energy E for our simple case. In the continuum, ΔE should be smaller than other relevant energy scales. Equation 5.286 can now be written as

$$P_E = \sum_E \frac{N_E}{\Delta E} P(i \rightarrow E) \Delta E \quad (5.287)$$

The quantity $N_E/\Delta E$ represents the number of final states per unit energy $g_f(E)$. For convenience, drop the subscript “f.” So, this last equation can be rewritten in the continuum limit as

$$P_E = \sum_E g(E) P(i \rightarrow E) \Delta E \Rightarrow \int dE g(E) P(i \rightarrow E) \quad (5.288)$$

Normalize out the volume of the crystal so that g in this last equation becomes the number of states per unit energy per unit volume. It should be clear that Equation 5.288 has the correct form based on the units involved.

$$P_V = \sum_E \left(\frac{\text{\#States}}{\text{Energy} \times \text{Vol}} \right) \left(\frac{\text{Prob}}{\text{State}} \right) \Delta E \quad (5.289)$$

where $P(i \rightarrow n) = P(E_i \rightarrow E_n)$ is the probability of transition (per state) and the integral must be over the energy of the final states.

Now, insert Equation 5.281 into Equation 5.289 to find

$$P_V = \int dE g(E) \left(\frac{\mu_{ni} E_0}{\hbar} \right)^2 \frac{\sin^2 \left[\frac{1}{2}(\omega_{ni} - \omega)t \right]}{(\omega_{ni} - \omega)^2}$$

where the transition frequency

$$\omega_{ni} = (E_n - E_i)/\hbar = (E - E_i)/\hbar$$

includes the energy of final states E and where ω symbolizes the angular frequency of the driving field. It is more convenient to write the integral in terms of the “transition” energy

$$E_T = E - E_i = \hbar\omega_{ni}$$

E_T , which is the energy between the initial state and final states as shown in Figure 5.61. We find

$$P_V = \int dE_T g(E_i + E_T) (\mu_{ni} E_0)^2 \frac{\sin^2 \left[\frac{1}{2\hbar}(E_T - \hbar\omega)t \right]}{(E_T - \hbar\omega)^2} \quad (5.290)$$

The quantity $\hbar\omega$ represents the energy of the electromagnetic wave inducing the transition. The dipole matrix element μ_{ni} depends on the energy of the final state E through the index n . Therefore, the dipole moment can be written as $\mu_{ni} = \mu(E)$ for fixed initial state i . In this section, we assume that the dipole matrix element to be independent of the energy of the final state. Therefore, we take $\mu_{ni} = \mu$

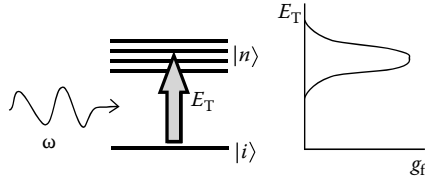


FIGURE 5.61 An electromagnetic wave induces a transition from state i to one of the final states.

to be a constant and remove it from the integral in Equation 5.290. This assumes that the final states all have the same transition characteristics; the interaction strength between the electromagnetic wave and the system (i.e., atom) remains the same for all possible final states under consideration.

Next, look at the last term in the integral in Equation 5.290

$$S = \frac{\sin^2 \left[\frac{1}{2\hbar} (E_T - \hbar\omega)t \right]}{(E_T - \hbar\omega)^2}$$

Section 5.11.6 shows that as time increases, the function S becomes sharper. For sufficiently large times t , the function S will become very sharp compared to the density of states g in Equation 5.290 as shown in Figure 5.62. The S function essentially becomes the Dirac delta function $S = \delta(E_T - \hbar\omega)$. The S function allows the density of states $g(E)$ to be removed from the integral with the substitution of $E_T = \hbar\omega$ in $g(E)$. Equation 5.290 becomes

$$P_V = (\mu E_0)^2 g(E = E_i + \hbar\omega) \int_{-\infty}^{\infty} dE_T \frac{\sin^2 \left[\frac{1}{2\hbar} (E_T - \hbar\omega)t \right]}{(E_T - \hbar\omega)^2}$$

Now evaluating the integral using a change of variable and checking the integration tables for $\int_{-\infty}^{\infty} dx (\sin^2 x)/x^2$, we find

$$P_V = (\mu E_0)^2 g(E = E_i + \hbar\omega) \frac{\pi t}{2\hbar}$$

which can also be written as

$$P_V = (\mu E_0)^2 g_f(E_f = E_i \pm \hbar\omega) \frac{\pi t}{2\hbar} \quad (5.291)$$

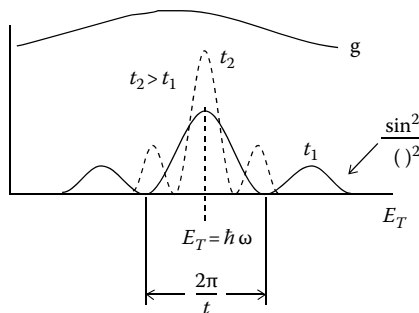


FIGURE 5.62 The S function becomes very narrow for larger times.

where E_f and E_i are the energy of the final and initial states, respectively. Equation 5.291 includes the “+” for absorption and the “−” for emission. Equation 5.288 provides the probability (per initial state per unit volume) of the system absorbing energy from the electromagnetic waves and making a transition from E_i to E_f . Notice how the probability depends on the frequency of the EM wave through the density of states. It is the energy of incident or emitted photons that connects initial states with final states.

“Fermi’s golden rule” provides the rate of stimulated emission and stimulated absorption from Equation 5.289. The rate of transition is found to be

$$R_{i \rightarrow f} = \frac{d}{dt} P_V = \frac{\pi}{2\hbar} (\mu E_0)^2 \rho_f(E_f = E_i \pm \hbar\omega) \quad (5.292)$$

Notice that the transition rate must be proportional to the optical power

$$\text{Optical power} \propto E_0^2$$

Fewer available final states at energy E_f implies a lower transition rate because of the density of states that appears in Equation 5.292. This fact has important applications for optoelectronics. For example, lowering the number of final states lowers the total rate of spontaneous emission. Tailoring the density of states, such as for photonic crystals, provides greater control over device functionality.

For a single final level, the density-of-states function must be a Dirac delta function centered at the energy E_f

$$R_{i \rightarrow f} = \frac{d}{dt} P_V = \frac{\pi}{2\hbar} (\mu E_0)^2 \delta(E_f = E_i + \hbar\omega)$$

The Dirac delta function ensures that transition process conserves energy. We could integrate this last equation over energy to find a rate of transition.

Example 5.25: An Initial Thought Experiment

Suppose a collection of atoms is excited by electrical discharge. Further suppose the light emitted by the atoms have N total states available for emission. The photon states here might refer to the modes of a Fabre-Perot cavity (similar to the modes on a violin string). Photon (or electromagnetic states) are defined by wave vector, direction and polarization). If $N=0$, then the atoms cannot emit light and the atoms either remain excited or find alternate paths to return to the ground state. Therefore, if $N=N(t)$ one would be able to modulate the emission from the collection of atoms. Interestingly, if it required very little energy to create and destroy these states, one would have a type of amplifier (of course the energy is supplied by the power source for exciting the atoms).

Example 5.26: A Second Thought Experiment Without an Immediate Solution

Suppose the collection of atoms in the previous example has very low resonance frequency (perhaps the wavelength is on the order of kilometers or more). Further suppose the collection has been placed near the center of a very long (order of kilometers or more) cylindrical tube and that the tube has a movable piston at one end so as to control the length (and the enclosed volume of the tube). The density of available states for electromagnetic (EM) emission (which would be light for shorter wavelengths) is then controlled by the position of the piston. The available states might for example correspond to wavelengths of $\lambda=L/n$ and therefore wave vectors $k=2\pi n/L$ where L represents the length of the tube at any time t and n represents an

integer. Similar to the previous example, moving the piston so that $L = L(t)$, changes the available EM states, and therefore modulates the optical emission. However, does this violate the principles of special relativity especially as concerns the speed of light? That is, the piston as the “source” of the modulation can be many kilometers (a galaxy?) away from the collection of atoms, and still has an “apparently” instantaneous effect on the emission.

5.13 DENSITY OPERATOR

The density operator and its associated equation of motion provide an alternate formulation for a quantum mechanical system. The density operator combines the probability functions of quantum and statistical mechanics into one mathematical object. The quantum mechanical part of the density operator uses the usual quantum mechanical wave function to account for the inherent particle probabilities. The statistical mechanics portion accounts for possible multiple wave functions attributable to random external influences. Typically, statistical mechanics deals with ensembles of many particles and only describes the dynamics of the system through statements of probability.

5.13.1 INTRODUCTION TO THE DENSITY OPERATOR

We usually assume we know the initial wave function of a particle or system. Consider the example wave function depicted in Figure 5.63 where the initial wave function consists of “two exactly specified basis functions with two exactly specified components.” Suppose the initial wave function can be written

$$|\psi(0)\rangle = 0.9|u_1\rangle + 0.43|u_2\rangle$$

As shown in Figure 5.64, the quantum mechanical probability of finding the electron in the first eigenstate must be

$$|\langle u_1 | \psi(0) \rangle|^2 = (0.9)^2 = 81\%$$

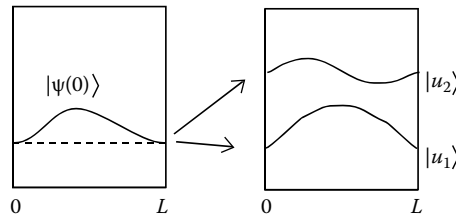


FIGURE 5.63 The initial wave function consists of exactly two basis functions.

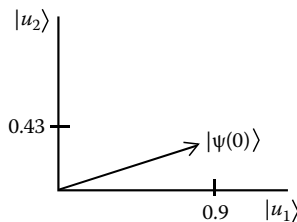


FIGURE 5.64 The components of the wave function.

Similarly, the quantum mechanical probability that the electron occupies the second eigenstate must be

$$|\langle u_2 | \psi(0) \rangle|^2 = (0.43)^2 = 19\%$$

We know the values of these probabilities with certainty since we know the decomposition of the initial wave function $|\psi(0)\rangle$ and the coefficients (0.9 and 0.43) with 100% certainty. We assume that the wave function $|\psi\rangle$ satisfies the time-dependent Schrödinger wave equation (SWE) while the basis states satisfy the time-independent SWE

$$\hat{H}|\psi\rangle = i\hbar\partial_t|\psi\rangle \quad \hat{H}|u_n\rangle = E_n|u_n\rangle$$

What if we do not exactly know the initial preparation of the system? For example, we might be working with an infinitely deep well. Suppose we try to prepare a number of identical systems. Suppose we make four such systems with parameters as close as possible to each other. Figure 5.65 shows the ensemble of systems all having the same width L . Unlike the present case with only four systems, we usually (conceptually) make an infinite number of systems to form an ensemble. Figure 5.65 shows that we were not able to prepare identical wave functions $|\psi\rangle$. Denote the wave function for system S by $|\psi_s\rangle$. Then the wave function $|\psi_s\rangle$ for each system must have different coefficients, as for example,

$$\begin{aligned} |\psi_1\rangle &= 0.98 |u_1\rangle + 0.19 |u_2\rangle \\ |\psi_2\rangle &= 0.90 |u_1\rangle + 0.43 |u_2\rangle \\ |\psi_3\rangle &= 0.95 |u_1\rangle + 0.31 |u_2\rangle \\ |\psi_4\rangle &= 0.90 |u_1\rangle + 0.43 |u_1\rangle \end{aligned} \tag{5.293}$$

The four wave functions appear in Figure 5.66. Notice how system $S=2$ and system $S=4$ both have the same wave function.

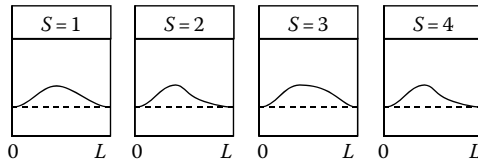


FIGURE 5.65 An ensemble of four systems.

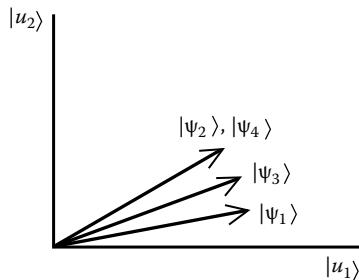


FIGURE 5.66 The different initial wave functions for the infinitely deep well.

What actual wave function $|\psi\rangle$ describes the system? Answer: An “actual” $|\psi\rangle$ does not exist; we can only talk about an average wave function. In fact, if we had prepared many such systems, we would only be able to specify the probability that the system has a certain wave function. For example, for the four systems described above, the probability of each type of wave function must be given by

$$P(S = 2) = \frac{1}{2} \quad P(S = 1) = \frac{1}{4} \quad P(S = 3) = \frac{1}{4}$$

For convenience, systems $S=2$ and $S=4$ have both been symbolized by $S=2$ since they have identical wave functions. Perhaps this would be clearer by writing

$$P\{0.90|u_1\rangle + 0.43|u_2\rangle\} = \frac{1}{2} \quad P\{0.98|u_1\rangle + 0.19|u_2\rangle\} = \frac{1}{4} \quad P\{0.95|u_1\rangle + 0.31|u_2\rangle\} = \frac{1}{4}$$

We can now represent the four systems by three vectors in our Hilbert space rather than four so long as we also account for the probability.

Now let us do something a little unusual. Suppose we try to define an “average wave function” to represent a typical system (think of the example with the four infinitely deep wells)

$$\text{Ave}\{|\psi\rangle\} = \sum_S P_S |\psi_S\rangle$$

Recall, the classical average of a quantity “ x_i ” or “ x ” can be written as $\langle x_i \rangle = \sum_i x_i P_i$ and $\langle x \rangle = \int dx x P(x)$ for the discrete and continuous cases, respectively (see Appendix D). The average wave function would represent an average system in the ensemble. We look at the entire ensemble of systems (there might be an infinite number of copies) and say that the wave function $\text{Ave}\{|\psi\rangle\}$ behaves like the average for all those systems. The wave function $\text{Ave}|\psi\rangle$ would represent the quantum mechanical stochastic processes while the probabilities P_S represent the macroscopic probabilities. No one actually uses this average wave function. The sum of the squares of the components of $\text{Ave}\{|\psi\rangle\}$ do not necessarily add to 1 since the probabilities P_i are squared (see the chapter review exercises).

Now here comes the really unusual part where we define an average probability. If we exactly know the wave function, then we can exactly calculate probabilities using the quantum mechanical probability density $\psi^*(x)\psi(x)$ (it is a little odd to be combining the words “exact” and “probability”). Now let us extend this idea of probability using our ensemble of systems. We change notation and let P_ψ be the probability of finding one of the systems to have a wave function of $|\psi\rangle$. We define an average probability density function according to

$$\text{Average}(\psi^*\psi) = \sum_\psi P_\psi (\psi^*(x)\psi(x)) \quad (5.294)$$

where P multiplies the product of wave functions in parenthesis (i.e., P is not a function of the product of wave functions).

This formula contains both the quantum mechanical probability density $\psi^*\psi$ and the macroscopic probability P_ψ . We could use the S subscripts on P_S so long as we include only one type of wave function for each S . Equation 5.294 assumes a discrete number of possible wave functions $|\psi_S\rangle$. However, the situation might arise with so many wave functions that they essentially form a continuum in Hilbert space (i.e., S must be a continuously varying parameter). In such a case, we talk about the classical probability density ρ_S which gives the probability per unit interval S of finding a particular wave function.

$$\text{Average}(\psi^*\psi) = \int dS \rho_S (\psi_S^*(x)\psi_S(x))$$

The probability p_S is similar to the density of states seen in later chapters; rather than a subscript of S , we would have a subscript of energy and units of “number of states per unit energy per unit volume.” We continue with Equation 5.294 since it contains all the essential ingredients.

Rearranging Equation 5.294, we obtain a “way to think of the average.” First switch the order of the wave function and its conjugate.

$$\text{Average}(\psi^*\psi) = \sum_{\psi} P_{\psi} \psi^*(x) \psi(x) = \sum_{\psi} P_{\psi} \psi(x) \psi^*(x)$$

Next write the wave functions in Dirac notation and factor out the basis kets $|x\rangle$

$$\text{Average}(\psi^*\psi) = \sum_{\psi} P_{\psi} \langle x | \psi \rangle \langle \psi | x \rangle = \langle x | \left\{ \sum_{\psi} P_{\psi} |\psi\rangle \langle \psi| \right\} | x \rangle$$

We define the density operator to be

$$\hat{\rho} = \sum_{\psi} P_{\psi} |\psi\rangle \langle \psi| \quad (5.295)$$

Equation 5.295 shows that the density operator represents an average of the possible projection operators. The density operator has the simultaneous attributes of the quantum through the wave functions and the macroscopic probability through P . The meaning will become clearer as we progress through the section.

Example 5.27

Find the initial density operator $\hat{\rho}(0)$ for the wave functions given in the following table. We assume four two-level atoms.

Initial Wave Function, $ \psi_S(0)\rangle$	Probability, P_S
$ \psi_1\rangle = 0.98 u_1\rangle + 0.19 u_2\rangle$	1/4
$ \psi_2\rangle = 0.90 u_1\rangle + 0.43 u_2\rangle$	1/2
$ \psi_3\rangle = 0.95 u_1\rangle + 0.31 u_2\rangle$	1/4

The initial density operator must be given by $\hat{\rho}(0) = \sum_{S=1}^3 P_S |\psi_S(0)\rangle \langle \psi_S(0)|$. Substituting the probabilities and initial wave functions, we find

$$\begin{aligned} \hat{\rho}(0) &= P_1 |\psi_1(0)\rangle \langle \psi_1(0)| + P_2 |\psi_2(0)\rangle \langle \psi_2(0)| + P_3 |\psi_3(0)\rangle \langle \psi_3(0)| \\ &= \frac{1}{4} [0.98|u_1\rangle + 0.19|u_2\rangle] [0.98\langle u_1| + 0.19\langle u_2|] \\ &\quad + \frac{1}{2} [0.90|u_1\rangle + 0.43|u_2\rangle] [0.90\langle u_1| + 0.43\langle u_2|] \\ &\quad + \frac{1}{4} [0.95|u_1\rangle + 0.31|u_2\rangle] [0.95\langle u_1| + 0.31\langle u_2|] \end{aligned}$$

Collecting terms

$$\hat{\rho}(0) = 0.86 |u_1\rangle \langle u_1| + 0.307 |u_1\rangle \langle u_2| + 0.307 |u_2\rangle \langle u_1| + 0.14 |u_2\rangle \langle u_2|$$

Example 5.28

Assume that the probability of any wave function is zero except for the particular wave function $|\psi_o\rangle$. Find the density operator in both the discrete and continuous cases.

SOLUTION

For the discrete case, the probability can be written as $P_\psi = \delta_{\psi, \psi_o}$ and the density operator becomes

$$\hat{\rho} = \sum_{\psi} P_{\psi} |\psi\rangle \langle \psi| = \sum_{\psi} \delta_{\psi, \psi_o} |\psi\rangle \langle \psi| = |\psi_o\rangle \langle \psi_o|$$

For the continuous case, the probability density can be written as $\rho_{\psi} = \delta(\psi - \psi_o)$ and the density operator becomes

$$\hat{\rho} = \int d\psi \rho_{\psi} |\psi\rangle \langle \psi| = \int d\psi \delta(\psi - \psi_o) |\psi\rangle \langle \psi| = |\psi_o\rangle \langle \psi_o|$$

5.13.2 DENSITY OPERATOR AND THE BASIS EXPANSION

The density operator can be written in the basis vector expansion. The density operator $\hat{\rho}$ has a range and domain within a single vector space. Suppose the set of basis vectors $\{|m\rangle = u_m\}$ spans the vector space of interest. People most commonly use the energy eigenfunctions as the basis set. Using the basis function expansion of an operator as described in Chapter 3, the density operator can be written as

$$\hat{\rho} = \sum_{mn} \rho_{mn} |m\rangle \langle n| \quad (5.296)$$

where $\langle n| = |n\rangle^+$. Recall that ρ_{mn} must be the matrix elements of the operator $\hat{\rho}$. We term the collection of coefficients $[\rho_{mn}]$ the “density matrix.” Apply the matrix methods from Chapter 3 to find

$$\langle a|\hat{\rho}|b\rangle = \sum_{mn} \rho_{mn} \langle a|m\rangle \langle n|b\rangle = \sum_{mn} \rho_{mn} \delta_{am} \delta_{bn} = \rho_{ab}$$

where $|a\rangle, |b\rangle$ are basis vectors. This section shows how the density operator can be expanded in a basis and provides an interpretation of the matrix elements.

The density operator provides two types of average. The first type consists of the quantum mechanical average and the second consists of the ensemble average. For the ensemble average, we imagine a large number of systems prepared as nearly the same as possible. We imagine a collection of wave functions $\{|\psi_S(t)\rangle\}$ with one for each different system S . Again, we imagine that P_S denotes the probability of finding a particular wave function $|\psi_S(t)\rangle$. Assume that all of the wave functions of the systems can be described by vector spaces spanned by the set $\{|m\rangle = u_m\}$ as shown in Figure 5.67. Assume the same basis functions for each system. Each wave function $|\psi_S(t)\rangle$ can be expanded in the complete orthonormal basis set for each system

$$|\psi_S(t)\rangle = \sum_m \beta_m^{(S)}(t) |m\rangle \quad (5.297)$$

The superscript (S) on each expansion coefficient refers to a different system. However, a single set of basis vectors applies to all of the systems S in the ensemble of systems. Therefore, if two systems

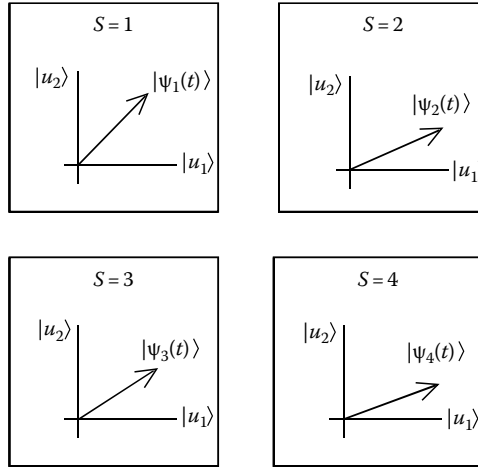


FIGURE 5.67 Four systems with the same basis functions.

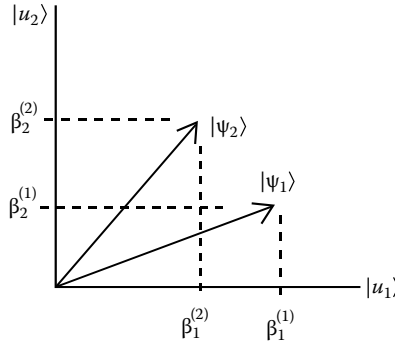


FIGURE 5.68 Two realizations of a system have different wave functions and therefore different components.

(a) and (b) have different wave functions, then the coefficients must be different $\beta_m^{(a)} \neq \beta_m^{(b)}$ (see Figure 5.68).

Using the definition of the density operator, we can write

$$\hat{\rho}(t) = \sum_S P_S |\psi_S(t)\rangle \langle \psi_S(t)| \quad (5.298)$$

Notice that the density operator in the Schrödinger picture can depend on time since the wave functions depend on time (it is also possible to have P_S depend on time, but neglect this for now). Using the definition of adjoint

$$\langle \psi_S(t) | = |\psi_S(t)\rangle^+ = \left[\sum_n \beta_n^{(S)} |n\rangle \right]^+ = \sum_n \beta_n^{(S)*} \langle n| \quad (5.299)$$

Substituting Equations 5.297 and 5.299 into Equation 5.298, we obtain

$$\hat{\rho}(t) = \sum_{mn} \sum_S P_S \beta_m^{(S)} \beta_n^{(S)*} |m\rangle \langle n|$$

Now, compare this last expression with Equation 5.296 to see that the matrix of the density operator (i.e., the density matrix) must be

$$\rho_{mn} = \langle m | \hat{\rho} | n \rangle = \sum_S P_S \beta_m^{(S)} \beta_n^{(S)*} = \langle \beta_m^{(S)} \beta_n^{(S)*} \rangle_e = \overline{\beta_m^{(S)} \beta_n^{(S)*}} \quad (5.300)$$

where the “e” subscript indicates the ensemble average.

Whereas the density “operator” $\hat{\rho}$ gives the “ensemble” average of the wave function projection operator $\overline{|\psi\rangle\langle\psi|} = \langle |\psi\rangle\langle\psi| \rangle_e$ the density “matrix” element ρ_{mn} provides the ensemble average of the wave function coefficients $\rho_{mn} = \overline{\beta_m^{(S)} \beta_n^{(S)*}} = \langle \beta_m^{(S)} \beta_n^{(S)*} \rangle_e$ (i.e., the average of the density matrix elements). The averages must be taken over all of the systems S in the ensemble.

The whole point of the density operator is to simultaneously provide two averages. We use the quantum mechanical average to find quantities such as average position, momentum, energy, or electric field using only the quantum mechanical state of a given system. The ensemble average takes into account nonquantum mechanical influences such as variation in container size or slight differences in environment that can be represented by a probability P_S . Notice in the definition of density operator

$$\hat{\rho}(t) = \sum_S P_S |\psi_S(t)\rangle \langle \psi_S(t)| \quad (5.301)$$

that if one of the systems occurs at the exclusion of all others (say $S = 1$) so that

$$\hat{\rho}(t) = |\psi_1(t)\rangle \langle \psi_1(t)| = |\psi(t)\rangle \langle \psi(t)| \quad (5.302)$$

then the density operator only provides quantum mechanical averages. In such a case, the wave functions for all the systems in the ensemble have the same form since macroscopic conditions do not differently affect any of the systems. Density operators as in Equation 5.302 without a statistical mixture will be called “pure” states. Sometimes people refer to a density operator of the form $|\psi(t)\rangle \langle \psi(t)|$ as a “state” or a “wave function” because it consists solely of the wave function $|\psi(t)\rangle$. The density operator and the wave function provide equivalent descriptions of the single quantum mechanical system and both obey the Schrödinger equation.

Now let us examine the conceptual meaning of the matrix elements $\rho_{mn} = \overline{\beta_m^{(S)} \beta_n^{(S)*}}$ in Equation 5.300. The diagonal matrix elements $\rho_{nn} = \overline{\beta_n^{(S)} \beta_n^{(S)*}} = \overline{P(n)}$ provide the average probability of finding the system in eigenstate n . In other words, even though the diagonal elements incorporate the ensemble average, we still “think” of them as $\rho_{nn} \sim |\beta_n|^2 \sim P(n)$ where $P(n)$ represents the usual quantum mechanical probability. For an ensemble of systems with different wave functions $|\psi^{(S)}\rangle$, we must average the quantum probability over the various systems.

The off-diagonal elements of the density operator appear to be similar to the probability amplitude that a particle simultaneously exists in two states. For simplicity, assume that the ensemble has only one type of wave function given by the superposition $|\psi\rangle = \sum_n \beta_n |u_n\rangle$ so that $\langle u_m | \psi \rangle = \sum_n \beta_n \langle u_m | u_n \rangle = \beta_m$. The off-diagonal elements have the form

$$\rho_{ab} = \langle u_a | \hat{\rho} | u_b \rangle = \langle u_a | \psi \rangle \langle \psi | u_b \rangle = \langle u_a | \psi \rangle \langle u_b | \psi \rangle^+ = \beta_a \beta_b^*$$

Recall that the classical probability of finding a particle in both states can be written as

$$P(a \text{ and } b) = P(a)P(b)$$

for independent events. But $P(a) = |\beta_a|^2$ and $P(b) = |\beta_b|^2$ so, combining the last several expressions provides

$$\rho_{ab} = \langle u_a | \hat{\rho} | u_b \rangle = \beta_a \beta_b^* \sim \sqrt{P(a \text{ and } b)}$$

Apparently (conceptually speaking), the off-diagonal elements of the density operator must be related to the probability of simultaneously finding the particle in both states a and b . This should remind the reader of a transition from one state to another when the particle can be quantum mechanically in both states at the same time. In fact, books on the physics of optoelectronics and light emitters/absorbers show that the off-diagonal elements can be related to the susceptibility which is related to the dipole moment and the gain or loss. That is, the off-diagonal elements describe the probability of transition between states while the diagonal elements describe the probability of finding the particle in a single state.

Example 5.29

For Example 5.27, find the density matrix.

SOLUTION

The density matrix can be written as

$$\underline{\rho} = \begin{bmatrix} 0.86 & 0.307 \\ 0.307 & 0.14 \end{bmatrix}$$

for the basis set $\{|u_1\rangle, |u_2\rangle\}$. Notice how the coefficients of the first and last term add to 1—this is not an accident. The diagonal elements of the density matrix correspond to the probability that a particle will be found in the level $|u_1\rangle, |u_2\rangle$.

Example 5.30

Find the coordinate and energy basis set representation for the density operator under the following conditions. Assume that the density operator can be written as $\hat{\rho} = |\psi\rangle\langle\psi|$. Assume also that the energy basis set can be written as $\{|u_a\rangle\}$ so that $|\psi\rangle = \sum_n \beta_n |u_n\rangle$. What is the probability of finding the particle in state $|a\rangle = |u_a\rangle$?

SOLUTION

First, the expectation of the density operator in the coordinate representation.

$$\langle x | \hat{\rho} | x \rangle = \langle x | \psi \rangle \langle \psi | x \rangle = \psi^*(x) \psi(x)$$

Second, the expectation of the density operator using a vector basis produces the probability of finding the particle in the corresponding state (i.e., diagonal matrix elements give the probability of occupying a state).

$$\langle u_a | \hat{\rho} | u_a \rangle = \langle u_a | \psi \rangle \langle \psi | u_a \rangle = \langle u_a | \psi \rangle \langle u_a | \psi \rangle^+ = |\beta_a|^2$$

Third, the probability of finding the particle in state $|a\rangle$ is

$$P(a) = |\beta_a|^2 = \langle u_a | \hat{\rho} | u_a \rangle = \rho_{aa}$$

as seen in the last equation. Therefore, the diagonal elements provide the probability of finding the electron in the corresponding state.

Example 5.31

Show that the diagonal terms of the density matrix add to 1. Assume that the wave function $|\psi^{(s)}\rangle = \sum_n \beta_n^{(s)} |n\rangle$ describes system s and the density operator has the form $\hat{\rho} = \sum_s P_s |\psi^{(s)}\rangle \langle \psi^{(s)}|$.

SOLUTION

The matrix element of the density operator can be written as

$$\rho_{aa} = \langle a | \hat{\rho} | a \rangle = \langle a | \left\{ \sum_s P_s |\psi^{(s)}\rangle \langle \psi^{(s)}| \right\} | a \rangle = \sum_s P_s \langle a | \psi^{(s)}\rangle \langle \psi^{(s)} | a \rangle = \sum_s P_s |\beta_a^{(s)}|^2$$

Now summing over the diagonal elements (i.e., equivalent to taking the trace)

$$\text{Tr}(\hat{\rho}) = \sum_a \rho_{aa} = \sum_a \sum_s P_s |\beta_a^{(s)}|^2 = \sum_s P_s \sum_a |\beta_a^{(s)}|^2 = \sum_s P_s 1 = \sum_s P_s$$

where the second to last result follows since the components for each individual wave function s must add to 1 (Figure 5.69). Finally, the sum of the probabilities P_s must add to 1 to get

$$\text{Tr}(\hat{\rho}) = \sum_s P_s = 1$$

This shows that the probability of finding the particle in any of the states must sum to 1.

5.13.3 ENSEMBLE AND QUANTUM MECHANICAL AVERAGES

For studies in solid state, the density operator most importantly provides averages of operators. We know averages of operators correspond to classically observed quantities. We will find the average of an operator has the form

$$\langle \langle \hat{O} \rangle \rangle = \text{Tr}(\hat{\rho} \hat{O}) \quad (5.303)$$

where for now the double brackets reminds us that the density operator involves two probabilities and therefore two types of average. This equation contains both the quantum mechanical and

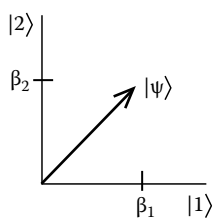


FIGURE 5.69 Wave function and components.

ensemble average. “Tr” means to take the Trace. The average of a Hermitian operator provides the expected classical value.

We define the quantum mechanical “q” and ensemble “e” averages for an operator $\hat{O}$ as follows

Quantum Mechanical	Ensemble
$\langle \hat{O} \rangle_q = \langle \psi \hat{O} \psi \rangle$	$\langle \hat{O} \rangle_e = \sum_S P_S \hat{O}_S$

where $|\psi\rangle$ denotes a typical quantum mechanical wave function. In what follows, we take the operator in the “ensemble” average to be just a number that depends on the particular system S (for example, it might be the system temperature that varies from one system to the next).

Now we will show that the ensemble and quantum mechanical average of an operator $\hat{O}$ can be calculated using $\langle \langle \hat{O} \rangle \rangle = \text{Tr}(\hat{\rho} \hat{O})$. Recall the definition of trace,

$$\text{Tr}(\hat{\rho} \hat{O}) = \sum_n \langle n | \hat{\rho} \hat{O} | n \rangle \quad (5.304)$$

Although the trace does not depend on the particular basis set, equations of motion use the energy basis $\{|n\rangle = |u_n\rangle\}$ where $\hat{H}|n\rangle = E_n|n\rangle$.

First let us find the quantum mechanical average of an operator for the specific system S starting with

$$\langle \hat{O} \rangle_q^{(S)} = \langle \psi_S | \hat{O} | \psi_S \rangle \text{ with } |\psi_S(t)\rangle = \sum_n \beta_n^{(S)}(t) |u_n\rangle \quad (5.305)$$

where as before, $|\psi_S(t)\rangle$ provides the wave function for the system S . Combining Equation 5.305 provides

$$\langle \hat{O} \rangle_q^{(S)} = \sum_n \beta_n^{*(S)} \langle u_n | \hat{O} \sum_m \beta_m^{(S)}(t) |u_m\rangle = \sum_{nm} \beta_n^{*(S)} \beta_m^{(S)} \langle u_n | \hat{O} | u_m \rangle = \sum_{nm} \beta_m^{(S)} \beta_n^{*(S)} O_{nm} \quad (5.306)$$

There is one such average for each different system S since there is a different wave function for each different system. For a given system S , this last expression gives the quantum mechanical average of the operator for that one system.

As a last step, take the ensemble average of Equation 5.306 using P_S as the probability.

$$\langle \langle \hat{O} \rangle \rangle = \left\langle \langle \hat{O} \rangle_q^{(S)} \right\rangle_e = \sum_S P_S \langle \hat{O} \rangle_q^{(S)} = \sum_S P_S \sum_{nm} \beta_m^{(S)} \beta_n^{*(S)} O_{nm}$$

Rearranging the summation and noting $\text{Tr}(\underline{\rho} \underline{O}) = \sum_{mn} \rho_{mn} O_{nm}$ provides the desired results.

$$\langle \langle \hat{O} \rangle \rangle = \sum_{mn} \left(\sum_S P_S \beta_m^{(S)} \beta_n^{*(S)} \right) O_{nm} = \sum_{mn} \overline{\beta_m^{(S)} \beta_n^{*(S)}} O_{nm} = \sum_{mn} \rho_{mn} O_{nm} = \text{Tr}(\hat{\rho} \hat{O})$$

Example 5.32

Find the average of an operator for a pure state with $\hat{\rho} = |\psi(t)\rangle \langle\psi(t)|$

SOLUTION

Equation 5.304 provides

$$\langle \hat{O} \rangle = \text{Tr}(\hat{\rho} \hat{O}) = \sum_n \langle u_n | \psi(t) \rangle \langle \psi(t) | \hat{O} | u_n \rangle = \sum_n \langle \psi(t) | \hat{O} | u_n \rangle \langle u_n | \psi(t) \rangle = \langle \psi(t) | \hat{O} | \psi(t) \rangle$$

where the first summation uses the definition of trace and the last step used the closure relation for the states $|u_n\rangle$. For the pure state, we see that the trace formula reduces to the ordinary quantum mechanical average of $\langle \hat{O} \rangle = \langle \psi(t) | \hat{O} | \psi(t) \rangle$.

Example 5.33: The Two Averages

The electron gun in a television picture tube has a filament to produce electrons and a high-voltage electrode to accelerate them toward the phosphorus screen (see top portion of Figure 5.70). Suppose the high-voltage section is slightly defective and produces small random voltage fluctuations. We therefore expect the momentum $p = \hbar k$ of the electrons to slightly vary similar to the bottom portion of Figure 5.70. Assume each individual electron is in a plane wave state $\psi^{(k)}(x, t) = \frac{1}{\sqrt{V}} e^{ikx - i\omega t}$ where the superscript “(k)” indicates the various systems rather than “(s)”. Find the average momentum.

SOLUTION

The quantum mechanical average can be found

$$\langle \psi^{(k)} | \hat{p} | \psi^{(k)} \rangle_q = \left\langle \psi^{(k)} \left| \frac{\hbar}{i} \frac{\partial}{\partial x} \right| \psi^{(k)} \right\rangle_q$$

Substituting for the wave function, we find

$$\langle \psi^{(k)} | \hat{p} | \psi^{(k)} \rangle_q = \frac{1}{V} \int_V dV e^{-ikx + i\omega t} \frac{\hbar}{i} \frac{\partial}{\partial x} e^{ikx - i\omega t} = \hbar k$$

where we assume that the wave function is normalized to the volume V . We still need to average over the various electrons (i.e., the systems or k values) leaving the electron gun. The bottom

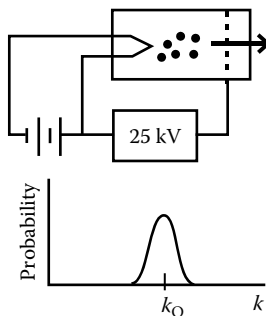


FIGURE 5.70 The electron gun (top) produces a slight variation in wave vector k (bottom).

portion of Figure 5.70 shows the k -vectors have a Gaussian distribution. Therefore, the average momentum must be $\langle \hat{p} \rangle_q = \hbar k_0$.

Example 5.34

Let $\hat{H}$ be the Hamiltonian for a two-level system with energy eigenvectors $\{|u_1\rangle, |u_2\rangle\}$ so that $\hat{H}|u_1\rangle = E_1|u_1\rangle$ and $\hat{H}|u_2\rangle = E_2|u_2\rangle$. What is the matrix of H with respect to the basis vectors $\{|u_1\rangle, |u_2\rangle\}$?

The matrix elements of $\hat{H}$ can be written as $H_{ab} = \langle u_a | \hat{H} | u_b \rangle = E_b \delta_{ab}$ which can be written as

$$\underline{H} = \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix}$$

Example 5.35

What is the ensemble-averaged energy $\langle \hat{H} \rangle \equiv \langle \langle \hat{H} \rangle \rangle$? Assume all of the information remains the same as for Examples 5.34, 5.27, and 5.29.

SOLUTION

We want to evaluate the average given by

$$\langle \hat{H} \rangle = \text{Tr}(\hat{\rho} \hat{H})$$

We can insert basis vectors as required by the trace and then insert the closure relation between the two operators. We would then end-up with the formula identical to taking the trace of the product of two matrices.

$$\text{Tr}(\hat{\rho} \hat{H}) = \text{Tr}(\underline{\rho} \underline{H}) = \text{Tr} \begin{bmatrix} 0.86 & 0.307 \\ 0.307 & 0.14 \end{bmatrix} \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} = \text{Tr} \begin{bmatrix} 0.86E_1 & 0.307E_2 \\ 0.307E_1 & 0.14E_2 \end{bmatrix}$$

Of course, in switching from operators to matrices, we have used the isomorphism between operators and matrices. Operations using the operators must be equivalent to operations using the corresponding matrices. Summing the diagonal elements provides the trace of a matrix and we find

$$\langle \hat{H} \rangle = \text{Tr}(\hat{\rho} \hat{H}) = 0.86E_1 + 0.14E_2$$

So the average differs from the eigenvalue E_1 or E_2 ! The average energy represents a combination of the energies dictated by both the quantum mechanical and ensemble probabilities.

Example 5.36

What is the probability that an electron will be found in the basis state $|u_1\rangle$? Assume all of the information remains the same as for Examples 5.35, 5.34, 5.29, and 5.27.

SOLUTION

We assume the density matrix

$$\underline{\rho} = \begin{bmatrix} 0.86 & 0.307 \\ 0.307 & 0.14 \end{bmatrix}$$

The answer is Probability of state #1 = $\langle u_1 | \hat{\rho} | u_1 \rangle = p_{11} = 0.86$. In fact, we can find the probability of the first state being occupied directly from the definition of the density operator

$$\langle 1 | \hat{\rho} | 1 \rangle = \langle 1 | \left[\sum_S P_S |\psi_S\rangle \langle \psi_S| \right] | 1 \rangle = \sum_S P_S \langle 1 | \psi_S \rangle \langle \psi_S | 1 \rangle = \sum_S P_S \beta_1^{(S)} \beta_1^{(S)*} = \overline{\beta_1 \beta_1^*}$$

5.13.4 LOSS OF COHERENCE

In some cases, the physical system introduces uncontrollable phase shifts in the various components of the wave functions. Suppose the wave functions have the form

$$|\psi^{(\phi_1, \phi_2, \dots)}\rangle = \sum_n \beta_n^{(\phi_n)} |n\rangle \quad (5.307a)$$

where the phases $(\phi_1, \phi_2, \dots)$ label the wave function and assume a continuous range of values. The components have the form

$$\beta_n^{(\phi_n)} = |\beta_n| e^{i\phi_n} \quad (5.307b)$$

Let $P_\phi(\phi_1, \phi_2, \dots) = P(\phi_1)P(\phi_2) \dots$ be the probability for $|\psi^{(\phi_1, \phi_2, \dots)}\rangle$. The density operator assumes the form

$$\hat{\rho} = \int d\phi_1 d\phi_2 \dots P(\phi_1, \phi_2, \dots) |\psi^{(\phi_1, \phi_2, \dots)}\rangle \langle \psi^{(\phi_1, \phi_2, \dots)}| \quad (5.308)$$

Now we can demonstrate the effects of the loss of coherence. One would expect the off-diagonal matrix elements to decrease as well as the probability of any transition between states. Expanding the terms in Equation 5.308 using Equation 5.307 produces

$$\hat{\rho} = \int d\phi_1 d\phi_2 \dots P(\phi_1)P(\phi_2) \dots \sum_{m,n} |\beta_m| |\beta_n| e^{i(\phi_m - \phi_n)} |m\rangle \langle n| \quad (5.309)$$

The exponential terms drop out for $m = n$. The integral over the probability density can be reduced using the property $\int d\phi_a P(\phi_a) = 1$.

$$\hat{\rho} = \sum_m |\beta_m|^2 |m\rangle \langle m| + \sum_{m \neq n} |\beta_m| |\beta_n| |m\rangle \langle n| \int d\phi_m P(\phi_m) e^{i\phi_m} \int d\phi_n P(\phi_n) e^{-i\phi_n} \quad (5.310)$$

Assume for a concrete example, a uniform distribution $P(\phi) = 1/2\pi$ on $(0, 2\pi)$. The integrals produce

$$\int_0^{2\pi} d\phi_m P(\phi_m) e^{i\phi_m} = 0$$

and the density operator in Equation 5.310 becomes diagonal

$$\hat{\rho} = \sum_m |\beta_m|^2 |m\rangle \langle m| \quad (5.311)$$

Some mechanisms produce a loss of coherence. For example, making a measurement causes the wave functions to collapse to a single state. The wave functions become $|m\rangle$ with quantum

mechanical probability $|\beta_m|^2$ so that the density operator appears as in Equation 5.311. Often the macroscopic and quantum probabilities are combined into a single number p_m and the density operator becomes

$$\hat{\rho} = \sum_m p_m |m\rangle\langle m| \quad (5.312)$$

Notice that the density matrix $\hat{\rho} = |\psi\rangle\langle\psi|$ for a pure state can always be reduced to a single entry by choosing a basis with $|\psi\rangle$ as one of the basis vectors. The mixed state in Equation 5.311 cannot be reduced from its diagonal form. Many processes cause decoherence including atomic collision/scattering processes.

Example 5.37

Suppose a system contains N independent two-level atoms (per unit volume). Each atom corresponds to one of the systems that make up the ensemble. Given the density matrix ρ_{mn} , find the number of two-level atoms in level #1 and level #2.

SOLUTION

The number of atoms in state $|a\rangle$ must be given by

$$N_a = (\text{Total number}) (\text{Prob of state } a) = N \rho_{aa} \quad (5.313)$$

Example 5.38

Suppose there are $N=5$ atoms as shown in Figure 5.71. Let the energy basis set be $\{|1\rangle = |u_1\rangle, |2\rangle = |u_2\rangle\}$. Assume that a measurement determines the number of atoms in each level. Find the density matrix based on the figure.

SOLUTION

Notice that the diagonal density-matrix elements can be calculated if we assume that the wave functions $|\psi_s\rangle$ can only be either $|u_1\rangle$ or $|u_2\rangle$. The density operator has the form

$$\hat{\rho} = \sum_{s=1}^2 P_s |\psi_s\rangle\langle\psi_s| = P_1 |u_1\rangle\langle u_1| + P_2 |u_2\rangle\langle u_2|$$

or, equivalently, the matrix must be

$$\rho_{aa} = \langle u_a | \hat{\rho} | u_a \rangle \rightarrow \underline{\rho} = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

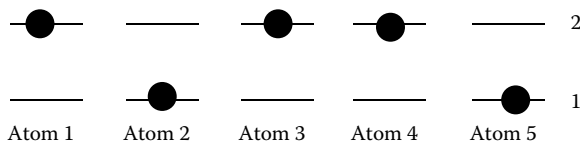


FIGURE 5.71 Ensemble of atoms in various states.

Figure 5.71 clearly shows that $\text{Prob}(1) = P_1 = 3/5$ and $\text{Prob}(2) = P_2 = 2/5$. Therefore, the probability of an electron occupying level #1 must be $\rho_{11} = 2/5$ and the probability of an electron occupying level #2 must be $\rho_{22} = 3/5$.

Example 5.39

What if we had defined the occupation number operator $\hat{n}$ to be

$$\hat{n}|1\rangle = 1|1\rangle, \quad \hat{n}|2\rangle = 2|2\rangle$$

Calculate the expectation value of $\hat{n}$ using the trace formula for the density operator.

SOLUTION

$$\langle \hat{n} \rangle = \text{Tr}(\hat{\rho}\hat{n}) = \text{Tr} \begin{bmatrix} 2/5 & 0 \\ 0 & 3/5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = \frac{8}{5}$$

This just says that the average state is somewhere between “1” and “2.” We can check this result by looking at the figure. The average state should be

$$1 \cdot \text{Prob}(1) + 2 \cdot \text{Prob}(2) = 1 \cdot \frac{2}{5} + 2 \cdot \frac{3}{5} = \frac{8}{5}$$

as found with the density matrix.

5.13.5 SOME PROPERTIES

1. If $P_\psi = 1$ so that $\hat{\rho} = |\psi\rangle\langle\psi|$ represents a pure state, then

$$\hat{\rho}\hat{\rho} = |\psi\rangle\langle\psi| |\psi\rangle\langle\psi| = |\psi\rangle\langle\psi| = \hat{\rho}$$

In this case, the operator $\hat{\rho}$ satisfies the property required for idempotent operators. The only possible eigenvalues for this particular density operator are 0 and 1.

$$\hat{\rho}|v\rangle = v|v\rangle \rightarrow \hat{\rho}\hat{\rho}|v\rangle = v|v\rangle \rightarrow v^2|v\rangle = v|v\rangle \rightarrow v^2 = v \rightarrow v = 0, 1$$

2. All density operators are Hermitian

$$\hat{\rho}^+ = \left\{ \sum_{\psi} P_{\psi} |\psi\rangle\langle\psi| \right\}^+ = \sum_{\psi} P_{\psi} \{ |\psi\rangle\langle\psi| \}^+ = \sum_{\psi} P_{\psi} |\psi\rangle\langle\psi| = \hat{\rho}$$

since the probability must be a real number.

3. Diagonal elements of the density matrix give the probability that a system will be found in a specific eigenstate. The diagonal elements take into account both ensemble and quantum mechanical probabilities. Let $\{|a\rangle\}$ be a complete set of states (basis states) and let the wave function for each system have the form

$$|\psi(t)\rangle = \sum_a \beta_a^{(\psi)}(t) |a\rangle$$

The diagonal elements of the density matrix must be

$$\begin{aligned}\rho_{aa} &= \langle a | \hat{\rho} | a \rangle = \langle a | \left\{ \sum_{\psi} P_{\psi} |\psi\rangle \langle \psi| \right\} | a \rangle = \sum_{\psi} P_{\psi} \langle a | \psi \rangle \langle \psi | a \rangle = \sum_{\psi} P_{\psi} \beta_a^{*(\psi)} \beta_a^{(\psi)} = \overline{|\beta_a|^2} \\ &= \overline{\text{Prob}(a)}\end{aligned}$$

4. The sum of the diagonal elements must be unity.

$$\text{Tr}(\hat{\rho}) = \sum_n \rho_{nn} = 1$$

since the matrix diagonal contains all of the system probabilities.

5.14 INTRODUCTION TO MULTIPARTICLE SYSTEMS

The quantum mechanics must include a description of multiple particles—the many-body problem. The multiparticle system plays a dominant role in the statistical mechanics where the distribution functions make it possible to determine the average behavior of the system. These distribution functions can be derived from elementary considerations on the behavior of the constituent particles and any identifiable distinctions between them. The quantum mechanics of the multiparticle system establish the foundations for the statistical mechanics.

The material presented in this section prepares the way for the second quantization. We develop the basis states appropriate for systems of multiple “bosons” and “fermions”; these basis states consist of a direct product of single-particle states. The theory leads naturally to the Fock state describing the distribution of an exact number of particles in the available states of the system. The section shows how the direct product states lead to the Fock states. The section primarily focuses on identical (i.e., indistinguishable) particles.

5.14.1 INTRODUCTION

Multiple particles share a direct product Hilbert space with each particle occupying its own space. Suppose particle # i can occupy the single-particle basis states $|\psi_m^{(i)}\rangle = |m\rangle_i$ (think of an infinitely deep quantum well with levels E_m). Each particle occupies its own Hilbert space. N independent particles therefore share the product space with basis set

$$\{|a\rangle_1 |b\rangle_2 \dots |c\rangle_N\} \quad (5.314a)$$

where $a, b, c, \dots$ represent the energy levels of the individual spaces, and the subscript represents the particular particle and hence the particular space. For example, a specific basis state for a two-particle system might look similar to the cartoon in Figure 5.72. Of utmost importance, the notation will later be changed for the multiparticle state whereby the position represents the state (such as

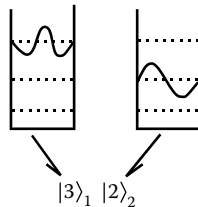


FIGURE 5.72 The basis vector for two independent particles in separate single-particle Hilbert spaces.

described by wave vector or energy) and the integer in the ket represents the number of particles in that state. However, this extension comes later in this section and the next section.

As discussed in the linear algebra, the general vector in the product space has the form

$$|\psi\rangle = \sum_{a,b,c,\dots} \beta_{a,b,c,\dots} |a\rangle_1 |b\rangle_2 |c\rangle_3 \dots \tag{5.314b}$$

This entangle state cannot be reduced in general. Independent electrons obey equations of motion strictly confined to there own spaces (no interaction terms). Without any previous interaction, Equation 5.314b can be reduced to

$$|\psi\rangle = \sum_a \beta_a |a\rangle_1 \sum_b \beta_b |b\rangle_2 \sum_c \beta_c |c\rangle_3 \dots \tag{5.315}$$

For the two-particle system for example, each abstract point in one space has a second space attached to it similar to the left-hand side of Figure 5.73. Or we might picture the spaces as adjacent to each other as indicated by the right-hand side of Figure 5.73. The pictures appear very similar to the order of the basis vectors in Equation 5.314a for a two-particle system.

Specializing to indistinguishable particles increases the symmetry of the system and thereby allows mathematical expressions such as Equation 5.314b to be reduced in the sense of finding relations between relevant coefficients β . The interchange of two indistinguishable particles (Figure 5.74) cannot affect the Hamiltonian of a system based on symmetry—the interchange of identical particles does not ultimately change anything. We will therefore see that the permutation operator and the Hamiltonian must commute. Hence, the basis functions for the multiparticle system must be simultaneous eigenfunctions of the Hamiltonian and the permutation operator.

For the full direct product space (Equation 5.314b), an interesting (and essential) subdivision occurs when dealing with fermions and bosons. The study begins by delineating the distinction between fermions and bosons. The study of angular momentum indicates the fermions have half-integral spin whereas the bosons have integral spin. A further classification concerns how the wave function for the multiparticle system transforms when two of the constituent particles are interchanged. Under the transformation, fermion wave functions are multiplied by a -1 , whereas the bosons wave functions are multiplied by $+1$. Notice that the interchange of identical particles does

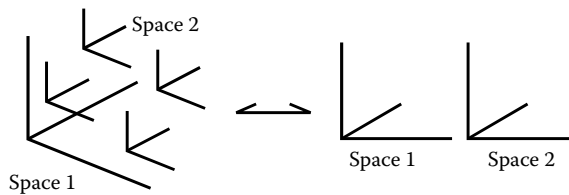


FIGURE 5.73 Two independent spaces.

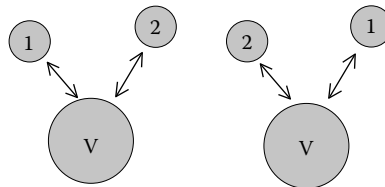


FIGURE 5.74 Interchanging identical particles cannot alter the Hamiltonian.

not have any effect on the Hamiltonian but does change the phase of the wave function. However, the change of phase does not have any effect on the probability.

The fermion and bosons occupy distinct types of states, which essentially divides the product space into two. The fermions occupy the odd-symmetry states and bosons occupy the even-symmetry states. We will see how the multiparticle fermion wave functions can be summarized using the so-called Slater determinant. These concepts prepare the way for the second quantization discussed in the next section.

Example 5.40

Consider a system of two electrons, with each capable of occupying only two states. Find the electron states.

SOLUTION

The state

$$\frac{1}{\sqrt{2}}(|1\rangle_1|2\rangle_2 - |2\rangle_1|1\rangle_2) \quad (5.316)$$

is the only one that can be formed from the vectors $\{|a\rangle_1 | b\rangle_2: a, b = 1, 2\}$ having the correct symmetry property. Notice that the symmetry is manifested by switching the subscripts as these represent the particle number. The symmetric linear combination

$$\frac{1}{\sqrt{2}}(|1\rangle_1|2\rangle_2 + |2\rangle_1|1\rangle_2)$$

does not describe fermions since it has even exchange symmetry. The other two odd combinations produce zero.

$$\frac{1}{\sqrt{2}}(|1\rangle_1|1\rangle_2 - |1\rangle_2|1\rangle_1) = 0 = \frac{1}{\sqrt{2}}(|2\rangle_1|2\rangle_2 - |2\rangle_2|2\rangle_1)$$

since the order of the kets is unimportant so long as the subscripts have been placed. Notice that the acceptable state can be written as the “Slater determinant”

$$\frac{1}{\sqrt{2}}(|1\rangle_1|2\rangle_2 - |2\rangle_1|1\rangle_2) = \frac{1}{\sqrt{2}} \begin{vmatrix} |1\rangle_1 & |2\rangle_1 \\ |1\rangle_2 & |2\rangle_2 \end{vmatrix}$$

5.14.2 PERMUTATION OPERATOR

The permutation operator interchanges two particles within a system. We first define the coordinate space function for a system with N particles and then define the permutation operator.

For the multiparticle system, a function of the form $f(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$ indicates two things. First, the “position” in the parenthesis indicates the particle number. Second, the vector $\vec{r}_a$ indicates the position of a specific particle. For example, $f(\vec{r}_2, \vec{r}_1, \dots, \vec{r}_N)$ indicates that particle # 1 has position $\vec{r}_2$ and particle #2 has position $\vec{r}_1$. For quantum theory, we do not usually think of the particle as definitely located at a specific point (except in the case of the delta-function type wave function as will be seen in more detail in the next section). Instead, the notation $|\psi(\vec{r}_2, \vec{r}_1, \dots, \vec{r}_N)|^2$ refers to the probability density that particle 1 is at $\vec{r}_2$ and particle 2 is at $\vec{r}_1$, etc. Because the particles cannot be distinguished, this must be the same as the probability density $|\psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)|^2$ that particle 2 is at $\vec{r}_2$ and particle 1 is at $\vec{r}_1$, etc.

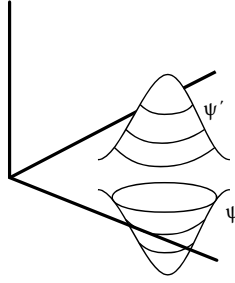


FIGURE 5.75 The effect of interchanging two fermion particles.

$$|\psi(\vec{r}_2, \vec{r}_1, \dots, \vec{r}_N)|^2 = |\psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)|^2 \quad (5.317a)$$

We would surmise that the two wave functions differ by at most a phase factor

$$\psi(\vec{r}_2, \vec{r}_1, \dots, \vec{r}_N) = e^{i\varphi} \psi(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) \quad (5.317b)$$

We will find the phase factor $e^{i\varphi}$ is $+1$ for bosons and -1 for fermions. Figure 5.75 shows the relation for a 2-D coordinate system with fermions.

Next we define the permutation operator. The symbol $\hat{P}(a, b, c \dots) = \hat{P}_{a,b,c,\dots} = \hat{P}_{1 \leftarrow a, 2 \leftarrow b, 3 \leftarrow c, \dots}$ means to replace particle #1 with particle # a , replace particle #2 with particle # b , and so on. Such an interchange means to switch the spatial coordinates of the particles. The set of all possible permutations forms a group and therefore every permutation must have an inverse. The inverse of the operator $\hat{P}_{a \leftarrow a, b \leftarrow b, c \leftarrow c, \dots}$ must be $\hat{P}_{a \leftarrow a, b \leftarrow b, c \leftarrow c, \dots}$. The permutation operator $\hat{P}_{i,j}$ for two particles produces new functions

$$\hat{P}_{i,j} \psi(\dots, \vec{r}_i, \dots, \vec{r}_j, \dots) = \psi(\dots, \vec{r}_j, \dots, \vec{r}_i, \dots) \quad (5.318a)$$

where the permutation operator switches the spatial coordinates which thereby defines the meaning of the interchange.

To see the effect of the permutation on the coordinate kets, consider the following, where the resolution of unity has been inserted.

$$\begin{aligned} \langle x_1, x_2 | \hat{P}_{1,2} | \psi \rangle &= \langle x_1, x_2 | \int dx_a dx_b \hat{P}_{1,2} | x_a, x_b \rangle \langle x_a, x_b | \psi \rangle \\ &= \int dx_a dx_b \langle x_1, x_2 | x_b, x_a \rangle \langle x_a, x_b | \psi \rangle \\ &= \int dx_a dx_b \delta(x_1 - x_b) \delta(x_2 - x_a) \psi(x_a, x_b) \\ &= \langle x_2, x_1 | \psi \rangle \end{aligned}$$

from which one can conclude for arbitrary ψ , that

$$\langle x_1, x_2 | \hat{P}_{1,2} = \langle x_2, x_1 | \quad \text{or equivalently} \quad \hat{P}_{1,2}^+ | x_1, x_2 \rangle = | x_2, x_1 \rangle \quad (5.318b)$$

The argument can then be extended to an arbitrary number of particles as follows.

Equation 5.318a can be written as

$$\langle \dots, \vec{r}_i, \dots, \vec{r}_j, \dots | \hat{P}_{i,j} = \langle \dots, \vec{r}_j, \dots, \vec{r}_i, \dots | \quad \text{or} \quad \hat{P}_{i,j}^+ | \dots, \vec{r}_i, \dots, \vec{r}_j, \dots \rangle = | \dots, \vec{r}_j, \dots, \vec{r}_i, \dots \rangle \quad (5.318c)$$

The permutation operator must be unitary (does not change the length of ψ).

$$\langle \vec{r}_i, \vec{r}_j | \hat{P}_{i,j} \hat{P}_{i,j}^+ | \vec{r}_i, \vec{r}_j \rangle = \langle \vec{r}_j, \vec{r}_i | \vec{r}_j, \vec{r}_i \rangle = \delta(\vec{r}_j - \vec{r}_j) \delta(\vec{r}_i - \vec{r}_i) = \langle \vec{r}_i, \vec{r}_j | \vec{r}_i, \vec{r}_j \rangle = \langle \vec{r}_i, \vec{r}_j | \hat{1} | \vec{r}_i, \vec{r}_j \rangle$$

Since we assume arbitrary coordinates $|\vec{r}_i, \vec{r}_j\rangle$, we conclude

$$\hat{P}_{i,j} \hat{P}_{i,j}^+ = \hat{1} \quad (5.319)$$

so that the operator must be unitary.

The interchange operator $\hat{P}_{i,j}$ must be Hermitian as well since if we apply it twice to a coordinate function we find

$$\hat{P}_{i,j} \hat{P}_{i,j} \psi(\vec{r}_i, \vec{r}_j) = \hat{P}_{i,j} \psi(\vec{r}_j, \vec{r}_i) = \psi(\vec{r}_i, \vec{r}_j)$$

For arbitrary ψ , we conclude $\hat{P}_{i,j} \hat{P}_{i,j} = \hat{1}$ and therefore $\hat{P}_{i,j}^{-1} = \hat{P}_{i,j}^+ = \hat{P}_{i,j}$. Then Equation 5.318b can also be written as $\hat{P}_{1,2} |x_1, x_2\rangle = |x_2, x_1\rangle$.

The interchange operator $\hat{P}_{i,j}$ can be seen to commute with any operator symmetrical under the interchange of coordinates. An operator is symmetric under the interchange of any two coordinates when

$$\hat{A}(\vec{r}_j, \vec{r}_i) = \hat{A}(\vec{r}_i, \vec{r}_j)$$

We can show that a symmetric operator always commutes with the interchange operator.

$$\hat{P}_{ij} \hat{A}(\vec{r}_i, \vec{r}_j) \psi(\vec{r}_i, \vec{r}_j) = \hat{A}(\vec{r}_j, \vec{r}_i) \psi(\vec{r}_j, \vec{r}_i) = \hat{A}(\vec{r}_j, \vec{r}_i) \hat{P}_{ij} \psi(\vec{r}_i, \vec{r}_j)$$

but for an arbitrary function ψ we have

$$\hat{P}_{ij} \hat{A}(\vec{r}_i, \vec{r}_j) = \hat{A}(\vec{r}_j, \vec{r}_i) \hat{P}_{ij}$$

Therefore for symmetric $\hat{A}$ we have

$$[\hat{A}(\vec{r}_i, \vec{r}_j), \hat{P}_{ij}] = 0 \quad (5.320a)$$

In particular, Equation 5.320a implies that the Hamiltonian commutes with the interchange operator for a system of identical particles

$$[\hat{H}, \hat{P}_{ij}] = 0 \quad (5.320b)$$

and therefore have simultaneous eigenvectors.

5.14.3 SIMULTANEOUS EIGENVECTORS OF THE HAMILTONIAN AND THE INTERCHANGE OPERATOR

We now show the eigenvalues of $\hat{P}$. Let $|\psi\rangle$ be an eigenfunction of the interchange operator $\hat{P}$. Suppose $\hat{P}|\psi\rangle = c|\psi\rangle$ then $\hat{P}^2|\psi\rangle = c^2|\psi\rangle$. However, we already know that $\hat{P}^2 = \hat{1}$ since $\hat{P}$ is both unitary and Hermitian. We conclude $c^2 = 1$. Therefore, the two possible eigenvalues must be $c = \pm 1$.

The introductory section in the present section discusses the symmetry of the Hamiltonian. We know that it must be symmetric under the interchange of two identical particles. The Hamiltonian and the interchange operators commute.

$$[\hat{H}(\vec{r}_i, \vec{r}_j), \hat{P}_{ij}] = 0$$

Therefore the Hamiltonian and interchange operators have simultaneous eigenfunctions. The bosons correspond to the +1 eigenvalues while the fermions correspond to the -1 ones. One can surmise these assignments using the creation and annihilation operators to be introduced in the next section. Basically, the boson creation/annihilation operators commute for distinct states which means creating identical particles in one combination must be the same as creating the same particles in a distinct combination. However, the fermion creation/annihilation operators anticommute for distinct states (as a result of the Pauli exclusion principle) which means creating particles in one combination must introduce a negative sign when compared with the interchanged combination.

Bosons correspond to the +1 eigenvalues. For N noninteracting bosons, the full Hamiltonian is $\hat{H} = \hat{H}_1 + \hat{H}_2 + \dots + \hat{H}_N$. The solutions to the time-independent Schrödinger

$$\hat{H}\psi(1, 2, \dots, N) = E\psi(1, 2, \dots, N) \quad (5.321a)$$

equation can be written as

$$|\psi(1, 2, \dots, N)\rangle = |a\rangle_1 |b\rangle_2 \dots |n\rangle_N \quad (5.321b)$$

and

$$E = E_a + E_b + \dots + E_n \quad (5.321c)$$

where the possible one-particles states are $\{|1\rangle, |2\rangle, \dots, |n\rangle, \dots\}$. Interchanging say particles 1 and 2 produces the same energy but a different wave function $|b\rangle_1 |a\rangle_2 \dots |n\rangle_N$ (exchange degeneracy). Any linear combination of the different wave functions (with the same number of states $a, b, \dots$) have the same total energy. The +1 eigenvalue of the permutation operator requires the properly normalized wave function to be

$$|\psi(1, 2, \dots, N)\rangle = \sqrt{\frac{N_a! N_b! \dots N_n!}{N!}} \{(|a\rangle_1 |b\rangle_2 \dots |n\rangle_N) + (|b\rangle_1 |a\rangle_2 \dots |n\rangle_N) + \dots\} \quad (5.322a)$$

where N_n represents the number of times the state $|n\rangle$ occurs. This state is symmetric under interchange of any two particles. Sometimes Equation 5.322a is written as

$$|\psi(1, 2, \dots, N)\rangle = \sqrt{\frac{N_a! N_b! \dots N_n!}{N!}} \sum_P P(|a\rangle_1 |b\rangle_2 \dots |n\rangle_N) \quad (5.322b)$$

where P represents all “different” permutations of the states (see Example 5.41).

Now consider a system of fermions. Suppose we have N fermions capable of occupying the states $\{u_1, u_2, u_3, \dots\} = \{|a\rangle, |b\rangle, \dots\}$. We need to take the antisymmetric combination of Equation 5.321b. The correctly normalized wave function is

$$|\psi(1, 2, \dots, N)\rangle = \sqrt{\frac{1}{N!}} \{+ (|a\rangle_1 |b\rangle_2 \dots |n\rangle_N) - (|b\rangle_1 |a\rangle_2 \dots |n\rangle_N) + \dots\} \quad (5.323a)$$

where the normalization comes from that in Equation 5.322 by noting that there can be at most one fermion per state and $1! = 1$ and $0! = 1$. Equation 5.323a can be written as

$$|\psi(1, 2, \dots, N)\rangle = \frac{1}{\sqrt{N!}} \sum_P (-1)^P (|a\rangle_1 |b\rangle_2 \dots |n\rangle_N) \quad (5.323b)$$

We can also write these last two equations as a Slater determinant.

$$|\psi(1, 2, \dots, N)\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} |a\rangle_1 & |a\rangle_2 & \dots & |a\rangle_N \\ |b\rangle_1 & |b\rangle_2 & \dots & |b\rangle_N \\ \vdots & \vdots & \ddots & \vdots \\ |n\rangle_1 & |n\rangle_2 & \dots & |n\rangle_N \end{vmatrix} \quad (5.323c)$$

5.14.4 INTRODUCTION TO FOCK STATES

In the previous notation, the ket $|3\rangle|1\rangle$ refers to particle #1 in state #3 and particle #2 in state #1. Often (especially in the theory of second quantization), the alternate notation of Fock states proves more convenient. Each “position” in the Fock ket

$$|n_1, n_2, \dots\rangle \quad (5.324)$$

refers to a different state with $n_1, n_2, \dots$ representing the number of particles in each state. We can think of the position as a type of receptacle to store particles as suggested by the buckets in Figure 5.76. The states might be degenerate in energy. For the example in the figure, the k_1 and k_2 refer to wave vectors for plane waves. They might have the same magnitude but refer to different directions of propagation. The states include the spin degree of freedom. Any number of bosons can occupy a boson state but only 0 or 1 fermion can occupy the fermion state. The state $|0\rangle \equiv |0, 0, \dots\rangle$ is the vacuum state without any particles. Also refer to the next section for a slightly different discussion on the various states. Books on the physics of optoelectronics and quantum optics discuss the states for photons (as bosons).

The Fock states satisfy the orthonormality condition

$$\langle m_1 m_2 \dots | n_1 n_2 \dots \rangle = \delta_{m_1 n_1} \delta_{m_2 n_2} \dots \quad \text{or equivalently} \quad \langle \{m_i\} | \{n_i\} \rangle = \delta_{\{m_i\}\{n_i\}} \quad (5.325)$$

The Fock states can be expressed in terms of the product states given in the previous section. Assume the one-particle states are $\{\phi_1, \phi_2, \phi_3, \dots\}$. For bosons, each state can accept an arbitrary number of particles. According to the prescription given in Equation 5.322

$$|n_1, n_2, \dots\rangle = \sqrt{\frac{n_1! n_2! \dots}{N!}} \sum_P P \left(\underbrace{|\phi_1\rangle_1 \dots |\phi_1\rangle_{n_1}}_{n_1} \underbrace{|\phi_2\rangle_{n_1+1} \dots |\phi_2\rangle_{n_1+n_2} \dots}_{n_2} \dots \right) \quad (5.326a)$$

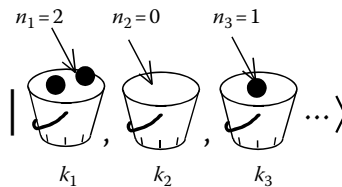


FIGURE 5.76 Example of two particles in momentum state k_1 and one particle in state k_3 .

where

$|\phi_i\rangle$ represents the one electrons states

P produces only different combinations (see Example 5.41)

On the other hand, only one fermion can occupy a given state. The Fock state for fermions has the form

$$|\psi(1, 2, \dots, N)\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} |\phi_1\rangle_1 & |\phi_1\rangle_2 & \cdots & |\phi_1\rangle_N \\ |\phi_2\rangle_1 & |\phi_2\rangle_2 & \cdots & |\phi_2\rangle_N \\ \vdots & \vdots & & \vdots \\ |\phi_N\rangle_1 & |\phi_N\rangle_2 & \cdots & |\phi_N\rangle_N \end{vmatrix} \quad (5.326b)$$

Example 5.41

Consider two bosons. Write $|1, 0, 2, 0, 0, \dots\rangle$ in terms of the one-electron wave functions.

SOLUTION

The Fock ket

$$|1, 0, 2, 0, 0, \dots\rangle = \sqrt{\frac{1!2!}{3!}} \sum_P P \left(\underbrace{|\phi_1\rangle_1 \cdots |\phi_1\rangle_{n_1}}_1 \underbrace{|\phi_3\rangle_{n_1+1} \cdots |\phi_3\rangle_{n_1+n_2}}_2 \cdots \right)$$

reduces to

$$|1, 0, 2, 0, 0, \dots\rangle = \frac{1}{\sqrt{3}} \sum_P P(|\phi_1\rangle_1 |\phi_3\rangle_2 |\phi_3\rangle_3)$$

The summation can be expanded to

$$|1, 0, 2, 0, 0, \dots\rangle = \frac{1}{\sqrt{3}} \{ |\phi_1\rangle_1 |\phi_3\rangle_2 |\phi_3\rangle_3 + |\phi_1\rangle_2 |\phi_3\rangle_1 |\phi_3\rangle_3 + |\phi_1\rangle_3 |\phi_3\rangle_2 |\phi_3\rangle_1 \}$$

Notice we did not include both $|\phi_1\rangle_3 |\phi_3\rangle_2 |\phi_3\rangle_1$ and $|\phi_1\rangle_3 |\phi_3\rangle_1 |\phi_3\rangle_2$ since they give the same result. If we include all six terms then the correct normalization would need to be $1/\sqrt{6}$ instead of $1/\sqrt{3}$.

Example 5.42

Write the Fermion Fock state $|1, 0, 1, 0, 0, \dots\rangle$ in terms of the single-particle states.

SOLUTION

$$|1, 0, 1, 0, 0, \dots\rangle = \frac{1}{\sqrt{2}} \{ |\phi_1\rangle_1 |\phi_3\rangle_2 - |\phi_3\rangle_1 |\phi_1\rangle_2 \}.$$

5.14.5 ORIGIN OF FOCK STATES

Assume a system of N particles. At this point, we do not care whether they are fermions or bosons. The particles have wave functions that depend on the coordinates x_k and the time. Assume that the Hamiltonian has the form

$$\hat{H} = \sum_k \hat{T}(x_k) + \frac{1}{2} \sum_{\substack{k,j \\ k \neq j}} \hat{V}(x_k, x_j) \quad (5.327)$$

where the kinetic energy $\hat{T}$ might have the form $\hat{T}_k = \frac{\hat{p}_k^2}{2m} = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x_k^2}$ and the potential term might have the form of Coulomb interaction

$$\hat{V}(x_k, x_j) \sim \frac{1}{|x_k - x_j|}$$

The summation over the potential terms does not include $j=k$ since that term is a self-interaction term and the potential would be infinite. The factor of $1/2$ occurs since $\hat{V}(x_k, x_j) = \hat{V}(x_j, x_k)$ and we do not want to include the same term twice.

The general wave function has the form

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{E_1, E_2, \dots, E_N} C(E_1, E_2, \dots, E_N, t) \phi_{E_1}(x_1) \phi_{E_2}(x_2) \dots \phi_{E_N}(x_N) \quad (5.328)$$

and solves the many-body Schrödinger equation

$$\hat{H}\psi = i\hbar \frac{\partial}{\partial t} \psi \quad (5.329)$$

The basis set $\{\phi_E(x)\}$ consists of single-body wave functions that account for the boundary conditions and the set $\{E\}$ consists of the corresponding energy eigenvalues. Notice, as usual, the basis set is independent of time. The subscripts on x and E in Equation 5.328 refer to the particle number. For example, an infinitely deep well has energy eigenstates that are sines or cosines with energy eigenvalues given by E_n as discussed in previous sections. We should include all of the quantum numbers in the summation (such as energy, angular momentum, etc). The reader should keep in mind that the position in the arguments of $\psi(\dots, x_i, \dots, x_j, \dots)$ or in $C(E_1, E_2, \dots, E_N, t)$ refers to a particular particle and not necessarily the x_i . In principle, the set of wave functions should be superscripted with an “(i)” to indicate the particle number so that the general wave function would read

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{E_1, E_2, \dots, E_N} C(E_1, E_2, \dots, E_N, t) \phi_{E_1}^{(1)}(x_1) \phi_{E_2}^{(2)}(x_2) \dots \phi_{E_N}^{(N)}(x_N)$$

since each particle “(i)” occupies its own Hilbert space spanned by the set $\{\phi_{E_i}^{(i)}\}$ where E_i takes on the range of eigenvalues. However, the i on the E_i consistently indicates the Hilbert space number.

We start with the observation that bosons and fermions obey different symmetry properties when two particles are interchanged; i.e., the position coordinates of the particles are interchanged. We require

$$\psi(\dots, x_i, \dots, x_j, \dots) = \pm \psi(\dots, x_j, \dots, x_i, \dots) \quad (5.330)$$

where

“+” refers to bosons

“−” refers to fermions

It can be shown that interchanging the particle coordinates in Equation 5.330 is equivalent to interchanging the energy labels in C according to

$$C(\dots, E_i, \dots, E_j, \dots, t) = \pm C(\dots, E_j, \dots, E_i, \dots, t) \quad (5.331)$$

where

“+” refers to bosons

“−” refers to fermions

To see this using only a two-particle system, start with Equation 5.328 and substitute Equation 5.330 for ψ on both sides to obtain

$$\sum_{E_1, E_2} C(E_1, E_2, t) \phi_{E_1}(x_1) \phi_{E_2}(x_2) = \pm \sum_{E_1, E_2} C(E_1, E_2, t) \phi_{E_1}(x_2) \phi_{E_2}(x_1)$$

On the right-hand side, interchange the dummy indices E_1, E_2 to obtain

$$\sum_{E_1, E_2} C(E_1, E_2, t) u_{E_1}(x_1) u_{E_2}(x_2) = \pm \sum_{E_1, E_2} C(E_2, E_1, t) u_{E_2}(x_2) u_{E_1}(x_1)$$

Compare both sides to obtain the results in Equation 5.331.

5.14.5.1 Bosons

Now use the symmetry of the coefficients to show the origin of the Fock state for bosons. First redefine the coefficients as follows. The energy basis sets for all N Hilbert spaces (i.e., N particles) correspond to the same set of eigenvalues. Here, one might imagine the range $\{E_i\} = \{1, 2, 3, \dots\}$ for every space i . For convenience, move the lowest values of the energies E_i to the left in the coefficients $C(E_1, E_2, \dots, t)$ (which can be accomplished by using the symmetry property in Equation 5.331). Let n_1 be the number of particles with energy “1” and so on. Then we would be able to write

$$C(E_1, E_2, \dots, t) = C\left(\underset{\leftarrow n_1 \rightarrow}{E_a, E_b, \dots}, \underset{\leftarrow n_2 \rightarrow}{E_c, E_d, \dots}, E_e, \dots\right)$$

Define a new coefficient $\bar{C}$ with an argument that has positions corresponding to energy rather than particle.

$$C(E_1, E_2, \dots, t) = \bar{C}(n_1, n_2, \dots, n_\infty, t)$$

where obviously

$$N = \sum_{i=1}^{n_\infty} n_i$$

represents the total number of particles.

Now we can rewrite the general wave function in Equation 5.328 as

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{n_1, n_2, \dots, n_\infty} \sum_{\substack{E_1, E_2, \dots, E_N \\ (n_1, n_2, \dots, n_\infty)}} \bar{C}(n_1, n_2, \dots, n_\infty, t) \phi_{E_1}(x_1) \phi_{E_2}(x_2) \dots \phi_{E_N}(x_N) \quad (5.332)$$

where the notation “ $(n_1, n_2, \dots, n_\infty)$ ” at the bottom of the second summation symbol means to hold the number of particles $n_1, n_2, \dots$ constant while performing the summation. The following examples show the meaning of the restricted summation and indicates that the summations in the previous equations are just an alternate method of adding over all energies.

Example 5.43

Suppose that there are three particles and five energy states

$$\{E_i: i = 1, 2, 3\} = \{1, 2, 3, 4, 5\}$$

then, for example, the coefficient $C(5, 4, 4)$ can be written

$$\begin{aligned} C(E_1 = 5, E_2 = 4, E_3 = 4) &= C(5, 4, 4) = C(4, 4, 5) \\ &= \bar{C}(n_1 = 0, n_2 = 0, n_3 = 0, n_4 = 2, n_5 = 1) = \bar{C}(0, 0, 0, 2, 1) \end{aligned}$$

Example 5.44

Consider the case of three particles and five energy levels. Assume the restriction that $n_1 = 2$ and $n_2 = 1$ and $n_i = 0$ for $i = 3, 4, 5$. The allowed configurations are

$$\begin{array}{lll} E_1 = 1 & E_2 = 1 & E_3 = 2 \\ E_1 = 1 & E_2 = 2 & E_3 = 1 \\ E_1 = 2 & E_2 = 1 & E_3 = 1 \end{array}$$

Therefore the restricted summation can be evaluated

$$\sum_{\substack{E_1, E_2, \dots, E_N \\ (n_1, n_2, \dots, n_\infty)}} C(E_1, E_2, E_3) = C(1, 1, 2) + C(1, 2, 1) + C(2, 1, 1) = 3 \quad C(1, 1, 2) = 3 \bar{C}(2, 1, 0, 0, 0)$$

The restricted summation adds over all the energy while keeping a constant number of particles with a particular energy.

The Fock states come from Equation 5.332, by defining new expansion coefficients

$$\beta(n_1, n_2, \dots, n_\infty, t) = \left(\frac{N!}{n_1! n_2! \dots n_\infty!} \right)^{1/2} \bar{C}(n_1, n_2, \dots, n_\infty, t) \quad (5.333)$$

and an alternate set of basis vectors according to the prescription

$$\phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N) = \left(\frac{n_1! n_2! \dots n_\infty!}{N!} \right)^{1/2} \sum_{\substack{E_1, E_2, \dots, E_N \\ (n_1, n_2, \dots, n_\infty)}} u_{E_1}(x_1) u_{E_2}(x_2) \dots u_{E_N}(x_N) \quad (5.334)$$

The new basis vector $\phi_{n_1, n_2, \dots, n_\infty}$ is the Fock state

$$|n_1, n_2, \dots, n_\infty\rangle = \left(\frac{n_1! n_2! \dots n_\infty!}{N!} \right)^{1/2} \sum_{\substack{E_1, E_2, \dots, E_N \\ (n_1, n_2, \dots, n_\infty)}} |u_{E_1}\rangle |u_{E_2}\rangle \dots |u_{E_N}\rangle$$

projected into coordinate space. Each Fock state for different n_i is a different basis vector as seen in the previous section. The general wave function now has the form

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{n_1, n_2, \dots, n_\infty} \beta(n_1, n_2, \dots, n_\infty, t) \phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N) \quad (5.335)$$

The Fock states are correctly normalized since

$$\langle \phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N) | \phi_{m_1, m_2, \dots, m_\infty}(x_1, x_2, \dots, x_N) \rangle = \delta_{n_1 m_1} \delta_{n_2 m_2} \dots$$

5.14.5.2 Fermions

It is possible to use the same reasoning for the fermion case. The antisymmetry of the wave function under interchange of coordinates in Equations 5.330 and 5.331

$$\psi(\dots, x_i, \dots, x_j, \dots) = -\psi(\dots, x_j, \dots, x_i, \dots) \quad (5.336)$$

$$C(\dots, E_i, \dots, E_j, \dots, t) = -C(\dots, E_j, \dots, E_i, \dots, t) \quad (5.337)$$

The fermion Fock states come from Equation 5.332

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{n_1, n_2, \dots, n_\infty} \sum_{\substack{E_1, E_2, \dots, E_N \\ (n_1, n_2, \dots, n_\infty)}} \bar{C}(n_1, n_2, \dots, n_\infty, t) u_{E_1}(x_1) u_{E_2}(x_2) \dots u_{E_N}(x_N)$$

by defining new expansion coefficients

$$\beta(n_1, n_2, \dots, n_\infty, t) = \left(\frac{N!}{n_1! n_2! \dots n_\infty!} \right)^{1/2} \bar{C}(n_1, n_2, \dots, n_\infty, t) \quad (5.338)$$

and an alternate set of basis vectors using the determinant

$$\phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N) = \left(\frac{n_1! n_2! \dots n_\infty!}{N!} \right)^{1/2} \begin{vmatrix} u_{E_1}(x_1) & \dots & u_{E_1}(x_N) \\ \vdots & & \vdots \\ u_{E_N}(x_1) & \dots & u_{E_N}(x_N) \end{vmatrix} \quad (5.339)$$

The last equation is the Fock state $|n_1, n_2, \dots, n_\infty\rangle$ projected into coordinate space. Each Fock state for different n_i is a different basis vector as seen in the previous section. The general wave function now has the form

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{n_1, n_2, \dots, n_\infty} \beta(n_1, n_2, \dots, n_\infty, t) \phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N) \quad (5.340)$$

The Fock states can be seen to be correctly normalize

$$\langle \phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N) | \phi_{m_1, m_2, \dots, m_\infty}(x_1, x_2, \dots, x_N) \rangle = \delta_{n_1 m_1} \delta_{n_2 m_2} \dots$$

by actually calculating the inner product.

5.15 INTRODUCTION TO SECOND QUANTIZATION

Quantization refers to the transition that occurs when describing a physical system by quantum mechanics rather than classical mechanics. The first quantization chapter converts the Hamiltonian and dynamical variables into operators and uses wave function to describe the characteristics of particles. Often but not always, the formalism applies to single particles. The second quantization converts the wave function into an operator. We must still find the energy basis set from the Schrödinger wave equation (SWE). Now however, the amplitudes of the wave functions become

operators. Essentially, the second quantization blends the particle-wave duality into equations that exhibit both particle and wave characteristics (see the Parker book on the Physics of Optoelectronics).

The second quantization generally applies to systems with many particles and seldom to those consisting of a single particle. The many-particle theory is required by the special theory of relativity although here, we will not make explicit use of Lorentz invariance. We use the second quantization as a conceptual simplification for understanding complex systems. Often, the second quantization and its applications fall under the subject of quantum field theory. The formalism provides the backbone of many modern theories of the solid state and condensed matter as well, and perhaps more commonly, for studies of elementary particles, and the physics of optoelectronics in the area of quantum optics and quantum electrodynamics.

5.15.1 FIELD COMMUTATORS

The present section starts with the results for the classical Lagrangian and Hamiltonian and shows the plausibility of the commutation relations for the fields. We start with bosons but stipulate similar results for fermions which use the anticommutator.

The wave functions ψ, ψ^* become operators $\hat{\psi}, \hat{\psi}^+$ in the quantum field theory. We will find the commutators below. The following sections will show that the operators $\hat{\psi}, \hat{\psi}^+$ destroy and create a particle at a specific point in space. The Lagrangian becomes

$$\mathcal{L} = \hat{\psi}^+ \left(i\hbar \partial_t + \frac{\hbar^2}{2m} \nabla^2 - V \right) \hat{\psi} \quad (5.341a)$$

which produces a Lagrangian-derived Hamiltonian density for the field.

$$\mathcal{H} = \hat{\pi} \hat{\psi} - \mathcal{L} = \hat{\psi}^+ \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \hat{\psi} \quad (5.341b)$$

The Lagrangian-derived Hamiltonian becomes

$$H = \int d^3x \mathcal{H} = \int d^3x \hat{\psi}^+ \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \hat{\psi} = \langle \hat{\psi}^+ | \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) | \hat{\psi} \rangle \quad (5.341c)$$

The Lagrangian-derived Hamilton looks more like an average. The canonical momentum becomes

$$\hat{\pi} = \frac{\partial \mathcal{L}}{\partial \dot{\hat{\psi}}} = i\hbar \hat{\psi}^+ \quad (5.341d)$$

The classical field theory (Section 4.6) shows how to divide space into cells so that the generalized coordinates have the form

$$q_i = \frac{1}{\Delta V_i} \int_{V_i} dV_i \psi(x_i) \xrightarrow{\Delta V_i \rightarrow 0} \psi(x_i) \quad (5.342a)$$

and the generalized momenta have the form

$$p_j = \Delta V_j \pi_j \quad (5.342b)$$

Classically we might think of p_j as the momentum associated with the volume ΔV_j . The classical dynamical variables satisfy the commutator in the form of the Poisson brackets.

$$[A, B] = \sum_i \left[\frac{\partial A}{\partial q_i} \frac{\partial B}{\partial p_i} - \frac{\partial B}{\partial q_i} \frac{\partial A}{\partial p_i} \right] \quad (5.343a)$$

The coordinates and momenta satisfy

$$[q_i, p_j] = \delta_{ij}, \quad [q_i, q_j] = 0 = [p_i, p_j] \quad (5.343b)$$

We assume that the quantum counterparts of the classical variables satisfy similar relations although without the derivatives.

$$i\hbar\delta_{ij} = [\hat{q}_i, \hat{p}_j] = \Delta V_j [\hat{\psi}(x_i), \hat{\pi}(x_j)] = \Delta V_j [\hat{\psi}(x_i), i\hbar\hat{\psi}^+(x_j)]$$

which gives

$$\delta_{ij} = \Delta V_j [\hat{\psi}(x_i), \hat{\psi}^+(x_j)]$$

We will take $\Delta V_j \rightarrow 0$. The last expression can be written as

$$\delta_{ij} = \int_{\Delta V_j} dV_j [\hat{\psi}(x_i), \hat{\psi}^+(x_j)]$$

We can satisfy this integral for “bosons” by requiring the commutator to be a Dirac delta function.

$$[\hat{\psi}(x_i), \hat{\psi}^+(x_j)] = \delta(x_i - x_j) \quad (5.344a)$$

Similarly, the remaining Equation 5.343b provides (at equal times)

$$[\hat{\psi}(x_i), \hat{\psi}(x_j)] = 0 = [\hat{\psi}^+(x_i), \hat{\psi}^+(x_j)] \quad (5.344b)$$

where $[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A}$

We assume “fermion” fields satisfy anticommutation relations of the form

$$\{\hat{\psi}(\vec{r}), \hat{\psi}^+(\vec{r}')\} = \delta(\vec{r} - \vec{r}'), \quad \{\hat{\psi}(\vec{r}), \hat{\psi}(\vec{r}')\} = 0 = \{\hat{\psi}^+(\vec{r}), \hat{\psi}^+(\vec{r}')\} \quad (5.345)$$

where $\{A, B\} = AB + BA$. The difference between the commutators and anticommutators produces different statistics for the two types of particles. The anticommutators allow only one fermion per state.

5.15.2 CREATION AND ANNIHILATION OPERATORS

We start with the energy basis set found from the Sturm-Liouville problem for the time-independent SWE (first quantization).

$$\hat{H}|\phi_E\rangle = E|\phi_E\rangle \quad (5.346a)$$

A solution to the time-dependent SWE takes the form

$$|\Psi(t)\rangle = \sum_E b_E(t) |\phi_E\rangle \quad (5.346b)$$

We interpret this as saying that a particle with wave function Ψ partly exists in each state ϕ at the same time. If all the b 's except one are zero, then

$$|\Psi(t)\rangle = b_E(t) |\phi_E\rangle \quad (5.346c)$$

This relation says that the single particle exists in the single state at time t . Here we use b for bosons.

In the second quantized theory, the boson wave function Ψ becomes an operator $\hat{\Psi}$ by changing the amplitudes b_E into operators $\hat{b}_E$.

$$|\hat{\Psi}(t)\rangle = \sum_E \hat{b}_E(t) |\phi_E\rangle \quad \text{or} \quad \hat{\Psi}(\vec{r}, t) = \sum_E \hat{b}_E(t) \phi_E(\vec{r}) \quad (5.347a)$$

Notice that we still use the same basis states $|\phi_E\rangle$ and must still solve the one-particle Schrödinger equation. The inverse relation can be written as

$$\hat{b}_E(t) = \int dV \phi_E^*(\vec{r}) \hat{\Psi}(\vec{r}, t) \quad (5.347b)$$

There are two types of Hilbert spaces involved with, for example, Equation 5.347a. The basis states ϕ_E live in one space. These states ϕ_E correspond to the typical basis states as eigenfunctions of the Schrödinger equation and studied in Chapters 2 and 3. The second Hilbert space corresponds to that on which the $\hat{b}_E$ operate. The $\hat{b}$ operators essentially provide the amplitude for a particular mode such as ϕ_E to be in the superposition. Perhaps if one considers ϕ_E to be a plane wave, it becomes more obvious as to the role of $\hat{b}_E$ if it were a number as a Fourier coefficient. However as an operator, $\hat{b}_E$ requires a Hilbert space on which to operate to provide the amplitudes. This “amplitude” space provides the characteristics of the actual wave function. For example, Fock states describe the number of particles with an exact value of energy (in this case) whereas for a second example, coherent states consist of a summation of Fock states and correspond to the closest quantum analog to a classically visualized localized particle. However, particles in Fock states are highly nonclassical.

Commutation relations apply to the amplitude operators whereas the modes $\phi(\vec{r})$ are treated as c-numbers. For this reason, the second form of Equation 5.347a is often preferable since it emphasizes the c-number aspect of ϕ_E . The commutation relations below will point out the distinctions between the two equations in Equation 5.347. Because elements of two distinct Hilbert spaces occur in Equation 5.347, two types of averages will be required. In addition, to find the amplitudes in the expansion Ψ , we will need to specify a Hilbert space for the amplitude-operators. Studies of quantum electromagnetic fields show examples for the Fock, coherent, and squeezed states. For now, we will use the Fock states. Often times, the set of basis states consists of plane waves and Equation 5.347 becomes

$$\hat{\Psi}(\vec{r}, t) = \sum_{\vec{k}} \hat{b}_{\vec{k}}(t) \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{V}} \quad (5.348)$$

where V represents the normalization volume. This has the form of a Fourier integral.

We demonstrate the commutation relations for the amplitude operators before continuing with the interpretation of the operators in Equation 5.347. The field operators $\hat{\Psi}, \hat{\Psi}^+$ satisfy commutation relations given in the previous section.

$$[\hat{\Psi}(\vec{r}), \hat{\Psi}^+(\vec{r}')] = \delta(\vec{r} - \vec{r}') \quad (5.349a)$$

Substituting Equation 5.347 provides

$$\left[\sum_m \hat{b}_m(t) \phi_m(\vec{r}), \sum_n \hat{b}_n^+(t) \phi_n^*(\vec{r}') \right] = \delta(\vec{r} - \vec{r}') \quad (5.349b)$$

Evaluating the commutator provides

$$\sum_{m,n} [\hat{b}_m(t), \hat{b}_n^+(t)] \phi_m(\vec{r}) \phi_n^*(\vec{r}') = \delta(\vec{r} - \vec{r}') \quad (5.349c)$$

Notice how the mode functions $\phi_m(\vec{r}) \phi_n^*(\vec{r}')$ have been freely commuted. Now use the Dirac notation for the mode functions and the delta function to find

$$\sum_{m,n} [\hat{b}_m(t), \hat{b}_n^+(t)] |\phi_m\rangle \langle \phi_n| = \sum_m |\phi_m\rangle \langle \phi_m| \quad (5.349d)$$

Notice how the amplitude operators remain in the commutator but the $|\phi_m\rangle \langle \phi_n|$ maintain the same order as that in the original commutator of Equation 5.349b. This points out the need for caution when using the first form of Equation 5.347a. Because $|\phi_m\rangle \langle \phi_n|$ forms a basis for linear operators on the function space, comparing both sides of Equation 5.349d requires

$$[\hat{b}_m(t), \hat{b}_n^+(t)] = \delta_{mn} \quad (5.350a)$$

Similar results can be demonstrated for the other equal-time commutation relations

$$[\hat{b}_m(t), \hat{b}_n(t)] = 0 = [\hat{b}_m^+(t), \hat{b}_n^+(t)] \quad (5.350b)$$

The fermion fields lead to anticommutation relations for the fermion amplitude operators $\hat{f}_m, \hat{f}_n^+$ where

$$|\hat{\Psi}(t)\rangle = \sum_E \hat{f}_E(t) |\phi_E\rangle \quad \text{or} \quad \hat{\Psi}(\vec{r}, t) = \sum_E \hat{f}_E(t) \phi_E(\vec{r}) \quad (5.351)$$

The anticommutation relations are

$$\{\hat{f}_m(t), \hat{f}_n^+(t)\} = \delta_{mn} \quad \{\hat{f}_m(t), \hat{f}_n(t)\} = 0 = \{\hat{f}_m^+(t), \hat{f}_n^+(t)\} \quad (5.352)$$

where $\{A, B\} = AB + BA$. Commuting the operators requires a multiplying minus sign.

5.15.3 INTRODUCTION TO FOCK STATES

The quantum fields and the Hamiltonian can be expressed by a traveling wave Fourier expansion with creation $\hat{b}^+$ and annihilation $\hat{b}$ operators for the Fourier amplitudes that satisfy commutation

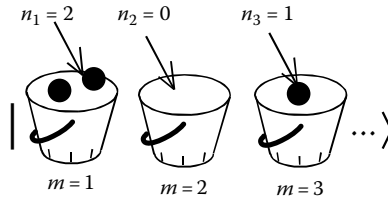


FIGURE 5.77 The Fock state describes the number of particles in the modes or states of the system. The diagram represents the ket $|2, 0, 1, \dots\rangle$.

relations. These operators act on “amplitude space.” The “Fock states” provide the first example of a basis set for this Hilbert space. The Fock states specify the exact number of particles in a given basic state of the system; the standard deviation of the number must be zero. The ket representing the Fock state consists of “place holders” for the number of particles in a given mode (basic state) $|n_1, n_2, \dots\rangle$. Figure 5.77 shows buckets that can hold particles where the mode numbers label the buckets. The figure shows the system has two particles (for example) in the $m = 1$ mode, none in the $m = 2$ mode, and so on. In proper notation, the state would be represented by the ket $|2, 0, 1, \dots\rangle$. The vacuum state, denoted by $|0, 0, 0, \dots\rangle = |0\rangle$ represents a system without any particles in any of the modes. The Fock state lives in a direct product space so that it can be written as $|n_1, n_2, \dots\rangle = |n_1\rangle|n_2\rangle \cdots$ with each ket representing a single mode. The Fock vectors for a system with only one mode characterized by the wave vector k produce only one position in the ket. For example, $|n_1\rangle$ represents n_1 particles in the mode k_1 and $|0\rangle$ represents the single mode vacuum state. The most important point of the Fock state is that it is an eigenstate of the number operator as we will see.

We should include the spin in the description of the Fock state. Assume the spin along the z -direction is represent by $s = 1$ (up) and $s = 2$ (down). Each index $\vec{k}$ value must be augmented with the polarization directions as indicated in Figure 5.78. Therefore, one can create a particle with a given wave vector and given spin.

For bosons, which are characterized by integer spin $(0, 1, 2, \dots)$, any number of them can occupy a mode. For a given set of modes, each Fock state is a basis vector for the amplitude space. The set

$$\{|n_1, n_2, n_3, \dots\rangle\}$$

represents the complete set of basis vectors where each n_i can range up to an infinite number of boson particles in the system. The orthonormality relation can be written as

$$\langle n_1, n_2, \dots | m_1, m_2, \dots \rangle = \delta_{n_1 m_1} \delta_{n_2 m_2} \cdots \quad (5.353)$$

and the closure relation as

$$\sum_{n_1, n_2, \dots=0}^{\infty} |n_1, n_2, \dots\rangle \langle n_1, n_2, \dots| = \hat{1} \quad (5.354)$$

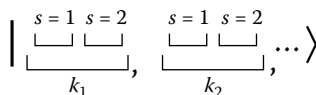


FIGURE 5.78 The modes must include polarization.

A general vector in the Hilbert space must have the form

$$|\xi\rangle = \sum_{n_1, n_2, \dots=0}^{\infty} \beta_{n_1, n_2, \dots} |n_1, n_2, \dots\rangle \quad (5.355)$$

where quantum mechanical wave functions must be normalized to unity as usual. The component $\beta_{n_1, n_2, \dots} = \langle n_1, n_2, \dots | \xi \rangle$ represents the probability amplitude of finding n_1 particles in state 1, n_2 particles in state 2, etc. when the system has wave function $|\xi\rangle$.

Fock states can also be constructed for fermions with half-integral spin, such as electrons with spin $1/2$; however, the Pauli exclusion principle limits the number per mode to at most 1. These properties originate in the commutation relations for the creation and annihilation operators.

5.15.4 INTERPRETATION OF THE AMPLITUDE AND FIELD OPERATORS

With considerable algebra, we can show the boson operators $\hat{b}_n^+$, $\hat{b}_n$ create and destroy a single boson in the state ϕ_n , respectively. The fermion operators $\hat{f}_n^+$, $\hat{f}_n$ create and destroy a single fermion in the state ϕ_n . We must ensure that the states have the proper symmetry properties as required for multiple-particle systems. The creation and destruction properties are best shown using the Fock states.

$$\hat{b}_i |n_1, n_2, \dots, n_i, \dots\rangle = \sqrt{n_i} |n_1, n_2, \dots, n_i - 1, \dots\rangle \quad (5.356a)$$

$$\hat{b}_i |n_1, n_2, \dots, n_i = 0, \dots\rangle = 0 \quad (5.356b)$$

$$\hat{b}_i^+ |n_1, n_2, \dots, n_i, \dots\rangle = \sqrt{n_i + 1} |n_1, n_2, \dots, n_i + 1, \dots\rangle \quad (5.356c)$$

Recall that the vacuum state $|0\rangle = |0, 0, 0, \dots\rangle$ does not have any particles at all. The fermion creation and annihilation operators do the same thing except the anticommutation relations permit no more than one particle per state.

$$0 = \{\hat{f}_i^+, \hat{f}_i^+\} |0\rangle = 2\hat{f}_i^+ \hat{f}_i^+ |0\rangle$$

Clearly, the general boson state can be constructed

$$|n_1, n_2, \dots\rangle = \left[\frac{(\hat{b}_1^+)^{n_1}}{\sqrt{n_1!}} \frac{(\hat{b}_2^+)^{n_2}}{\sqrt{n_2!}} \dots \right] |0\rangle \quad (5.357)$$

with a similar expression for fermions.

The creation and annihilation operators act differently than the ladder operators used for the simple harmonic oscillator. The ladder operators $\hat{a}^+$, $\hat{a}$ might at best be considered to “move” a particle from one state to another; however, they primarily map one basis vector to another one. The creation and annihilation operators must be used in the combination $\hat{b}_{n+1}^+ \hat{b}_n$ in order to move a particle from one state to another.

The number operator $\hat{n}_i = \hat{b}_i^+ \hat{b}_i$ gives the number of particles in state $|i\rangle$. For example, Equation 5.356a yield

$$\hat{n}_2 |n_1, n_2, n_3, \dots\rangle = \hat{b}_2^+ \hat{b}_2 |n_1, n_2, n_3, \dots\rangle = \sqrt{n_2} \hat{b}_2^+ |n_1, n_2 - 1, n_3, \dots\rangle = n_2 |n_1, n_2, n_3, \dots\rangle$$

The total-number operator must be

$$\hat{N} = \sum_i \hat{n}_i \quad (5.358)$$

An alternate expression for the number operator comes from the field operators and the definition for the particle-density operator

$$\rho(\vec{r}) = \hat{\Psi}^+(\vec{r}, t) \hat{\Psi}(\vec{r}, t) \quad (5.359)$$

For example, using Equation 5.347a and integrating over space provides

$$\begin{aligned} \int dV \rho(\vec{r}) &= \int dV \hat{\Psi}^+(\vec{r}, t) \hat{\Psi}(\vec{r}, t) = \sum_{m,n} \int dV \hat{b}_m^+(t) \hat{b}_n(t) \phi_m^*(\vec{r}) \phi_n(\vec{r}) \\ &= \sum_n \hat{b}_n^+(t) \hat{b}_n(t) = \hat{N} \quad \text{since } \langle \phi_m | \phi_n \rangle = \delta_{mn} \end{aligned}$$

The field operators $\hat{\Psi}^+(\vec{r}, t), \hat{\Psi}(\vec{r}, t)$ can be interpreted as creating, annihilating a particle at point $\vec{r}$ and time t . To see this, consider the state for a single particle localized to a single point defined by the coordinate ket

$$|\vec{r}, t\rangle = \Psi(\vec{r}, t) |0\rangle \quad (5.360)$$

We can show that the state $|\vec{r}, t\rangle$ is a eigenstate of the number operator with the eigenvalue of 1 for an infinitesimally small volume; therefore, the particle must be at point $\vec{r}$. Let $\Delta V \rightarrow 0$ be a small volume. Define the number operator

$$\hat{N}_{\Delta V} = \int_{\Delta V} dV' \hat{\Psi}^+(\vec{r}', t) \hat{\Psi}(\vec{r}', t) \quad (5.361)$$

to be the expected number of boson particles expected in the volume ΔV . First note

$$[N_{\Delta V}, \Psi^+(r, t)] = \begin{cases} \Psi^+(r, t) & \vec{r} \in \Delta V \\ 0 & \vec{r} \notin \Delta V \end{cases} \quad (5.362a)$$

Apply this last equation for the case $\Delta V \rightarrow 0$ to see

$$N_{\Delta V} |\vec{r}, t\rangle = N_{\Delta V} \Psi^+(r, t) |0\rangle = \Psi^+(r, t) N_{\Delta V} |0\rangle + \begin{cases} \Psi^+(r, t) |0\rangle & \vec{r} \in \Delta V \\ 0 & \vec{r} \notin \Delta V \end{cases} \quad (5.362b)$$

The vacuum does not have any particles so that $N_{\Delta V} |0\rangle = 0$. Therefore, substituting $|\vec{r}, t\rangle$ shows

$$N_{\Delta V} |\vec{r}, t\rangle = \begin{cases} |\vec{r}, t\rangle & \vec{r} \in \Delta V \rightarrow 0 \\ 0 & \vec{r} \notin \Delta V \rightarrow 0 \end{cases} \quad (5.362c)$$

So that Ψ^+ must create one particle at $\vec{r}, t$ and nowhere else.

5.15.5 FERMION–BOSON OCCUPATION AND INTERCHANGE SYMMETRY

The previous section used the fact that any number of bosons can occupy a state whereas only one fermion could do so. The restrictions on the occupation number can be related to the phase of the wave functions upon interchange of identical particles.

For fermions, the Pauli exclusion principle does not allow more than one particle per state. Then we must have $f_n^+ f_n^+ |0\rangle = 0$ so that $2f_n^+ f_n^+ |0\rangle = 0$ and then $\{f_n^+, f_n^+\} |0\rangle = 0$ and finally

$\{f_n^+, f_n^+\} = 0$. This also shows that the state $|0, \dots, 2, 0, \dots\rangle$ does not exist for fermions. Assuming the anticommutator relations hold in general $\{f_m^+, f_n^+\} = 0$, the effects of interchange can be seen. Consider a two-particle, two-state system for simplicity. Include a superscript of “p1” or “p2” to indicate the particle number so that $|1, 1\rangle$ becomes $|1^{(p1)}, 1^{(p2)}\rangle$. The designation of “p1” and “p2” has no real meaning since the two fermions live in an entangled product state and share equally in all aspects of the state; that is, they live as essentially one entity in the product state. The previous section provides

$$|1^{(p1)}, 1^{(p2)}\rangle \sim \frac{1}{\sqrt{2}} \left\{ \phi_{E_a}^{(p1)}(x_1) \phi_{E_b}^{(p2)}(x_2) - \phi_{E_a}^{(p2)}(x_1) \phi_{E_b}^{(p1)}(x_2) \right\}$$

Now consider the effect of using a permutation operator $\hat{P}$ to interchange the labels

$$\hat{P}|1^{(p1)}, 1^{(p2)}\rangle = \hat{P}f_1^{(p1)+}f_2^{(p2)+}|0, 0\rangle = -\hat{P}f_2^{(p2)+}f_1^{(p1)+}|0, 0\rangle = -f_2^{(p1)+}f_1^{(p2)+}|0, 0\rangle = -|1^{(p2)}, 1^{(p1)}\rangle$$

where the transition from the second to third term used the anticommutator, whereas the transition from the third term to the fourth used an interchange of labels “p1” and “p2.” Usually people leave off the labels and credit the anticommutator for the minus sign.

For bosons, the Pauli exclusion principle does not apply and more than one boson can occupy any given state. In this case, for a simple two-particle system, we have from a previous section

$$|1^{(p1)}, 1^{(p2)}\rangle \sim \frac{1}{\sqrt{2}} \left\{ \phi_{E_a}^{(p1)}(x_1) \phi_{E_b}^{(p2)}(x_2) + \phi_{E_a}^{(p2)}(x_1) \phi_{E_b}^{(p1)}(x_2) \right\}$$

Now consider the effect of using a permutation operator $\hat{P}$ to interchange the labels

$$\hat{P}|1^{(p1)}, 1^{(p2)}\rangle = \hat{P}f_1^{(p1)+}f_2^{(p2)+}|0, 0\rangle = \hat{P}f_2^{(p2)+}f_1^{(p1)+}|0, 0\rangle = f_2^{(p1)+}f_1^{(p2)+}|0, 0\rangle = |1^{(p2)}, 1^{(p1)}\rangle$$

where the transition from the second to third term used the commutator, whereas the transition from the third term to the fourth used an interchange of labels “p1” and “p2.” As with fermions, people usually leave off the labels and credit the commutator for the plus sign.

5.15.6 SECOND QUANTIZED OPERATORS

The Schrödinger operators $\hat{O}_s$ must be converted into those for the second quantization $\hat{O}_q$. Averages in the second quantization appear as $\langle \text{Fock} | \hat{O}_q | \text{Fock} \rangle$, for example. However, the transition from the Schrödinger wave functions to the field operators (Equation 5.357a) involves two types of Hilbert space. Therefore, we expect the averages in the second quantized theory to already implement an average over the c-number functions. This behavior can be seen from the Hamiltonian.

The second quantized form of the Hamiltonian can be found from Equations 5.341c and 5.347a.

$$\hat{H}_q = \langle \hat{\psi}^+ | H_s | \hat{\psi} \rangle = \sum_{m,n} \hat{b}_m^+ \hat{b}_n \langle \phi_m | H_s | \phi_n \rangle \quad (5.363a)$$

Notice how the mode average $\langle \phi_m | H_s | \phi_n \rangle$ appears in the formula for the second quantized operator. Applying the amplitude states to Equation 5.363a then gives the average in the second quantization. For ϕ_n an eigenfunction of H_s , Equation 5.363a reduces to the form

$$\hat{H}_q = \sum_n E_n \hat{b}_n^+ \hat{b}_n \quad (5.363b)$$

This last formula says to multiply the energy E_n of a state by the number of particles in the state $\hat{N}_n = \hat{b}_n^\dagger \hat{b}_n$, and then add them all together. For the Fock state $|n_1, n_2, \dots\rangle$, for example, we find

$$\hat{H}_q |n_1, n_2, \dots\rangle = \sum_i E_i \hat{b}_i^\dagger \hat{b}_i |n_1, n_2, \dots\rangle = \left(\sum_i n_i E_i \right) |n_1, n_2, \dots\rangle$$

A similar form holds for fermions.

The second quantization simplifies some calculations. For example, suppose an electron can be in either of two states and can make transitions by absorbing or emitting a photon. Then we can immediately write down the interaction Hamiltonian as

$$\hat{H}_{\text{int}} = c_a \hat{f}_2^\dagger \hat{f}_1 \hat{a}_\lambda + c_e \hat{f}_1^\dagger \hat{f}_2 \hat{a}_\lambda^\dagger \quad (5.364)$$

where $\hat{H}_{\text{int}}$ is Hermitian so long as $c_e^* = c_a$. The first term destroys a photon using the photon annihilation operator $\hat{a}_\lambda$ and uses the absorbed energy to promote an electron from state 1 to state 2. The second term transitions an electron from state 2 to 1 and conserves energy by emitting a photon by creating one using $\hat{a}_\lambda^\dagger$.

A prescription similar to Equation 5.363a works for changing the general Schrödinger operator into the second quantized form. Operators are classified according to the number of coordinates (i.e., the number of particles involved). A one-body operator $\hat{O}_1$ such as the kinetic energy or momentum of a single particle, follows the rule

$$\hat{O}_{1q} = \sum_{m,n} \hat{b}_m^\dagger \langle \phi_m | \hat{O}_{1S} | \phi_n \rangle \hat{b}_n \quad (5.365)$$

A two-body operator $\hat{O}_2$ such as the potential energy $V(\vec{r}_1, \vec{r}_2)$ takes the form

$$\hat{V} = \frac{1}{2} \iint dV dV' \hat{\Psi}^\dagger(\vec{r}) \hat{\Psi}^\dagger(\vec{r}') V(\vec{r}, \vec{r}') \hat{\Psi}(\vec{r}') \hat{\Psi}(\vec{r}) = \frac{1}{2} \sum \hat{b}_\alpha^\dagger \hat{b}_\beta^\dagger V_{\alpha\beta\gamma\delta} \hat{b}_\delta \hat{b}_\gamma \quad (5.366a)$$

where

$$V_{\alpha\beta\gamma\delta} = \langle \phi_\alpha \phi_\beta | V | \phi_\gamma \phi_\delta \rangle = \iint dV dV' \phi_\alpha(\vec{r}) \phi_\beta(\vec{r}') V \phi_\gamma(\vec{r}') \phi_\delta(\vec{r}) \quad (5.366b)$$

and the $\frac{1}{2}$ occurs to prevent double counting terms in the summation. Especially notice the order of the indices in Equation 5.366a. For bosons, the order does not matter, but for fermions, the anticommutation relations will insert a negative sign.

The current density can be written in second quantized form by converting the standard quantum mechanical expression into one with the field operators.

$$\hat{J} = \frac{q\hbar}{2mi} \left[\hat{\Psi}^\dagger(\vec{r}) \nabla \hat{\Psi}(\vec{r}) - \nabla \hat{\Psi}^\dagger(\vec{r}) \cdot \hat{\Psi}(\vec{r}) \right] \quad (5.367)$$

The previous equation is seen as an extension of the first quantized form.

5.15.7 OPERATOR DYNAMICS

The previous sections in this chapter indicate that operators obey equations of motion using commutators with the Hamiltonian. For the Heisenberg picture,

$$\frac{d\hat{O}_h}{dt} = \frac{i}{\hbar} [\hat{\mathcal{H}}, \hat{O}_h] \quad (5.368a)$$

while for the interaction picture

$$\frac{d\hat{O}}{dt} = \frac{i}{\hbar} [\hat{\mathcal{H}}_0, \hat{O}] \quad (5.368b)$$

where $\hat{\mathcal{H}}_0$ agrees with the Schrödinger Hamiltonian.

For second quantization, we take

$$\hat{\mathcal{H}}_0 = \sum_n E_n \hat{b}_n^+ \hat{b}_n(t) \quad (5.369)$$

The equation of motion for the annihilation operators becomes

$$\frac{d\hat{b}_m}{dt} = \frac{i}{\hbar} \left[\sum_n E_n \hat{b}_n^+ \hat{b}_n, \hat{b}_m \right] = \frac{i}{\hbar} \sum_n E_n [\hat{b}_n^+, \hat{b}_m] \hat{b}_n = \frac{i}{\hbar} \sum_n E_n (-\delta_{mn}) \hat{b}_n = \frac{-iE_m}{\hbar} \hat{b}_m$$

Solving this ordinary differential equation provides

$$\hat{b}_m(t) = \hat{b}_m e^{\frac{E_m t}{i\hbar}} \quad (5.370)$$

where the coefficient $\hat{b}_m$ does not depend on time.

Because of Equation 5.370, the time dependence of operators (in the interaction representation) drops out for operators of the form $\hat{b}_n^+ \hat{b}_n$. For example, the Hamiltonian becomes

$$\hat{\mathcal{H}}_0 = \sum_n E_n \hat{b}_n^+ \hat{b}_n \quad (5.371)$$

5.15.8 ORIGIN OF BOSON CREATION AND ANNIHILATION OPERATORS

We now investigate the origin of Fock states and apply the results to the creation and annihilation operators. We continue to work with an N -particle system but do not distinguish between fermions and bosons. The development follows the excellent book by Fetter and Walecka. Recall the Hamiltonian and general wave function have the form

$$\hat{H} = \sum_k \hat{T}(x_k) + \frac{1}{2} \sum_{\substack{k,j \\ k \neq j}} \hat{V}(x_k, x_j) \quad (5.372)$$

The general wave function satisfying the many body Schrödinger equation

$$\hat{H}\psi = i\hbar \frac{\partial}{\partial t} \psi \quad (5.373)$$

has the form

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{W_1, W_2, \dots, W_N} C(W_1, W_2, \dots, W_N, t) u_{W_1}(x_1) u_{W_2}(x_2) \dots u_{W_N}(x_N) \quad (5.374)$$

where the notation has been changed for later convenience. The W_i denotes the energy eigenvalue for the particle i . Substituting Equation 5.374 into Equation 5.373, provides

$$\begin{aligned} i\hbar \sum_{W_1, W_2, \dots, W_N} \left[\frac{\partial}{\partial t} C(W_1, \dots, W_N, t) \right] u_{W_1}(x_1) \dots u_{W_N}(x_N) \\ = \sum_{W_1, \dots, W_N} C(W_1, \dots, W_N, t) \left[\sum_k \hat{T}(x_k) + \frac{1}{2} \sum_{\substack{k,j \\ k \neq j}} \hat{V}(x_k, x_j) \right] u_{W_1}(x_1) \dots u_{W_N}(x_N) \end{aligned}$$

Factor out the two summations on the right-hand side, multiply from the right by the operator

$$\int dx_1 \dots dx_N u_{E_1}^*(x_1) \dots u_{E_N}^*(x_N)$$

(where $E_1, E_2, \dots$ are now specific energy values) to find

$$\begin{aligned} i\hbar \sum_{W_1, W_2, \dots, W_N} \left[\frac{\partial}{\partial t} C(W_1, \dots, W_N, t) \right] \int dx_1 \dots dx_N u_{E_1}^*(x_1) \dots u_{E_N}^*(x_N) u_{W_1}(x_1) \dots u_{W_N}(x_N) \\ = \sum_{W_1, \dots, W_N} C(W_1, \dots, W_N, t) \int dx_1 \dots dx_N u_{E_1}^*(x_1) \dots u_{E_N}^*(x_N) \left[\sum_k \hat{T}(x_k) \right] u_{W_1}(x_1) \dots u_{W_N}(x_N) \\ + \sum_{W_1, \dots, W_N} C(W_1, \dots, W_N, t) \int dx_1 \dots dx_N u_{E_1}^*(x_1) \dots u_{E_N}^*(x_N) \left[\frac{1}{2} \sum_{\substack{k,j \\ k \neq j}} \hat{V}(x_k, x_j) \right] u_{W_1}(x_1) \dots u_{W_N}(x_N) \end{aligned}$$

The functions $u_{E_j}(x_j)$ are a particular choice of the basis functions so that the orthonormality relations

$$\int dx_j u_{E_j}^*(x_j) u_{W_j}(x_j) = \delta_{E_j, W_j}$$

can be used to simplify the equations (notice both functions in the integral have the same coordinates). The result is

$$\begin{aligned} i\hbar \frac{\partial}{\partial t} C(E_1, \dots, E_N, t) \\ = \sum_k \sum_{W_k} C(E_1, \dots, E_{k-1}, W_k, E_{k+1}, \dots, t) \int dx_k u_{E_k}^*(x_k) \hat{T}(x_k) u_{W_k}(x_k) \\ + \sum_{\substack{k,j \\ k \neq j}} \sum_{W_k W_j} C(E_1, \dots, W_j, E_{j+1}, \dots, W_k, E_{k+1}, \dots, t) \int dx_j dx_k u_{E_j}^*(x_j) u_{E_k}^*(x_k) \frac{1}{2} \hat{V}(x_k, x_j) u_{W_j}(x_j) u_{W_k}(x_k) \end{aligned}$$

Once again restrict the argument to bosons. Consider the coefficient $C(E_1, \dots, E_{k-1}, E_k, E_{k+1}, \dots, t)$ with the corresponding number coefficient given by

$$\overline{C}(n_1, n_2, \dots, n_{E_k}, \dots, t) = C(E_1, \dots, E_k, \dots, t)$$

where n_{E_k} means the number of particles with the energy E_k . The coefficient $C(E_1, \dots, E_{k-1}, W_k, E_{k+1}, \dots, t)$ changes the energy E_k of particle $\#k$ to the new energy W_k . There is one less particle with energy E_k and one more with W_k . Therefore,

$$C(E_1, \dots, E_{k-1}, W_k, E_{k+1}, \dots, t) = \overline{C}(n_1, \dots, n_{E_k} - 1, \dots, n_{W_k} + 1, \dots, t)$$

This can be incorporated in the kinetic energy term

$$\text{ke} = \sum_k \sum_{W_k} C(E_1, \dots, E_{k-1}, W_k, E_{k+1}, \dots, t) \int dx_k u_{E_k}^*(x_k) \hat{T}(x_k) u_{W_k}(x_k)$$

by considering a general sum of the form

$$\sum_k f(E_k) = f(a) + f(b) + \dots$$

where the symbols $a, b, c \dots$ represent one of the possible energy values E . Suppose $a, b, c \dots$ have energy E_1 , and $k, l, m \dots$ have energy E_2 , and so on. The terms in the sum can be grouped according to the different energy values

$$\sum_k f(k) = f(a) + f(b) + f(c) + \dots + f(k) + f(l) + f(m) + \dots + \dots = \sum_E n_E f(E)$$

Therefore, the kinetic energy term becomes

$$\begin{aligned} \text{ke} &= \sum_k \sum_{W_k} C(E_1, \dots, E_{k-1}, W_k, E_{k+1}, \dots, t) \int dx_k u_{E_k}^*(x_k) \hat{T}(x_k) u_{W_k}(x_k) \\ &= \sum_E n_E C(n_1, n_2, \dots, n_E - 1, \dots, n_W + 1, \dots, t) \langle E | \hat{T} | W \rangle \end{aligned}$$

Let i, j now represent the energy values, we can write

$$\begin{aligned} \text{ke} &= \sum_{EW} n_E \overline{C}(n_1, n_2, \dots, n_E - 1, \dots, n_W + 1, \dots, t) \langle E | \hat{T} | W \rangle \\ &= \sum_{ij} n_i \langle i | \hat{T} | j \rangle \overline{C}(n_1, n_2, \dots, n_i - 1, \dots, n_j + 1, \dots, t) \end{aligned}$$

Fetter and Walecka also evaluate the potential energy term. When the two results are combined with the coefficients from Equation 5.365

$$\beta(n_1, n_2, \dots, n_\infty, t) = \left(\frac{N!}{n_1! n_2! \dots n_\infty!} \right)^{1/2} \overline{C}(n_1, n_2, \dots, n_\infty, t)$$

they end up with a messy looking equation.

$$\begin{aligned}
 i\hbar \frac{\partial}{\partial t} \beta(n_1, n_2, \dots, n_\infty, t) &= \sum_i \langle i | \hat{T} | i \rangle n_i \beta(n_1, \dots, n_i, \dots, n_\infty, t) \\
 &+ \sum_{\substack{ij \\ i \neq j}} \langle i | \hat{T} | j \rangle \sqrt{n_i} \sqrt{n_j + 1} \beta(n_1, n_2, \dots, n_i - 1, \dots, n_j + 1, \dots, t) \\
 &+ \sum_{i \neq j \neq k \neq m} \langle ij | \hat{V} | km \rangle \frac{1}{2} \sqrt{n_i} \sqrt{n_j} \sqrt{n_k + 1} \sqrt{n_m + 1} \beta(\dots, n_i - 1, \dots, n_j - 1, \dots, n_k \\
 &\quad + 1, \dots, n_m + 1, \dots, t) \\
 &+ \sum_{i \neq j \neq k \neq m} \langle ii | \hat{V} | km \rangle \frac{1}{2} \sqrt{n_i} \sqrt{n_i - 1} \sqrt{n_k + 1} \sqrt{n_m + 1} \beta(\dots, n_i - 2, \dots, n_k \\
 &\quad + 1, \dots, n_m + 1, \dots, t) \\
 &+ ETC
 \end{aligned}$$

There is one of these long equations for each set of occupation numbers $n_1, n_2, \dots$

We can now proceed as follows. Using the Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle = \hat{H} |\Psi(t)\rangle \quad (5.375)$$

where

$$\psi(x_1, x_2, \dots, x_N, t) = \sum_{n_1, n_2, \dots, n_\infty} \beta(n_1, n_2, \dots, n_\infty, t) \phi_{n_1, n_2, \dots, n_\infty}(x_1, x_2, \dots, x_N)$$

or

$$|\psi(t)\rangle = \sum_{n_1, n_2, \dots, n_\infty} \beta(n_1, n_2, \dots, n_\infty, t) |n_1, n_2, \dots, n_\infty\rangle \quad (5.376)$$

By substituting Equation 5.376 in Equation 5.375 and working with the Hamiltonian,

$$i\hbar \sum_{n_1, n_2, \dots, n_\infty} \frac{\partial \beta(n_1, n_2, \dots, n_\infty, t)}{\partial t} |n_1, n_2, \dots, n_\infty\rangle = \hat{H} |\Psi(t)\rangle \quad (5.377)$$

The expression for the derivative of β (long equation above) can be substituted into Equation 5.377 to yield an alternate expression for $\hat{H}$. The second kinetic energy term

$$i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle = \dots + \sum_{n_1, n_2, \dots, n_\infty} \sum_{\substack{ij \\ i \neq j}} \langle i | \hat{T} | j \rangle \beta(\dots, n_i - 1, \dots, n_j + 1, \dots, t) \sqrt{n_i} \sqrt{n_j + 1} |n_1, \dots, n_\infty\rangle + \dots \quad (5.378)$$

Notice that the square roots and the Fock state are almost the form required for creation and annihilation operators. Redefine the dummy indices according to

$$n_i - 1 \rightarrow n_i, \quad n_j + 1 \rightarrow n_j$$

to get

$$i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle = \cdots + \sum_{n_1, n_2, \dots, n_\infty} \sum_{\substack{ij \\ i \neq j}} \langle i | \hat{T} | j \rangle \beta(\dots, n_i, \dots, n_j, \dots, t) \sqrt{n_i + 1} \sqrt{n_j} |\dots, n_i + 1, \dots, n_j - 1, \dots\rangle + \cdots$$

Now we can substitute the creation and annihilation operators to get

$$i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle = \cdots + \sum_{n_1, n_2, \dots, n_\infty} \sum_{\substack{ij \\ i \neq j}} \langle i | \hat{T} | j \rangle \beta(\dots, n_i, \dots, n_j, \dots, t) \hat{b}_i^+ \hat{b}_j |\dots, n_i, \dots, n_j, \dots\rangle + \cdots$$

All of the terms in the expansion Equation 5.378 can be rewritten in terms of the creation and annihilation operators. The result is

$$\hat{H} = \sum_{i,j} b_i^+ \langle i | \hat{T} | j \rangle b_j + \frac{1}{2} \sum_{ijk} b_i^+ b_j^+ \langle ij | \hat{V} | km \rangle b_k b_m$$

5.16 PROPAGATOR

The propagator represents a conditional probability that a particle will be found at one point given that it started at another. Similar to a Green function, the propagator can be viewed as a function that moves a wave function in space and time. As a Green function, it satisfies Schrödinger's equation with a Dirac delta forcing function. Green functions find common applications in electromagnetics, control theory, and especially in particle theory. The notions of the propagator and the Feynman path integral stress the fact that the wave function “samples” all regions of space in traveling from one point to another.

5.16.1 IDEA OF THE GREEN FUNCTION

The Green function makes solving partial differential equations more convenient. Suppose $\hat{L}(x, t)$ represents a linear differential operator in space and time such as for the Schrödinger equation $\hat{L} = \hat{\mathcal{H}} - i\hbar\partial_t$. A partial differential equation can be solved for a variety of forcing functions $f(t)$

$$\hat{L}\psi(x, t) = f(x, t) \quad (5.379)$$

once finding the solution G to the same equation with Dirac delta functions replacing the forcing function

$$\hat{L}G(x, t) = \delta(x)\delta(t) \quad (5.380)$$

Note, if a clock starts at $t = 0$ (actually, infinitesimally before zero denoted by 0^-), then the right-hand side of Equation 5.380 represents a specific initial condition of creating a unit disturbance at $t = 0$ and localized to $x = 0$. We can show a solution to Equation 5.379 must be

$$\psi(x, t) = \int dx' dt' G(x - x', t - t') f(x', t') \quad (5.381)$$

We can easily show that the function ψ in this last equation satisfies Equation 5.380 just by substituting (and remember that the operator depends on x, t and not x', t'). In the case of

Equation 5.381, the green function has the interpretation of moving the disturbance in space and time to provide the solution.

Example 5.45

Find the charge density as a function of time from the conservation equation $\partial_t \rho - \nabla \cdot \vec{j} = 0$ where ρ and $\vec{j}$ represent the charge density and current density, respectively. Assume an impulse of charge created exactly at $t=0$. Assume the charge does not flow once created.

SOLUTION

The charge generation term has the form $\delta(x)\delta(t)$, which produces the conservation equation $\partial_t \rho - \nabla \cdot \vec{j} = \delta(x)\delta(t)$. Setting the current to zero and integrating over space yields a differential equation $\partial_t Q = \delta(t)$ for the charge $Q(t)$. Integrating over time shows that the total charge must be $Q = \theta(t)$ at $x=0$ where θ gives the step function.

5.16.2 PROPAGATOR FOR A CONSERVATIVE SYSTEM

The propagator moves a wave function through space and time. In this section, we present the algebra.

Consider a conservative (i.e., closed) system. The evolution operator is $\hat{u}(t) = e^{\hat{H}t/(i\hbar)}\theta(t)$. The wave function at a later time can be written as

$$|\psi(t)\rangle = \hat{u}(t - t')|\psi(t')\rangle$$

The probability amplitude for finding a particle at x can be written as

$$\langle x|\psi(t)\rangle = \langle x|\hat{u}(t - t')|\psi(t')\rangle$$

Substituting the resolution of 1 for the coordinate basis provides

$$\langle x|\psi(t)\rangle = \int dx' \langle x|\hat{u}(t - t')|x'\rangle \langle x'|\psi(t')\rangle = \int dx' \langle x|\hat{u}(t - t')|x'\rangle \psi(x', t') \quad (5.382)$$

The propagator is seen to be ($t > t'$)

$$G(x, x'; t, t') = \langle x|\theta(t - t')e^{\frac{\hat{H}(t-t')}{i\hbar}}|x'\rangle \quad (5.383)$$

The form of Equation 5.382 shows that the propagator produces a wave function at the point x at time t provided an initial wave function $\psi(x', t')$ is known. The integral over all the initial points x' shows that all portions of the wave can propagate to the final point x . This behavior is reminiscent of Huygen's principle from optics (see Figure 5.79). A wave passing through a slit behaves as if all

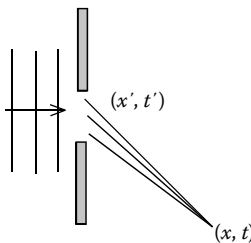


FIGURE 5.79 Points within the slit scatter the incident optical waves in all directions.

points within the slit scatter the wave in all forward directions. We must sum over all of these individual wave amplitudes to find the resultant wave at the forward point x at time t .

We can see that G is also a Green function by performing the following calculation:

$$(i\hbar\partial_t - \hat{\mathcal{H}})e^{\frac{i\hat{\mathcal{H}}(t-t')}{\hbar}}\theta(t-t') = i\hbar\delta(t-t')$$

Operating with $\langle x|$ and $|x'\rangle$ yields

$$(i\hbar\partial_t - \hat{\mathcal{H}}(x))\langle x|e^{\frac{i\hat{\mathcal{H}}(t-t')}{\hbar}}|x'\rangle\theta(t-t') = i\hbar\delta(x-x')\delta(t-t')$$

where the Hamiltonian has been projected onto a coordinate basis. Therefore the propagator is also a green function.

5.16.3 ALTERNATE FORMULATION

Assume a particle definitely starts at the point x_0 at time t_0 (or in some small volume centered on the point). The ket $|x_0, t_0\rangle$ can be used to represent this initial position. We are interested in the probability of finding the particle at point x at time t as represented by the ket $|x, t\rangle$. We will find the propagator from the probability amplitude $\langle x, t|x_0, t_0\rangle$ that a particle starting at x_0 at time t_0 ends up at point x at time t . This clearly shows that the propagator has the form of a conditional probability. The ket $|x, t\rangle$ is in the Heisenberg representation.

The first thing to realize is that a coordinate ket in the Schrödinger representation does not carry any time dependence. The wave functions carry the time dependence and so the coordinate projectors do not need any. We can easily find the coordinate kets in the Heisenberg representation by rewriting $\langle x|\psi(t)\rangle$. We require (h = Heisenberg)

$$\langle x|\psi(t)\rangle = \langle x_h|\psi_h\rangle \equiv \langle x, t|\psi_h\rangle$$

Substituting the evolution operator we find

$$\langle x, t|\psi_h\rangle = \langle x|\psi(t)\rangle = \langle x|\hat{u}(t)|\psi_h\rangle$$

from which we can identify the relation between the Heisenberg projector $\langle x, t|$ and the Schrödinger one $\langle x|$.

$$\langle x, t| = \langle x|\hat{u}(t) \quad \rightarrow \quad |x, t\rangle = \hat{u}^+(t)|x\rangle \quad (5.384)$$

The propagator can now be written as

$$G(x, t; x_0, t_0) = \langle x, t|x_0, t_0\rangle = \langle x|\hat{u}(t)\hat{u}^+(t_0)|x_0\rangle = \langle x|e^{\frac{i\hat{\mathcal{H}}(t-t_0)}{\hbar}}|x_0\rangle \quad (5.385)$$

as previously found in Equation 5.383. This time, we did not introduce the integral over all initial positions. We could do so though, by noting that we started with a particle definitely located at point x_0, t_0 and generalize to the case that the particle is smeared across space (using the wave function). Then an integral is clearly indicated.

A couple of comments should be made. First, the propagator can be represented as the trace of a transition operator $G = \text{Tr}|x_0, t_0\rangle\langle x, t| = \langle x, t|x_0, t_0\rangle$. And second, Equation 5.385 shows that the propagator approaches the Dirac delta function in the limit.

$$\lim_{t \rightarrow t_0} \langle x|e^{\frac{i\hat{\mathcal{H}}(t-t_0)}{\hbar}}|x_0\rangle = \langle x|x_0\rangle = \delta(x-x_0)$$

5.16.4 PROPAGATOR AND THE PATH INTEGRAL

We now illustrate the propagator found in Equations 5.385 and 5.383 using a path integral approach. Now we include all space-time points between the initial and final points. Figure 5.80 shows two of the many paths. The initial point is (x_0, t_0) and the final point is $(x, t) = (x_4, t_4)$. Actually, we want to find the amplitude of the wave function reaching the point (x, t) regardless of where it originates along the x -axis. Technically, we are working with a 1-D problem (in spatial coordinates) which means we are asking how the wave travels along the single x -axis starting at any point and traveling in either direction to reach the final destination at (x, t) . The line segments are made small enough that they closely approximate the actual curved paths.

The propagators resemble conditional probabilities. The probability of reaching x given that the wave made it to x_3 must be $\langle x, t | x_3, t_3 \rangle$. The probability of reaching point 3 given that the wave reached point 2 must be $\langle x_3, t_3 | x_2, t_2 \rangle$. Therefore the probability of reaching point x given that it reached point 2 must be the product of the two small path segments $\langle x, t | x_3, t_3 \rangle \langle x_3, t_3 | x_2, t_2 \rangle$. However, there exist a large number of other paths spanning the distance between points 2 and 4. We must sum over these paths in accordance with the basic principles of quantum theory. We now have

$$\langle x, t | x_2, t_2 \rangle = \int dx_3 \langle x, t | x_3, t_3 \rangle \langle x_3, t_3 | x_2, t_2 \rangle$$

The process continues for the x_1 points. We find the propagator $\langle x, t | x_0, t_0 \rangle$ (probability amplitude) for a particle starting at the point (x_0, t_0) and reaching the point (x, t) along the four path segments.

$G(x, t; x_0, t_0) = \langle x, t | x_0, t_0 \rangle = \int dx_3 \int dx_2 \int dx_1 \langle x, t | x_3, t_3 \rangle \langle x_3, t_3 | x_2, t_2 \rangle \langle x_2, t_2 | x_1, t_1 \rangle \langle x_1, t_1 | x_0, t_0 \rangle$
We can substitute the results for each small propagator from Equation 5.385 to find

$$G(x, t; x_0, t_0) = \int dx_3 \int dx_2 \int dx_1 \langle x | e^{\frac{i\hat{H}(t-t_3)}{\hbar}} | x_3 \rangle \langle x_3 | e^{\frac{i\hat{H}(t_3-t_2)}{\hbar}} | x_2 \rangle \langle x_2 | e^{\frac{i\hat{H}(t_2-t_1)}{\hbar}} | x_1 \rangle \langle x_1 | e^{\frac{i\hat{H}(t_1-t_0)}{\hbar}} | x_0 \rangle$$

Using the closure relations for x_3 , x_2 , and x_1 produces the results

$$G(x, t; x_0, t_0) = \langle x | e^{\frac{i\hat{H}(t-t_3)}{\hbar}} e^{\frac{i\hat{H}(t_3-t_2)}{\hbar}} e^{\frac{i\hat{H}(t_2-t_1)}{\hbar}} e^{\frac{i\hat{H}(t_1-t_0)}{\hbar}} | x_0 \rangle$$

The arguments of the exponentials all commute and the exponentials can all be combined to find

$$G(x, t; x_0, t_0) = \langle x | e^{\frac{i\hat{H}(t-t_0)}{\hbar}} | x_0 \rangle$$

just as we found previously. Here we see the intermediate times drop from consideration for a conservative system.

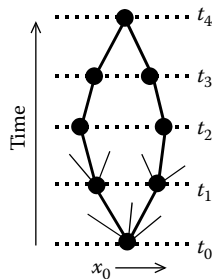


FIGURE 5.80 Two of many possible paths spanning between the initial and final points.

In general, the propagator has the form

$$G(x, t; x_0, t_0) = \lim_{\substack{N \rightarrow \infty \\ \varepsilon \rightarrow 0}} \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1}, t_{n+1} | x_n, t_n \rangle = \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1} | e^{\frac{j\hat{H}(t_{n+1}-t_n)}{\hbar}} | x_n \rangle \quad (5.386)$$

where point # N is (x, t) , $N\varepsilon = t - t_0$ where ε is the small interval of time between the time slices appearing in Figure 5.81, and the measure $\mathcal{D}x = dx_1 dx_2 \dots dx_{N-1}$ integrates over the intermediate spatial points.

5.16.5 FREE-PARTICLE PROPAGATOR

Consider a single particle moving through space void of any potential energy. The Hamiltonian can be written as $\mathcal{H} = \frac{\hat{p}^2}{2m}$. We calculate the propagator using the more complicated method (Equation 5.386) rather than the four easy steps required by Equation 5.385 (see chapter problems). We will find the propagator to be

$$G(x, t; x_0, t_0) = \langle x | e^{\frac{j\hat{H}(t-t_0)}{\hbar}} | x_0 \rangle = \sqrt{\frac{m}{2\pi i \hbar (t - t_0)}} e^{\frac{im(x-x_0)^2}{2\hbar(t-t_0)}} \quad (t > t_0)$$

We need to calculate many integrals such as the Fourier transform in the momentum representation.

$$|f\rangle = \int dp |p\rangle \langle p | f \rangle \quad \text{where} \quad \int dp |p\rangle \langle p| = 1$$

where $\hat{P}|p\rangle = p|p\rangle$, $p = \hbar k$, and the momentum basis functions projected on the x -axis have a form very similar to the Fourier basis set $\left\{ \langle x | p \rangle = \frac{e^{ipx/\hbar}}{\sqrt{2\pi\hbar}} \right\}$.

Let us calculate $\langle x_{n+1} | e^{\frac{j\hat{H}(t_{n+1}-t_n)}{\hbar}} | x_n \rangle$. Assume equal spacing between times $t_{i+1} - t_i = \varepsilon$ and the total spacing of $t - t_0 = N\varepsilon$.

$$\langle x_{n+1} | e^{\frac{j\hat{H}(t_{n+1}-t_n)}{\hbar}} | x_n \rangle = \langle x_{n+1} | e^{\frac{\varepsilon}{\hbar} \frac{\hat{p}_n^2}{2m}} | x_n \rangle$$

where $\hat{p}_n$ represents the momentum operator on the path length connecting points x_n and x_{n+1} . Next insert the closure relation for the momentum basis set between x_{n+1} and the exponential so that the operator can be written as a c-number.

$$\langle x_{n+1} | e^{\frac{j\hat{H}(t_{n+1}-t_n)}{\hbar}} | x_n \rangle = \langle x_{n+1} | e^{\frac{\varepsilon}{\hbar} \frac{\hat{p}_n^2}{2m}} | x_n \rangle = \int dp_n \langle x_{n+1} | p_n \rangle \langle p_n | e^{\frac{\varepsilon}{\hbar} \frac{\hat{p}_n^2}{2m}} | x_n \rangle = \int dp_n \langle x_{n+1} | p_n \rangle \langle p_n | e^{\frac{\varepsilon}{\hbar} \frac{p_n^2}{2m}} | x_n \rangle$$

since

$$e^{-\frac{\varepsilon}{\hbar} \frac{\hat{p}_n^2}{2m}} | p \rangle = e^{-\frac{\varepsilon}{\hbar} \frac{p^2}{2m}} | p \rangle \quad \rightarrow \quad \langle p | e^{\frac{\varepsilon}{\hbar} \frac{\hat{p}_n^2}{2m}} = \langle p | e^{\frac{\varepsilon}{\hbar} \frac{p^2}{2m}}$$

The last propagator results essentially assumes that p_n is constant over the small path length. Now the projector $\langle p_n |$ can be moved past the c-number (the exponential)

$$\langle x_{n+1} | e^{\frac{j\hat{H}(t_{n+1}-t_n)}{\hbar}} | x_n \rangle = \int dp_n \langle x_{n+1} | p_n \rangle \langle p_n | x_n \rangle e^{\frac{\varepsilon}{\hbar} \frac{p_n^2}{2m}} = \int \frac{dp_n}{2\pi\hbar} e^{\frac{ip_n(x_{n+1}-x_n)}{\hbar}} e^{\frac{\varepsilon}{\hbar} \frac{p_n^2}{2m}} \quad (5.387)$$

Integrals of the type in Equation 5.387 (integrated over the entire axis) can be evaluated using the results for a Gaussian. The integral of the Gaussian can be written as

$$\int_{-\infty}^{\infty} dx e^{-ax^2+bx} = \sqrt{\frac{\pi}{a}} e^{\frac{b^2}{4a}} \quad \text{when } \text{Re}(a) > 0 \quad (5.388)$$

The chapter problems evaluate the integral and we find

$$\langle x_{n+1} | e^{\frac{ji(t_{n+1}-t_n)}{\hbar}} | x_n \rangle = \int \frac{dp_n}{2\pi\hbar} e^{\frac{i}{\hbar} p_n (x_{n+1}-x_n)} e^{\frac{p_n^2}{2m}} = \sqrt{\frac{m}{2\pi i\hbar\epsilon}} e^{\frac{im}{2\hbar} \left(\frac{x_{n+1}-x_n}{\epsilon}\right)^2} \quad (5.389)$$

Now we can work with the entire propagator in Equation 5.386, specifically

$$G(x, t; x_0, t_0) = \lim_{\substack{N \rightarrow \infty \\ \epsilon \rightarrow 0}} \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1}, t_{n+1} | x_n, t_n \rangle = \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1} | e^{\frac{ji(t_{n+1}-t_n)}{\hbar}} | x_n \rangle$$

where $\mathcal{D}x = dx_1 dx_2 \dots dx_{N-1}$ and $N\epsilon = t - t_0$. The single term

$$G(x_1, t_1; x_0, t_0) = \langle x_1, t_1 | x_0, t_0 \rangle = \sqrt{\frac{m}{2\pi i\hbar\epsilon}} e^{\frac{im}{2\hbar} \left(\frac{x_1-x_0}{\epsilon}\right)^2} = \sqrt{\frac{m}{2\pi i\hbar(\epsilon)}} e^{\frac{im}{2\hbar(\epsilon)} (x_1-x_0)^2} \quad (5.390)$$

does not require an integral. The second two terms

$$\langle x_2, t_2 | x_0, t_0 \rangle = \int_{-\infty}^{\infty} dx_1 \langle x_2, t_2 | x_1, t_1 \rangle \langle x_1, t_1 | x_0, t_0 \rangle = \int_{-\infty}^{\infty} dx_1 \sqrt{\frac{m}{2\pi i\hbar\epsilon}} e^{\frac{im}{2\hbar} \left(\frac{x_2-x_1}{\epsilon}\right)^2} \sqrt{\frac{m}{2\pi i\hbar\epsilon}} e^{\frac{im}{2\hbar} \left(\frac{x_1-x_0}{\epsilon}\right)^2}$$

require an integral over a Gaussian

$$\int_{-\infty}^{\infty} dx e^{a(x-x_0)^2-b(x-x_1)^2} = \sqrt{\frac{-\pi}{a+b}} e^{\frac{ab}{a+b}(x_0-x_1)^2} \quad (5.391)$$

We find

$$G(x_2, t_2; x_0, t_0) = \langle x_2, t_2 | x_0, t_0 \rangle = \sqrt{\frac{m}{2\pi i\hbar(2\epsilon)}} e^{\frac{im}{2\hbar(2\epsilon)} (x_2-x_0)^2}$$

This is the same as $G(x_1, t_1; x_0, t_0)$ in Equation 5.390 except $\epsilon \rightarrow 2\epsilon$. By induction, the remainder of the integral in Equation 5.386 must have the form

$$G(x, t; x_0, t_0) = \lim_{\substack{N \rightarrow \infty \\ \epsilon \rightarrow 0}} \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1}, t_{n+1} | x_n, t_n \rangle = \sqrt{\frac{m}{2\pi i\hbar(N\epsilon)}} e^{\frac{im}{2\hbar(N\epsilon)} (x_N-x_0)^2} \quad (5.392)$$

The limit produces $N\epsilon = t - t_0$ and $x_N = x$.

$$G(x, t; x_0, t_0) = \sqrt{\frac{m}{2\pi i\hbar(t-t_0)}} e^{\frac{im}{2\hbar(t-t_0)} (x-x_0)^2} \quad (5.393)$$

5.17 FEYNMAN PATH INTEGRAL

The Feynman path integral provides a beautiful link between the classical and quantum theory. This path integral treats the classical action as a phase (the argument of a complex exponential). The path of a quantum particle in configuration space (or Euclidean space) nearly follows the classical path since the action is nearly stationary there and therefore provides coherent summations across nearby paths. Because of the role of the action, the Feynman path integral can provide an alternate means of developing the Hamiltonian and the Schrödinger equation. The propagator and the path integral play key roles in Feynman diagrams for interactions. The reader would likely enjoy reading the Feynman easy-to-read but-full-of-wisdom book titled “QED”. Don’t confuse this title with the one spelled out as “Quantum Electrodynamics” as this later one could not be termed easy-reading.

5.17.1 DERIVATION OF THE FEYNMAN PATH INTEGRAL

One method of developing the Feynman path integral starts with the propagator for a single particle in a potential $\hat{V}$, which depends only on the 1-D spatial variable x (and not time).

$$\hat{H} = \frac{\hat{p}^2}{2m} + V(\hat{x}) \tag{5.394}$$

Similar to the development of the propagator, we consider the many paths from a point $|x_0, t_0\rangle$ (Heisenberg coordinate) to the point $|x, t\rangle$ (see Figure 5.81). The propagator can be written as

$$\langle x, t | x_0, t_0 \rangle = \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1} | e^{-\frac{i}{\hbar} \hat{H}(t_{n+1} - t_n)} | x_n \rangle \tag{5.395}$$

where

$\mathcal{D}x = dx_1 dx_2 \cdots dx_{N-1}$
 N denotes the number of small path lengths
 $x = x_N$

The key step concerns the method of evaluating the matrix elements in Equation 5.395. A number of treatments can be found (see chapter references) including those that use (1) the Weyl ordering, (2) normal ordering with small time steps, (3) a constant potential energy, (4) very close path elements with commutators of kinetic and potential energy that are zero so that there is no transfer between

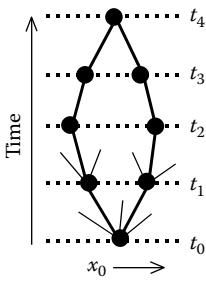


FIGURE 5.81 Two of many possible paths spanning between the initial and final points.

kinetic and potential energy. We assume infinitesimally small time steps $\varepsilon = t_{n+1} - t_n$ so that the exponential can be approximated to first order. The path integral becomes

$$\langle x, t | x_0, t_0 \rangle = \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1} | e^{-\frac{i}{\hbar} \hat{H}(t_{n+1} - t_n)} | x_n \rangle = \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1} | e^{-\frac{i}{\hbar} \hat{H} \varepsilon} | x_n \rangle \quad (5.396)$$

Consider the n th term and expand to first order in ε to find

$$\langle x_{n+1} | e^{\frac{\varepsilon}{i\hbar} \hat{H}} | x_n \rangle \cong \langle x_{n+1} | \left\{ 1 + \frac{\varepsilon}{i\hbar} \hat{H} \right\} | x_n \rangle = \langle x_{n+1} | \left\{ 1 + \frac{\varepsilon}{i\hbar} \left(\frac{\hat{p}^2}{2m} + V(\hat{x}) \right) \right\} | x_n \rangle \quad (5.397)$$

We know from previous sections to insert the closure relation in the momentum basis. We can do this later but until then, don't evaluate any inner products between coordinates.

The matrix element for the potential in Equation 5.397 should be handled first. A couple of variations occur in the literature. For one, the inner product can be written as

$$\langle x_{n+1} | V(\hat{x}) | x_n \rangle = V(x_n) \langle x_{n+1} | x_n \rangle$$

This form comes from the “normal ordering” method as well. Most commonly, the matrix element takes on a symmetric appearance using the Weyl ordering. We can see how this happens by making a linear approximation of $V = 1 + c_1 x$ and then computing the matrix elements. We find

$$\begin{aligned} \langle x_{n+1} | V(\hat{x}) | x_n \rangle &= \langle x_{n+1} | 1 + c_1 \hat{x} | x_n \rangle = \langle x_{n+1} | x_n \rangle + c_1 \langle x_{n+1} | \left(\frac{\hat{x} + \bar{x}}{2} \right) | x_n \rangle = \langle x_{n+1} | 1 + c_1 \bar{x}_n | x_n \rangle \\ &\cong \langle x_{n+1} | V(\bar{x}_n) | x_n \rangle \end{aligned}$$

where the average value of the position along the small path element $\bar{x}_n = (x_{n+1} + x_n)/2$ is a real number. The potential is essentially constant and therefore commutes with the kinetic energy. Substituting back into Equation 5.397 produces

$$\langle x_{n+1} | e^{\frac{\varepsilon}{i\hbar} \hat{H}} | x_n \rangle \cong \langle x_{n+1} | \left\{ 1 + \frac{\varepsilon}{i\hbar} \left(\frac{\hat{p}^2}{2m} + V(\bar{x}_n) \right) \right\} | x_n \rangle \cong \langle x_{n+1} | e^{\frac{\varepsilon}{i\hbar} \hat{H}(\hat{p}, \bar{x}_n)} | x_n \rangle \quad (5.398)$$

We need to remove the momentum operator from the Hamiltonian. This can be accomplished by inserting the closure relation for the momentum between the bra and the exponential.

$$1 = \int dp |p\rangle \langle p| \quad \text{where } \langle x | p \rangle = \frac{e^{ipx/\hbar}}{\sqrt{2\pi\hbar}} \quad \text{and} \quad p = \hbar k \quad (5.399)$$

The matrix element in Equation 5.398 becomes

$$\begin{aligned} \langle x_{n+1} | e^{\frac{\varepsilon}{i\hbar} \hat{H}} | x_n \rangle &= \int dp_{n+1} \langle x_{n+1} | p_{n+1} \rangle \langle p_{n+1} | e^{\frac{\varepsilon}{i\hbar} \hat{H}(\hat{p}, \bar{x}_n)} | x_n \rangle \\ &= \int dp_{n+1} \langle x_{n+1} | p_{n+1} \rangle \langle p_{n+1} | x_n \rangle e^{\frac{\varepsilon}{i\hbar} \mathcal{H}(p_{n+1}, \bar{x}_n)} \\ &= \int dp_{n+1} \frac{1}{2\pi\hbar} e^{ip_{n+1}(x_{n+1} - x_n)} e^{\frac{\varepsilon}{i\hbar} \mathcal{H}(p_{n+1}, \bar{x}_n)} \end{aligned} \quad (5.400)$$

This last integral can be evaluated by substituting $\mathcal{H} = p_{n+1}^2/(2m) + V(\bar{x})$ and collecting the momentum terms.

$$\langle x_{n+1} | e^{\frac{i\epsilon}{\hbar} \hat{\mathcal{H}}} | x_n \rangle = \int dp_{n+1} \frac{1}{2\pi\hbar} e^{-\frac{i\epsilon}{2m\hbar} \left[p_{n+1}^2 - \frac{2mp_{n+1}(x_{n+1} - x_n)}{\epsilon} \right]} e^{-\frac{i\epsilon}{\hbar} V(\bar{x}_n)}$$

Completing the square and integrating gives

$$\langle x_{n+1} | e^{\frac{i\epsilon}{\hbar} \hat{\mathcal{H}}} | x_n \rangle = \sqrt{\frac{m}{2\pi i \hbar \epsilon}} e^{\frac{im\epsilon}{2\hbar} \left(\frac{x_{n+1} - x_n}{\epsilon} \right)^2} e^{-\frac{i\epsilon}{\hbar} V(\bar{x}_n)} \quad (5.401)$$

Now we are in a position to work with the full propagator in Equation 5.396.

$$\langle x, t | x_0, t_0 \rangle = \int \mathcal{D}x \prod_{n=0}^{N-1} \langle x_{n+1} | e^{\frac{i\epsilon}{\hbar} \hat{\mathcal{H}}} | x_n \rangle = \lim_{\substack{\epsilon \rightarrow 0 \\ N \rightarrow \infty}} \int \mathcal{D}x \left(\frac{m}{2\pi i \hbar \epsilon} \right)^{\frac{N}{2}} e^{\frac{i\epsilon}{\hbar} \sum_{n=0}^{N-1} \left[\frac{m}{2} \left(\frac{x_{n+1} - x_n}{\epsilon} \right)^2 - V(\bar{x}_n) \right]} \quad (5.402)$$

where $\mathcal{D}x = dx_1 dx_2 \cdots dx_{N-1}$ and where we take the limits $N \rightarrow \infty$, $\epsilon \rightarrow 0$ such that $N\epsilon = t - t_0$. We can make the following definitions

$$\lim_{\substack{\epsilon \rightarrow 0 \\ N \rightarrow \infty}} \frac{x_{n+1} - x_n}{\epsilon} = v_n, \quad \lim_{\substack{\epsilon \rightarrow 0 \\ N \rightarrow \infty}} \bar{x}_n = \lim_{\substack{\epsilon \rightarrow 0 \\ N \rightarrow \infty}} \frac{x_{n+1} + x_n}{2} = x_n \quad (5.403a)$$

and therefore in the limit, the summation becomes an integral

$$\sum_{n=0}^{N-1} \epsilon \left[\frac{m}{2} \left(\frac{x_{n+1} - x_n}{\epsilon} \right)^2 - V(\bar{x}_n) \right] = \sum_{n=0}^{N-1} \epsilon \left[\frac{m}{2} \dot{x}_n^2 - V(\bar{x}) \right] \Rightarrow \int dt \left[\frac{m}{2} \dot{x}^2 - V(x) \right] \quad (5.403b)$$

Therefore the propagator (a.k.a., the Feynman path integral) becomes

$$\langle x, t | x_0, t_0 \rangle = A \int \mathcal{D}x e^{\frac{i}{\hbar} \int_{t_0}^t dt \left[\frac{m}{2} \dot{x}^2 - V(x) \right]} = A \int \mathcal{D}x e^{\frac{i}{\hbar} S[x]} \quad (5.404)$$

where A is a constant. The quantity $S[x]$ is clearly the classical action

$$S[x] = \int_{t_0}^t dt \left[\frac{m}{2} \dot{x}^2 - V(x) \right] = \int_{t_0}^t dt L(x, \dot{x}) \quad (5.405)$$

since L is the classical Lagrangian for a single particle in a potential V .

5.17.2 CLASSICAL LIMIT

The classical limit corresponds to $\hbar \rightarrow 0$. Let us examine the propagator. Quantum mechanically, a particle initially located at (x_0, t_0) can propagate along many different paths to reach the final point (x, t) . Figure 5.82 shows the classical path (#0) surrounded by a number of other quantum mechanically possible paths. The classical path makes the action an extremum (hopefully a minimum). This means neighboring paths do not change the phase of the exponential very much in Equation 5.404. Consequently, paths close enough to the classical path produce phases that

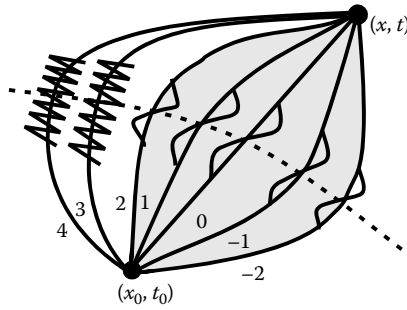


FIGURE 5.82 Cartoon representation of multiple paths leading from the initial to final points. Path #0 corresponds to the classical path minimizing the action. Paths in the shaded area coherently add phases.

coherently add in the propagator such as those in the shaded area # (-2) to # 2 . Notice how Figure 5.82 illustrates the coherence between paths by showing how sinusoid-like waves match each other along the dotted curve. Those paths further from the classical path produce large variations in phase and incoherently add so as to cancel in the propagator. Therefore, the classical particle cannot follow paths too far from the classical path. Now, $\hbar \rightarrow 0$ makes the exponential more sensitive to small changes in the phase. Consequently, the group of “allowed” paths becomes smaller. In the limit, only the classical path survives.

5.17.3 SCHRÖDINGER EQUATION FROM THE PROPAGATOR

The path integral should be capable of reproducing the results of the quantum theory. The Schrödinger wave equation (SWE) represents a significant amount of the quantum theory. It is a partial differential equation that describes the character of the wave function based on infinitesimal changes of the coordinates. The path integral represents the entire set of paths possibly followed by a particle. Therefore, to recover the Schrödinger equation, we must consider infinitesimally small paths and reduce the integral to a differential form.

We will be interested in infinitesimal times $t - t' = \varepsilon$. The propagator $G(x, t; x', t')$ from Equation 5.401 provides

$$G(x, t; x', t') = \langle x | e^{\frac{i\hat{H}(t-t')}{\hbar}} | x' \rangle = \sqrt{\frac{m}{2\pi i \hbar \varepsilon}} e^{\frac{im\varepsilon}{2\hbar} \left(\frac{x-x'}{\varepsilon}\right)^2} e^{-\frac{i\varepsilon}{\hbar} V(\bar{x}_n)} \quad (5.406)$$

Recall that the wave function at the later time $\psi(x, t)$ is related to the wave function at the earlier time $\psi(x', t')$ by

$\langle x | \psi(t) \rangle = \langle x | \hat{U}(t - t') | \psi(t') \rangle = \int dx' \langle x | \hat{U}(t - t') | x' \rangle \langle x' | \psi(t') \rangle = \int dx' G(x, t; x', t') \psi(x', t')$ Substituting Equation 5.406 into this last equation provides the starting point for finding the SWE.

$$\psi(x, t) = \int dx' \sqrt{\frac{m}{2\pi i \hbar \varepsilon}} e^{\frac{im\varepsilon}{2\hbar} \left(\frac{x-x'}{\varepsilon}\right)^2} e^{-\frac{i\varepsilon}{\hbar} V(\bar{x}_n)} \psi(x', t') \quad (5.407)$$

For infinitesimal differences in time $t - t' = \varepsilon$ and space $x' - x = \eta$, we find

$$\psi(x, t' + \varepsilon) = \sqrt{\frac{m}{2\pi i \hbar \varepsilon}} \int d\eta e^{\frac{im\varepsilon}{2\hbar} \left(\frac{\eta}{\varepsilon}\right)^2} e^{-\frac{i\varepsilon}{\hbar} V\left(x + \frac{\eta}{2}\right)} \psi(x + \eta, t') \quad (5.408)$$

The integral can be used to show that ε small means that η must be small since otherwise the phase would rapidly vary and the integral would average to zero.

Now we can start with Equation 5.408 to reproduce the SWE. Expanding the exponential and the wave function in η yields

$$\psi(x, t' + \varepsilon) = \sqrt{\frac{m}{2\pi i\hbar\varepsilon}} \int d\eta e^{\frac{im\varepsilon}{2\hbar} \left(\frac{\eta}{\varepsilon}\right)^2} \left[1 - \frac{i\varepsilon}{\hbar} V\left(x + \frac{\eta}{2}\right) \right] \left[\psi(x, t') + \eta \frac{\partial}{\partial x} \psi(x, t') + \frac{\eta^2}{2} \frac{\partial^2}{\partial x^2} \psi(x, t') \right]$$

Expanding the potential and keeping lowest orders gives

$$\psi(x, t' + \varepsilon) = \sqrt{\frac{m}{2\pi i\hbar\varepsilon}} \int d\eta e^{\frac{im\varepsilon}{2\hbar} \left(\frac{\eta}{\varepsilon}\right)^2} \left[1 - \frac{i\varepsilon}{\hbar} V(x) \right] \left[\psi(x, t') + \eta \frac{\partial}{\partial x} \psi(x, t') + \frac{\eta^2}{2} \frac{\partial^2}{\partial x^2} \psi(x, t') \right]$$

Distributing terms on the right-hand side and keeping lowest order terms gives

$$\psi(x, t' + \varepsilon) = \sqrt{\frac{m}{2\pi i\hbar\varepsilon}} \int d\eta e^{\frac{im\varepsilon}{2\hbar} \left(\frac{\eta}{\varepsilon}\right)^2} \left[\psi(x, t') - \frac{i\varepsilon}{\hbar} V(x) \psi(x, t') + \eta \frac{\partial}{\partial x} \psi(x, t') + \frac{\eta^2}{2} \frac{\partial^2}{\partial x^2} \psi(x, t') \right]$$

Evaluating the integrals over η (including a convergence factor where necessary) yields

$$\psi(x, t' + \varepsilon) = \psi(x, t') + \frac{i\hbar\varepsilon}{2m} \frac{\partial^2}{\partial x^2} \psi(x, t') - \frac{i\varepsilon}{\hbar} V(x) \psi(x, t')$$

Rearranging the equation and taking $\varepsilon \rightarrow 0$ gives

$$i\hbar \frac{\partial}{\partial t'} \psi(x, t') = i\hbar \lim_{\varepsilon \rightarrow 0} \frac{\psi(x, t' + \varepsilon) - \psi(x, t')}{\varepsilon} = \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] \psi(x, t')$$

Or, replacing the dummy variable t' with t produces the Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \psi(x, t) = \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] \psi(x, t) \quad (5.409)$$

5.18 INTRODUCTION TO QUANTUM COMPUTING

The size of electronic components and the systems continues to decrease. Thus far, these components generally obey the laws of classical physics. Inevitably, the reduced sizes will require new quantum operating principles. This translates to new operating principles for computers as well. The new principles must address and incorporate the ultimate probabilistic nature of the elementary particle.

Quantum computing is an interdisciplinary endeavor. It encompasses theoretical computer science, physics, and engineering. The reader will find a wealth of information and simulation software in the book *Explorations in Quantum Computing* by C.P. Williams and S.H. Clearwater with over 300 references.

5.18.1 TURING MACHINES

The Turing machine originated as a conceptual means to reduce mathematical proofs to a mechanical computation. The results apply to modern computers regardless of size and speed.

The classical (deterministic) Turing machine consists of a “tape” as a type of memory that moves forward and backwards across a read–write head as shown in Figure 5.83. The tape contains 0s and 1s arranged in sequential order. These bits can represent program steps or data bits. The head has

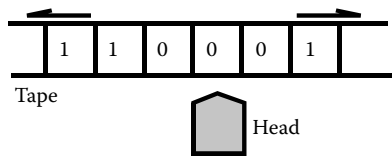


FIGURE 5.83 The classical Turing machine.

the responsibility to interpret the “meaning” of the bits. For example, if the head is in such a “state” that it must read in 8 data bits then it will interpret the next sequence of 8 bits as data. The history of the calculations performed and the program steps executed determine the “state of the head” which gives meaning to the sequence of bits on the tape. With these machines, there is a trade-off between computational accuracy and the length of time required to perform a calculation.

The probabilistic Turing machine differs from the classical one in that the machine can produce several possible responses for a given head state and tape bit pattern. The possible responses will be controlled by a probability function. For example, if the head is in state 1 and the tape has bit pattern X , then the head might write bit pattern Y or Z depending on the probability. Machines can be defined for which the head state or direction of tape travel also depends on a probability. Basically, the result from the machine represents a possible path through a calculation as controlled by a probability distribution. The resulting state of the head (etc.) will be a probability that is related to the probabilities for all possible past states. However, only one path is actually followed which distinguishes this machine from its quantum counterpart. Any problem solvable on the probabilistic Turing machine can also be solved on the classical one (and vice versa).

The quantum Turing machine replaces the “bit” with “quantum bits” (qbits). The qbits most often represent quantum properties that can assume two possible configurations although an observable with any number of discrete states will work. For the purposes of this chapter, we envision the qbit as representing the up or down spin of an electron confined to a trap. When the electron occupies the “up state” denoted by $|0\rangle$ then this will correspond to a logical 0 or false. The down state, denoted by $|1\rangle$, represents the logical 1 or true. The bits can encode a range of values between 0 and 1 since the actual quantum mechanical state of the spin particle can have the form $|\psi\rangle = \beta_0|0\rangle + \beta_1|1\rangle$ where β_i represents a complex number.

The original quantum Turing machine considered the head to be interacting with a given qbit for a fixed period of time but leaves it in a collapsed state (i.e., in one of the basis states $|0\rangle$ or $|1\rangle$). The quantum Turing machine attempts to use the fact that an electron will sample all possible trajectories through Hilbert space similar to the idea behind the Feynman path integral only applied to spin space in this case. Therefore, the particle reaches time t bearing the influence of all possible paths represented by a superposition of basis states. Making an observation forces the particle wave function to collapse to one of the basis states with a probability determined by its history. This process does not have any classical analog (Figure 5.84).

In the section on the relation between linear algebra and quantum theory, we discuss the collapse of the wave function. The quantum mechanical system without outside influences and observers evolves according to the dynamics in the Schrödinger equation. This evolution causes the system,

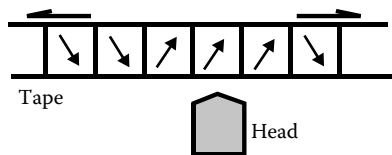


FIGURE 5.84 Bennett’s original quantum Turing machine replaces bits with quantum bits characterized by a 2-D Hilbert space.

perhaps initially in an energy basis state, to evolve to some superposition of the basis states. We view the particle as simultaneously in these states. Making an observation on the system causes the wave function to instantaneously and randomly collapse to one of the basis sets without following the evolution described by the Schrödinger equation. Making such an observation is the same as “checking the answer” from the computer. So long as we do not check for an answer, the quantum computer can be reversed at any time since the evolution operator $\hat{U} = e^{(i\hat{H}t/\hbar)}$ is unitary so that

$$|\psi(t)\rangle = \hat{U}|\psi(0)\rangle \quad \Leftrightarrow \quad |\psi(0)\rangle = \hat{U}^\dagger |\psi(t)\rangle \tag{5.410}$$

The original Turing machine only allows the qbit to evolve according to the evolution operator only during the time that the head interacts with it. Therefore, this machine could not make full use of the ability of the electron to make large superpositions with many different qbits.

5.18.2 BLOCK DIAGRAMS FOR THE QUANTUM COMPUTER

We now fix our ideas on how a quantum computer might physically appear. In classical computers, logic gates have an input and an output. The input signal might come from a register of bits. The output usually goes to a separate location as transformed bits. Applying this classical view to the quantum gate results in Figure 5.85. In this case, the gate transforms the qbit into another separate qbit. Several present designs for the quantum computer do not allow for this capability. In fact, a register of qbits might be pictured as a series of electrons confined to traps. The quantum computer has an input starting with this register of qbits and an output ending with these qbits. A scheme similar to Figure 5.85 might become viable for the quantum computer if the teleportation technology becomes viable (refer to the next section). This technology might one-day be able to extract all of the quantum information from a particle, modify and transmit the information through a quantum gate, and reconstruct the state at a new location.

For now, we use a register consisting of spin particles and design a Hamiltonian to evolve the spins. The Hamiltonian represents the program. An interaction begins at $t = 0$ and qbits evolve in time according to the evolution operator

$$\hat{U}(t) = e^{\frac{i\hat{H}t}{\hbar}} \tag{5.411}$$

This form of the evolution operator requires a closed system. A time-independent Hamiltonian therefore represents a type of “hardwired” gate. In order to change the programming, the Hamiltonian would need to depend on time and the evolution operator would use the time-ordered product discussed in the quantum mechanical representation theory.

The Feynman processor uses a closed system. The design starts with logic operations. In order to determine when to stop the processor, the register of qbits is divided into two sections. The r-qbits make up the data and the p-qbits serve as a program step counter. The r-qbits (r =register, number of bits= r) store the data and interact with the processor in parallel fashion. The p-qbits (p =program counter, number of bits= $p=k+1$) keep track of the number of steps that the computer has executed. The number of p-qbit corresponds to the number of “gates” in Figure 5.86 (plus one). When the cursor resides in the $k+1$ qbit then the calculation is complete.

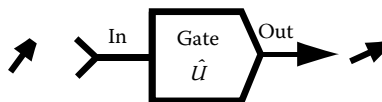


FIGURE 5.85 Classical view of a quantum gate.

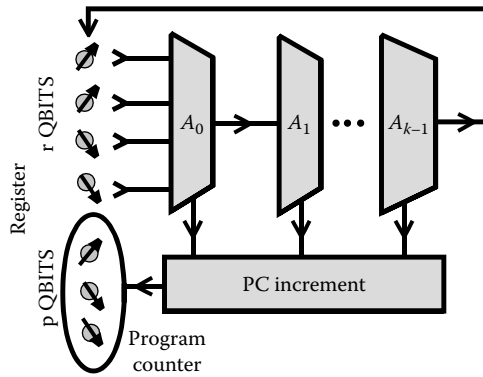


FIGURE 5.86 Idea behind the Feynman processor. In actuality, the depicted gates are part of the Hamiltonian. The evolution operator actually operates on the register.

The Feynman computer cannot be reprogrammed once the circuitry has been set since it uses the time-independent Hamiltonian. Figure 5.86 sets the basic computer architecture. Once having decided on the computation to be performed, the basic block diagram can be laid out using quantum gates. The machine performs the function $\hat{A}_{k-1}\hat{A}_{k-2} \cdots \hat{A}_1\hat{A}_0$. Next, the Hamiltonian and evolution operator can be calculated. For a closed system, the product is implemented using a Hamiltonian of the form

$$\hat{H} = \frac{1}{2} \sum_{i=0}^{k-1} \left[\hat{a}_{i+1}^+ \hat{a}_i \hat{A}_i + (\hat{a}_{i+1}^+ \hat{a}_i \hat{A}_i)^+ \right] \quad (5.412)$$

where $\hat{a}^+, \hat{a}$ represent creation and annihilation operators, respectively. The adjoint operator appears in Equation 5.412 to ensure the Hamiltonian is Hermitian. As will become clear below, each operator can act on a separate Hilbert space and therefore the products must be direct products. The creation and annihilation operators change the state of the program counter. Once knowing the number of gates, the number p-qbits can be determined. Once the mechanics have been built, one can initialize the data in the r-qbits (i.e., memory register) and let the computer run. We periodically check the p-qbits until the $(k+1)$ qbit sets and we then read off the answer from the memory register.

Alternate version of the quantum computer can be envisioned. One radically different model uses the Feynman path integral for coordinate space rather than for the configuration space used above. A person might imagine an electron entering a region of space with a number of obstacles. The Feynman path integral indicates that the electron arriving at the output of the box, must carry with it information from all possible paths through the box. By an appropriate choice of “innards” (i.e., interactions), the resulting electron will carry the results of a computation. One advantage of this scheme would be that the “box” could be reduced to 100s of Angstroms and the computer would have separate inputs and outputs.

The following sections continue the Feynman computer.

5.18.3 MEMORY REGISTER WITH MULTIPLE SPINS

In this section, we model the memory qbits after two-state spin but realize that memory can be implemented using any number quantized levels. Rather than use the notation of $|1\rangle, |2\rangle$ for spin up

and spin down, we use $|0\rangle, |1\rangle$ as a reminder of logic 0 and logic 1, respectively. The superposition wave function has the form

$$\begin{aligned} |\psi^{(1)}\rangle &= \beta_0^{(1)}|0\rangle^{(1)} + \beta_1^{(1)}|1\rangle^{(1)} \rightarrow \underline{\psi}^{(1)} = \begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} \\ |\psi^{(2)}\rangle &= \beta_0^{(2)}|0\rangle^{(2)} + \beta_1^{(2)}|1\rangle^{(2)} \rightarrow \underline{\psi}^{(2)} = \begin{pmatrix} \beta_0^{(2)} \\ \beta_1^{(2)} \end{pmatrix} \end{aligned}$$

The linear algebra shows that the direct product of the two wave functions has the form

$$|\psi^{(1)}\rangle \otimes |\psi^{(2)}\rangle = \beta_0^{(1)}\beta_0^{(2)}|0\rangle^{(1)}|0\rangle^{(2)} + \beta_1^{(1)}\beta_0^{(2)}|1\rangle^{(1)}|0\rangle^{(2)} + \beta_0^{(1)}\beta_1^{(2)}|0\rangle^{(1)}|1\rangle^{(2)} + \beta_1^{(1)}\beta_1^{(2)}|1\rangle^{(1)}|1\rangle^{(2)}$$

which produces the matrix

$$\underline{\psi} = \underline{\psi}^{(1)} \otimes \underline{\psi}^{(2)} = \begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} \otimes \begin{pmatrix} \beta_0^{(2)} \\ \beta_1^{(2)} \end{pmatrix} = \begin{pmatrix} \beta_0^{(1)}\beta_0^{(2)} \\ \beta_0^{(1)}\beta_1^{(2)} \\ \beta_1^{(1)}\beta_0^{(2)} \\ \beta_1^{(1)}\beta_1^{(2)} \end{pmatrix}$$

The basis vectors for the direct product space becomes

$$\begin{aligned} |00\rangle &= |0\rangle^{(1)}|0\rangle^{(2)} \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} & |01\rangle &= |0\rangle^{(1)}|1\rangle^{(2)} \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \\ |10\rangle &= |1\rangle^{(1)}|0\rangle^{(2)} \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} & |11\rangle &= |1\rangle^{(1)}|1\rangle^{(2)} \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \end{aligned}$$

In general, can write a sequence of memory qbits as $|011010001\dots\rangle$ where each location in the ket corresponds to a different spin particle. We anticipate the basis vector $|\dots b_3b_2b_1b_0\rangle$ produces a 1 in location $\sum_{n=0}^{N-1} 2^n b_n = \dots 2^3 b_3 + 2^2 b_2 + 2^1 b_1 + 2^0 b_0$.

For multiple spins that interact with each other (or other multiple systems that interact with each other), the wave function becomes a coherent state that cannot be factored. Any measurement will destroy the state. These are entangled states. Classical computing does not incorporate this feature.

5.18.4 FEYNMAN COMPUTER FOR NEGATION WITHOUT A PROGRAM COUNTER

One of the simplest examples of the Feynman computer calculates the “negation” of an input bit as shown in Figure 5.87. For this example, we do not include the program counter in order to make the computation. We compute the negation of a single qbit initially assumed to be in a zero state $|0\rangle$ corresponding to spin up. The book *Explorations in Quantum Computing* by C.P. Williams and S. H. Clearwater discusses the case of $\sqrt{\text{NOT}}$ as a purely quantum mechanical operation and provides references for Feynman’s two bit adder.

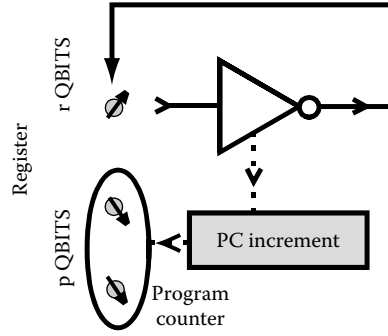


FIGURE 5.87 The Feynman processor for calculating the “NOT” of a qbit.

The “not” operator has the form $\hat{N} = |1\rangle\langle 0| + |0\rangle\langle 1|$ which should be recognized as the Pauli x -component spin operator $\hat{\sigma}_x$. The Hamiltonian in Equation 5.412 reduces to

$$\hat{H} = \frac{1}{2}(\hat{\sigma}_x + \hat{\sigma}_x^+) = \hat{\sigma}_x \quad (5.413)$$

since we only need the single “NOT” gate defined by $\hat{\sigma}_x$ which is already Hermitian. We do not include Planck’s constant $\hbar$. The unitary operator in Equation 5.411 becomes

$$\hat{U}(t) = e^{-i\hat{H}t} = e^{-i\hat{\sigma}_x t} \quad (5.414a)$$

We can see that this operator rotates an “up” spin to a “down” spin in Hilbert space by making a Taylor series expansion and using

$$\underline{\sigma}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad (5.414b)$$

Expanding the evolution operator gives

$$\underline{U}(t) = e^{-i\underline{\sigma}_x t} = 1 + \frac{(-i)}{1!}\underline{\sigma}_x t + \frac{(-i)^2}{2!}\underline{\sigma}_x^2 t^2 + \frac{(-i)^3}{3!}\underline{\sigma}_x^3 t^3 + \dots$$

Next, separate the real and imaginary parts and note

$$\underline{\sigma}_x^n = \begin{cases} \underline{1} & n = \text{even} \\ \underline{\sigma}_x & n = \text{odd} \end{cases} \quad (5.415)$$

to find

$$\underline{U}(t) = e^{-i\underline{\sigma}_x t} = \underline{1} \left(1 - \frac{1}{2!}t^2 + \dots \right) - i\underline{\sigma}_x \left(\frac{1}{1!}t - \frac{1}{3!}t^3 + \dots \right) = \underline{1} \cos(t) - i\underline{\sigma}_x \sin(t) \quad (5.416)$$

Now if we could monitor the progress of the interaction, we would find that near $t = \pi/2$

$$\underline{U}\left(\frac{\pi}{2}\right) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -i\underline{\sigma}_x \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -i \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -i \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (5.417)$$

which shows that the qbit is inverted apart from an unimportant phase factor $-i$.

We can show how this inverter can be physically implemented. The discussion of spin from Chapter 5 shows the spin Hamiltonian has the form

$$\hat{\mathcal{H}}_s = -\frac{q}{m}\vec{B} \cdot \hat{\vec{S}} = \mu_B \vec{B} \cdot \hat{\vec{\sigma}}$$

The x Pauli spin matrix needs to appear in the unitary operator (Equation 5.416), so choose the magnetic field to point along the positive x -direction

$$\hat{\mathcal{H}}_s = \mu_B \vec{B} \cdot \hat{\vec{\sigma}} = \mu_B B_x \hat{\sigma}_x \quad (5.418)$$

The unitary operator becomes

$$\hat{U}(t) = e^{\frac{\hat{\mathcal{H}}_s t}{\hbar}} = \exp\left\{\frac{\mu_B B_x \hat{\sigma}_x t}{i\hbar}\right\} \quad (5.419)$$

which uses a physical Hamiltonian and so the expression must use Planck's constant. Expanding the exponential using the results from Equation 5.416 with

$$t \rightarrow \frac{\mu_B B_x t}{\hbar}$$

produces

$$\underline{U}(t) = \underline{1} \cos\left(\frac{\mu_B B_x t}{\hbar}\right) - i\hat{\sigma}_x \sin\left(\frac{\mu_B B_x t}{\hbar}\right) \quad (5.420)$$

When

$$\frac{\mu_B B_x t}{\hbar} = \frac{\pi}{2} \rightarrow t = \frac{\pi\hbar}{2\mu_B B_x} \quad (5.421)$$

we find that the spin has flipped. Notice that we can control the rate at which the spin flips by adjusting the magnitude of the magnetic field.

Figure 5.88 shows why the magnetic field B_x causes the spin to flip. The external magnetic field produces a torque on the spin particle in order to align the two magnetic fields. The Hamiltonian does not include any damping. From a classical point of view, the spin will overshoot the lowest energy configuration and point downward at the time given in Equation 5.421. If left to itself, the spin would return to its original configuration. The process explains the sine and cosine in Equation 5.420.

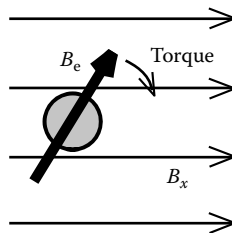


FIGURE 5.88 The external field causes the spin to flip.

5.18.5 EXAMPLE PHYSICAL REALIZATIONS OF QUANTUM COMPUTERS

We now very briefly summarize several physical implementations of quantum computers and logic gates. The interested reader can find in-depth information in the Nielsen and Chuang book *Quantum Computation and Quantum Information*, published by Cambridge University Press in 2000. An abbreviated version appears in the Willams and Clearwater book *Explorations in Quantum Computing*, published by Springer in 1997. Also check the references in these books. We briefly present the heteropolymer-based, ion-trap based, QED-based, and NMR-based computers.

The heteropolymer-based computer uses an array of atoms for the memory register. The atoms have three levels as shown in Figure 5.89. The ground state $|0\rangle$ is stable. The highest state $|2\rangle$ decays rapidly to either the ground state $|0\rangle$ or to the metastable first excited state $|1\rangle$. A pulse of light with center optical frequency of ω_{02} will transition an electron to state $|2\rangle$. The excited electron can decay to either state $|0\rangle$ or state $|1\rangle$. This three-level arrangement is actually considered to be two levels since $|2\rangle$ decays so rapidly.

Adjacent atoms (say A,B) affect the energy levels of each other through an electric dipole interaction. Figure 5.90 shows how the state of atom B affects the states of atom A . The notation $|A/B\rangle$ refers to the state of A given the state of B . Notice the state of atom B shifts the energy of $|0\rangle$ and $|1\rangle$ with respect to $|2\rangle$. The energy difference between $|0\rangle$ and $|1\rangle$ is smaller for $B = 1$ than for $B = 0$ for Figure 5.90. The frequency of the light required to induce a transition from level $|a/b\rangle$ to $|a'/b\rangle$ is denoted by $\omega_{a'b}^{B=b}$. Notice that the transition a to the frequency of light controls the operation of the device and represents the program. For example, we can make a controlled inverter. Suppose $B = 1$ then an electron in state $A = 0$ will make a transition to $A = 1$ when $\omega = \omega_{02}^{B=1}$. However, if $B = 0$ then the same process cannot occur. The sequence of pulses determines the overall function of the computer.

The ion-trap computer uses lasers to excite atoms in a well. NIST made the wells from RF waves rather than atomic barriers. These wells have parabolic shape and the well levels (restricted to 2) can encode a qbit. Additionally NIST encoded a second qbit in the energy levels of the valence electrons. The scheme worked 90% of the time. Interaction between neighboring atoms can produce

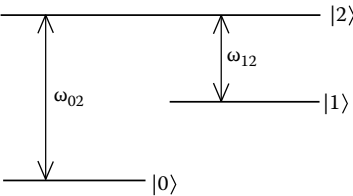


FIGURE 5.89 The three-level atom with the angular frequency ω given by the relation $E = \hbar\omega$.

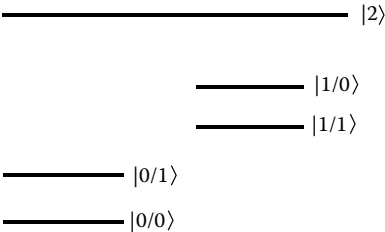


FIGURE 5.90 The energy levels for atom A given the state of atom B . The symbol “/” represents “given.” Notice the four short lines refer to atom A and shows that the spacing between states 0, 1 depend on the state of atom B (not shown).

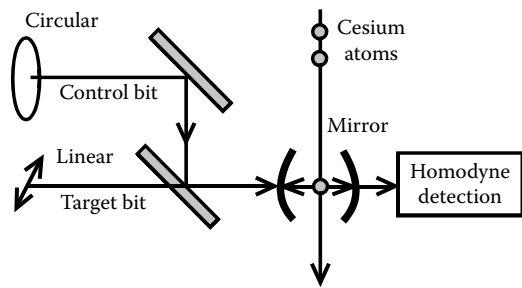


FIGURE 5.91 A block diagram of the QED-based computer.

a type of bus to carry the quantum information from one location to another. Other groups have considered a range of atoms and have discovered Yb would have a long enough lifetime to factor 385 bits.

The Cal-Tech QED-based (photonic) computer implements an XOR function. Figure 5.91 shows the gate. The target bit consists of linearly polarized photons, which can be decomposed into right and left circularly polarized components. The control bit is circularly polarized. On average, only a single control bit, target bit, and cesium atom occupy the cavity at any time. The cavity resonant frequency matches the cesium transition energy and the energy of the two photons. The control and target bits interact with a cesium atom in a cavity. The phase of the shift of one component of the linearly polarized target bit depends on the atomic excitation and upon the polarization (right or left) of the control photon.

The nuclear magnetic resonance (NMR) computer uses the spin of the nucleus. The large number of nuclei in a molecule along with the large number of molecules means that the answer occurs as an ensemble average. The state of the nucleus can be read-out by observing an the NMR spectrum. The shift of the resonance peak corresponds to a change of state in the spin.

5.19 INTRODUCTION TO QUANTUM TELEPORTATION

Science fiction depicts teleportation as a method of deconstructing an object, transmitting it as a form of RF or light waves, and reconstructing it again at a distant location. However, here we transmit qbits of information but not the physical particle itself. This is especially astonishing since any observation of a particle storing the qbit must cause the wave function to collapse and the observer would not know the exact qbit from the single measurement. Teleportation allows the full original qbit as a superposition to be reconstructed at a distant location. It opens the way for a quantum computer to operate on a qbit of information and move it through a distance after possibly performing an operation.

We first examine Bell’s theorem that draws a distinction between the classical and quantum worlds. It gives a condition that can be checked as to whether the physical world conforms to a local versus nonlocal theory.

5.19.1 LOCAL VERSUS NONLOCAL

Until the 1960s, physics was based on the notion of a “local” universe. This means that an action must have some cause in the immediate vicinity. For example, gravity exerts an influence on a nearby mass through the gravitational force at the position of the mass. Modern physics postulates the existence of gravitons that mediate the gravitational force between two masses. In this view, the direct interaction of the graviton with the mass at the location of this mass produces the force. Similarly, the electric field produces a force on a charge by virtue of photons. In either case, we

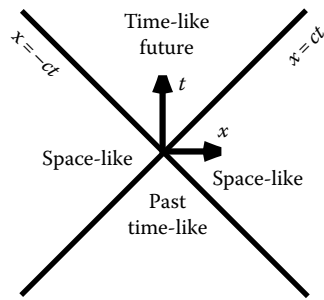


FIGURE 5.92 The light-cone with the vertex at $x=0=t$.

often envision the lines of force as radiating from one object to another. The contact of the object with the force-lines produces a force.

The theory of special relativity divides space-time into two regions, namely, the time-like and space-like regions. The regions come from the fact that a signal cannot travel faster than the speed of light. Consider a single spatial dimension x and a source of disturbance situated at $x=0$. What points x could possibly experience the disturbance at time t ? The maximum possible rate the distance could move away from $x=0$ must be the speed of light c so that $x=\pm ct$ gives the maximum possible distance the effects of the disturbance could move. The time-like region $-ct \leq x \leq +ct$ in Figure 5.92 marks the space-time position of events that can be causally related to an event occurring at $x=0=t$. The speed of light limits the slope of any path followed by a particle or a signal from an event. The space-like regions cannot be casually related since signals of any kind cannot reach the points there without exceeding the speed of light.

The “locality” of the universe requires the cause to be at the position of the event. The cause can only be the effect of another cause so long as they fall within the time-like portion of the light cone. We next set up a situation whereby two correlated particles separate and occupy positions within each others space-like region. The collapse of the wave function for one particle produces a collapse for the other. Apparently the collapse connects the two space-like points. This means that some type of disturbance traveled faster than the speed of light. Physical signals do not behave this way. Furthermore, the interaction must be nonlocal since the cause does not appear to have a physical intermediary.

5.19.2 EPR PARADOX

Einstein, Podolsky, and Rosen (EPR) posed a thought experiment in an attempt to show that the quantum theory does not fully describe nature; they expected that some variables must be hidden.

Suppose a source produces two electrons (or photons) with correlated spin as shown in Figure 5.93. The source puts the electrons in an entangled (i.e., nonseparable) state given by

$$|\psi\rangle = \frac{|01\rangle - |10\rangle}{\sqrt{2}} \tag{5.422}$$

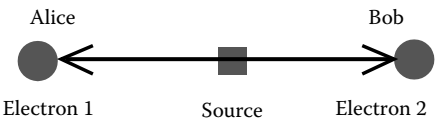


FIGURE 5.93 A source produces correlated electrons.

We cannot say that electron 1 has spin up or down and the same for electron 2 because this last equation cannot be separated into distinct states for the two particles. We can say that if electron 1 is found in state $|0\rangle$ (say spin up) then electron 2 must be in state $|1\rangle$ (spin down) as shown by the $|01\rangle$ ket in Equation 5.422. Similarly, ket $|10\rangle$ indicates that electron 1 occupies state $|1\rangle$ and therefore electron 2 occupies state $|0\rangle$.

The source sends the two electrons far across space, say several light years. During this time, the electron states remain entangled. According to the quantum theory, a measurement of the spin state of particle 1 causes the wave function to collapse. The effect instantly travels across space so that particle 2 must be in a collapsed state. Observing electron 1 in say $|0\rangle$ immediately forces electron 2 into $|1\rangle$.

EPR objected to this effect on the basis of special relativity. They claimed that the two electrons could not coordinate their collapse since it would require a signal to travel faster than the speed of light. From their point of view, the source places electrons 1 and 2 into motion with predefined spin. When observer 1, Alice, makes a measurement of electron 1, she finds the predetermined state of the electron. If the source placed electron 1 in state $|0\rangle$ then naturally electron 2 must be in state $|1\rangle$. In this way, a signal does not need to travel faster than light and we do not need to worry that the collapse of the wave function is anything more than a mathematical artifact.

Bell came up with an argument that shows the conditions under which the quantum interpretation is correct. Later, a number of researchers showed that the quantum interpretation was in fact the best explanation.

5.19.3 BELL'S THEOREM

A variety of versions of Bell's theorem have been developed. A large number use optical polarizers and rotation angles to calculate probabilities. These developments provide greater physical intuition and show a range of values for which the classical theory fails. However, we only need one such value to indicate physical reality is not local. In its most basic form, Bell's theorem is a simple math-only proof regarding probability. The theorem implicitly assumes locality and independent events. The genius of the work comes from comparing the results with the predictions of quantum theory.

Suppose we have four classical random variables A, B, C, D where Alice deals with A, B and Bob deals with C, D . Further assume that these random variables can only have values of ± 1 . Consider the sum of products

$$AC + BC + BD - AD = (A + B)C + (B - A)D \quad (5.423)$$

Since $A, B = \pm 1$ then either $(A + B)C = 0$ or $(B - A)D = 0$ but not both. Therefore the sum of products must have the value

$$AC + BC + BD - AD = \pm 2 \quad (5.424)$$

and hence, the expected value of the sum of products must satisfy

$$\langle AC \rangle + \langle BC \rangle + \langle BD \rangle - \langle AD \rangle = \langle AC + BC + BD - AD \rangle \leq +2 \quad (5.425)$$

Now compute the same quantity in a quantum setting. Assume the two electrons live in the entangled state in Equation 5.423. Identify the following observables

$$A = \hat{\sigma}_z^{(1)} \quad B = \hat{\sigma}_x^{(1)} \quad C = (-\hat{\sigma}_z^{(2)} - \hat{\sigma}_x^{(2)})/\sqrt{2} \quad D = (\hat{\sigma}_z^{(2)} - \hat{\sigma}_x^{(2)})/\sqrt{2} \quad (5.426)$$

where $\hat{\sigma}_x, \hat{\sigma}_z$ represent the Pauli spin operators for the x - and z -directions, and the superscripts refer to observer 1, Alice, and to observer 2, bob. When a measurement is made of any of the quantities A, B, C, D , the wave function collapses to one of the eigen vectors for the respective operator and gives a value of ± 1 . Furthermore we can see that

$$\langle AC \rangle = \frac{1}{\sqrt{2}} \quad \langle BC \rangle = \frac{1}{\sqrt{2}} \quad \langle BD \rangle = \frac{1}{\sqrt{2}} \quad \langle AD \rangle = \frac{-1}{\sqrt{2}} \quad (5.427)$$

The first relation, for example, in Equation 5.427 comes from

$$\langle AB \rangle = -\langle \psi | \frac{\hat{\sigma}_z^{(1)}(\hat{\sigma}_z^{(2)} + \hat{\sigma}_x^{(2)})}{\sqrt{2}} | \psi \rangle = -\frac{1}{2\sqrt{2}} (\langle 01 | - \langle 10 |) [\hat{\sigma}_z^{(1)}\hat{\sigma}_z^{(2)} + \hat{\sigma}_z^{(1)}\hat{\sigma}_x^{(2)}] (|01 \rangle + |10 \rangle)$$

with quantities of the form $\hat{\sigma}_z^{(1)}|01 \rangle = +1|01 \rangle$ since $\hat{\sigma}_z^{(1)}|0 \rangle = +1|0 \rangle$ etc. Combining the terms in Equation 5.427 produces

$$\langle AC \rangle + \langle BC \rangle + \langle BD \rangle - \langle AD \rangle = \langle AC + BC + BD - AD \rangle = +2\sqrt{2} \quad (5.428)$$

Clearly, the quantum theory does not reproduce the results of the classical theory as can be seen by comparing Equations 5.428 and 5.425. The scientist named Aspect experimentally verified the discrepancy. We conclude that either the observables do not have well defined values or there exists an element of nonlocality.

5.19.4 QUANTUM TELEPORTATION

Suppose Alice wants to send a qbit in an arbitrary superposition state to Bob. We might as well assume the qbit is encoded on a spin particle such as an electron. This would be somewhat equivalent to having a computer backplane that transports qbits around a computer or perhaps a signal transport system for communications around the country. Unfortunately, if Alice has a single qbit then any measurement will cause the superposition to collapse and Alice will only observe a single value and not the entire superposition. She will only be able to transmit that single value to Bob and neither Alice nor Bob will be able to reconstruct the original qbit.

Suppose Alice wants to transmit a data qbit given by

$$|\phi \rangle = a|0 \rangle + b|1 \rangle \equiv \begin{pmatrix} a \\ b \end{pmatrix} \quad (5.429)$$

where

$|0 \rangle$ represents spin up and logic 0

$|1 \rangle$ represents spin down and logic 1

A method exists to transmit this qbit as shown in Figure 5.94. Alice prepares an entangled spin state with electrons 2 and 3 given by

$$|\psi_{23} \rangle = \frac{|01 \rangle - |10 \rangle}{\sqrt{2}} = \frac{|0 \rangle_2 |1 \rangle_3 - |1 \rangle_2 |0 \rangle_3}{\sqrt{2}} \quad (5.430)$$

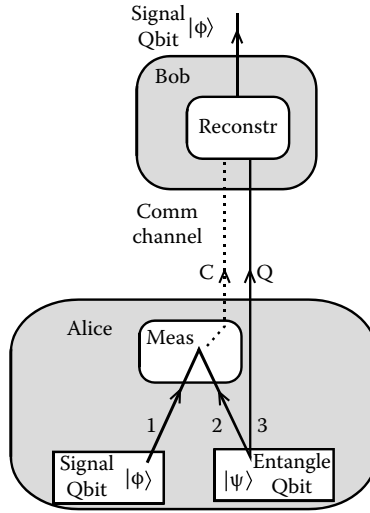


FIGURE 5.94 Setup for quantum teleportation that uses a conventional C and quantum Q communications channel.

Alice then combines electrons 1 and 2 producing the combined wave function as the direct product

$$|\psi\rangle = |\varphi\rangle \otimes |\psi_{23}\rangle$$

$$= \frac{a}{\sqrt{2}} \{ |0\rangle_1 |0\rangle_2 |1\rangle_3 - |0\rangle_1 |1\rangle_2 |0\rangle_3 \} + \frac{b}{\sqrt{2}} \{ |1\rangle_1 |0\rangle_2 |1\rangle_3 - |1\rangle_1 |1\rangle_2 |0\rangle_3 \} \quad (5.431)$$

Because she will combine electrons 1 and 2, she uses the Bell basis set defined by

$$|\psi^A\rangle = \frac{1}{\sqrt{2}} \{ |0\rangle_1 |1\rangle_2 - |1\rangle_1 |0\rangle_2 \} \quad |\psi^B\rangle = \frac{1}{\sqrt{2}} \{ |0\rangle_1 |1\rangle_2 + |1\rangle_1 |0\rangle_2 \} \quad (5.432a)$$

$$|\psi^C\rangle = \frac{1}{\sqrt{2}} \{ |1\rangle_1 |1\rangle_2 - |0\rangle_1 |0\rangle_2 \} \quad |\psi^D\rangle = \frac{1}{\sqrt{2}} \{ |1\rangle_1 |1\rangle_2 + |0\rangle_1 |0\rangle_2 \} \quad (5.432b)$$

Writing the three electron combination in Equation 5.431 in terms of the Bell basis produces

$$|\psi\rangle = \frac{1}{2} \{ |\psi^A\rangle (-a|0\rangle_3 - b|1\rangle_3) + |\psi^B\rangle (-a|0\rangle_3 + b|1\rangle_3) + |\psi^C\rangle (a|1\rangle_3 + b|0\rangle_3) + |\psi^D\rangle (a|1\rangle_3 - b|0\rangle_3) \} \quad (5.433)$$

Alice sends particle 3 (uncollapsed) to Bob via the quantum channel Q in Figure 5.94. She makes a measurement of the combined system of particles 1 and 2. The particles drop into one of the four basis vectors appearing in Equation 5.432. She then sends a conventional message to Bob via a conventional communications channel C in Figure 5.94. The message contains the name of the state in Equation 5.432 to which particles 1 and 2 collapsed. Bob has four choices for the state that particle 3 might occupy from Equation 5.433

Alice's State	State for Particle 3	Bob's Operator
$ \psi^A\rangle$	$-a 0\rangle_3 - b 1\rangle_3 \equiv \begin{pmatrix} -a \\ -b \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$ \psi^B\rangle$	$-a 0\rangle_3 + b 1\rangle_3 \equiv \begin{pmatrix} -a \\ +b \end{pmatrix}$	$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$
$ \psi^C\rangle$	$a 1\rangle_3 + b 0\rangle_3 \equiv \begin{pmatrix} +b \\ +a \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
$ \psi^D\rangle$	$a 1\rangle_3 - b 0\rangle_3 \equiv \begin{pmatrix} -b \\ +a \end{pmatrix}$	$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$

Bob uses the convention information to apply an operation to the received particle. If Alice says that particles 1 and 2 dropped to state B , then Bob applies the corresponding operation to correct the qbit and thereby reconstruct the original data qbit.

5.20 REVIEW EXERCISES

- 5.1 Normalize the following functions (i.e., find A) to make them a probability density. Note that they are not a wave function (i.e., not a probability amplitude) and therefore do not need to be squared.
- a. $y = Ae^{ax}$ for $a < 0, x \in (0, \infty)$.
 - b. $y = A\delta(x - 1) + (1 - A)\delta(x - 2), x \in (0, 3)$.
 - c. Repeat part b for $x \in (1, 2)$.
 - d. $y = A \sin(\pi x), x \in (0, 1)$.
 - e. Describe what each one looks like.
- 5.2 For each of the density functions in Problem 5.16, find $\bar{x}$.
- 5.3 Suppose an engineer has a mechanism to place an electron in an initial state defined by

$$\Psi(x, 0) = \begin{cases} x & x \in (0, 1) \\ 2 - x & x \in (1, 2) \end{cases}$$

for an infinitely deep quantum well with width $L = 2$. The bottom of the well has potential $V = 0$.

- a. Is this state normalized to have unit length? If not, normalize it.
 - b. At $t = 0$, what is the probability that the electron will be found in the $n = 2$ state?
 - c. What is the probability of finding $n = 2$ at time t ?
- 5.4 Suppose a time-independent wave function $y(x)$ is given by $y(x) = \sqrt{3}x$ for $x \in (0, 1)$ (Figure P5.4)
- a. Write a correctly normalized wave function.
 - b. What is the probability of finding an electron in the region $x \in (0, 0.5)$.

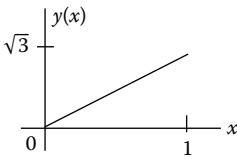


FIGURE P5.4 The wave function.

- 5.5** Find the commutator $[x, p_x^2]$.
- 5.6** Using the coordinate representation, find the Heisenberg uncertainty relations for
- The position x and x -momentum p_x .
 - The position x and y -momentum p_y .
 - The energy $\hat{H}$ and time t . *Hint:* Schrödinger's equation provides the identity $\hat{H} = i\hbar \frac{\partial}{\partial t}$.
- 5.7** Consider a superposed wave function $|\psi(t)\rangle = \sum_n \beta_n(t)|n\rangle$ where orthonormal set $\{|n\rangle\}$ spans a vector space. Suppose we multiply the wave function by the number $C = \frac{1}{2}(e^{i\alpha} - 1)$ where α is real. What values of α do not affect the probability of finding the particle in state n ?
- 5.8** Suppose $\Psi(x, t) = \sum_n C_n X_n(x) T_n(t)$ solves the SWE

$$-\frac{\hbar}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V(x)\Psi = i\hbar \frac{\partial \Psi}{\partial t}$$

where

$X_n(x)$ are stationary states

$$T_n(t) = e^{-iE_n t/\hbar}$$

Assume the collection of $X_n(x)$ form a basis set. Define $D_n(t) = C_n T_n(t)$.

- Show that the normalization $\|\Psi(x, t)\|^2 = \langle \Psi(t) | \Psi(t) \rangle = 1$ requires $\sum_n |D_n|^2 = 1$
Hint: Write $|\Psi(t)\rangle = \sum_n D_n(t) |X_n\rangle$ and use the adjoint.
 - Show $\sum_n |C_n|^2 = 1$ by using the results of part a.
- 5.9** Suppose a physical problem requires a continuous basis set $\{|\phi_k\rangle\}$. Assume

$$|\psi\rangle = \int_{-\infty}^{\infty} dk \beta_k(t) |\phi_k\rangle$$

Determine what $\langle \psi | \psi \rangle = 1$ implies about the components β_k .

- 5.10** A student has 10 exact copies of a one-particle system. She makes measurements to find the following results for the energy $E_1, E_2, E_2, E_1, E_2, E_2, E_1, E_2, E_1, E_2$. Find a wave function describing the initial system (do not use the density operator).
- Find specific probability amplitudes that will produce the 10 observations.
 - Find an expression for all possible initial wave functions assuming only two possible energy levels.
- 5.11** A student has 10 exact copies of a one-particle system. She makes measurements to find the following results for the energy $E_1, E_3, E_2, E_2, E_1, E_3, E_2, E_1, E_3, E_3$. Find a wave function describing the initial system (do not use the density operator).
- Find specific probability amplitudes that will produce the 10 observations.
 - Find an expression for all possible initial wave functions assuming only three possible energy levels.
- 5.12** A student makes measurements on a wave and finds it consists of two possible plane waves. They have the same energy but two different wave vectors $\vec{k}_1$ and $\vec{k}_2$. Assume a normalization volume V .
- Find specific probability amplitudes that will produce the two amplitudes.
 - Find an expression for all possible initial wave functions.
 - Explain why and under what conditions the specific choice of probability amplitude can have physical consequences (if it does). For example, maybe the waves can be recombined by tailoring the propagation path.
- 5.13** Show the vectors in the basis $\left\{ \phi_n(x) = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} : n = 1, 2, \dots; x \in (0, L) \right\}$ are orthonormal.

- 5.14** Electrons traveling at speed v (much slower than the speed of light) in plane wave states are incident on two very narrow, infinitely long slits separated by distance d . A phosphorus screen is located a distance $D \gg d$. Without solving Schrödinger's equation, find the probability of an electron hitting the screen a distance y from the center. Assume a wave function decrease as $1/R$ from a slit. Retain the R dependence.
- 5.15** A particle starts in the state $|\psi(0)\rangle = \int_{-\infty}^{\infty} dk C_k |\phi_k\rangle$ where $|\phi_k\rangle$ satisfies the eigenvector equation for the Hamiltonian $\hat{H}|\phi_k\rangle = E_k|\phi_k\rangle$. Show that the wave function at time t has the form $|\psi(t)\rangle = \int_{-\infty}^{\infty} dk \psi(k) e^{\frac{E_k t}{\hbar}} |\phi_k\rangle$. Hint: Consider the evolution operator and the definition of the Fourier transform.
- 5.16** A free particle starts in the state $\psi(x, 0) = \int_{-\infty}^{\infty} dk \tilde{\psi}(k) \frac{e^{ikx}}{\sqrt{2\pi}}$. Find the wave function at time t .
- 5.17** Using the definition $p = \hbar k$, rewrite the answers to Problems 5.53 and 5.54 in terms of p rather than k .
- 5.18** Consider a free particle in a plane wave state $\psi(x, 0) = \frac{e^{ikx}}{\sqrt{2\pi}}$.
- Find the wave function $\psi(x, t)$.
 - What is the Fourier transform of $\psi(x, 0)$?
 - What is the Fourier transform of $\psi(x, t)$? Keep in mind the E depends on k .
- 5.19** Consider the infinitely deep quantum well in one dimension. Show that an electron in the $n = 1$ state satisfies the Heisenberg uncertainty relation $\sigma_x \sigma_p \geq \hbar/2$.
- 5.20** Assume a particle is in a 1-D well with basis states $\{|\phi_n\rangle\}$ given in

$$\left\{ \phi_n(x) = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} : n = 1, 2, \dots; x \in (0, L) \right\}$$

- Find the average position $\bar{x}$ and average momentum $\bar{p}_x$ for each basis state.
 - Find the value of the standard deviation for x and p for each basis state.
 - What is the exact value for the Heisenberg uncertainty $\sigma_x \sigma_p$ for each basis state.
- 5.21** A student measures the position of a particle in a 1-D square well of width L and finds the value $L/2$ (i.e., the student finds the wave function collapses to the coordinate ket $|x_0\rangle = |L/2\rangle$). Using $|\psi\rangle = \sum_{n=1}^{\infty} \beta_n |\phi_n\rangle$ and the fact that projecting a wave function onto a basis state produces the probability amplitude, explain why the particle could only have been in the $n = \text{odd}$ states. Assume the states are eigenvectors of the Hamiltonian.
- 5.22** A particle is confined to an infinitely deep well. The particle is initially in the state

$$|\psi(x, 0)\rangle = \sqrt{\frac{2}{3}} |X_1\rangle + \sqrt{\frac{1}{3}} |X_2\rangle$$

where, as usual, $|X_n\rangle$ are the energy eigenfunctions satisfying $\hat{H}|X_n\rangle = E_n|X_n\rangle$ (Figure P5.22).

- A measurement is made to determine the actual energy of the particle. What is the probability of finding the particle in state X_2 ?
- What is the average value of the energy

$$\langle \hat{H} \rangle = \langle \psi(x, 0) | \hat{H} | \psi(x, 0) \rangle \text{ at } t = 0?$$

- Starting with the fact that sine waves exactly fit into the well, explain why

$$X_n(x) = \sqrt{\frac{2}{L}} \sin(k_n x) \quad k_n = \frac{n\pi}{L} \quad \text{and} \quad E_n = \frac{\hbar^2 k_n^2}{2m}$$

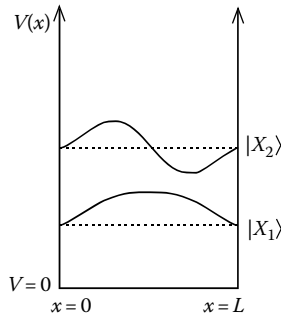


FIGURE P5.22 A quantum well.

- 5.23** An engineering student goes into the fabrication and growth facility and makes a quantum well laser with a single well of width L . Use the effective mass of the electron and hole for GaAs. Assume electrons and holes drop to the lowest possible energy levels as shown. What wavelength of light does the student find when the electron and hole recombine? (Figure P5.23)
- a. Use the infinitely deep well approximation.
 - b. Use the finitely deep well model.

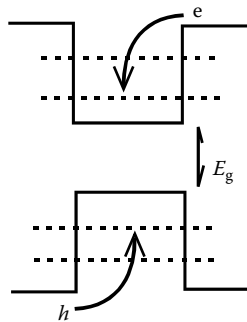


FIGURE P5.23 Electron and hole wells.

- 5.24** A student makes an “electron trap.” First the student makes a box with metallic screen (metal with many small holes). Second the student places the screen box inside a second larger screen box and prevents the two boxes from touching by installing plastic supports. The student applies a voltage between the inner and outer conductors. The student finds the interior of the screened region to have potential energy $V_1 = 0$ and the potential of the top of the well is V . The inner box has sides of length L and the outer box has sides of length L_o (Figure P5.24).
- a. Set $L = 20 \text{ \AA}$ with $L \cong L_o$, $V = 2 \text{ eV}$. Find the energy of the first allowed energy using the infinitely deep well.
 - b. Using the numbers in step a, find the first allowed energy using the finitely deep well.
 - c. For the finitely deep well, how far outside of the inner box does the wave function penetrate?
 - d. What is the ionization energy for the electron?

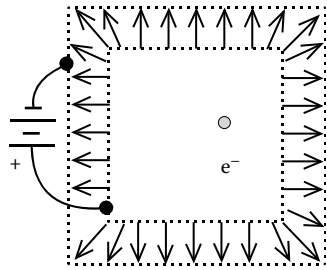


FIGURE P5.24 An electron trap.

5.25 Assume that a particle is in a 1-D well with basis states $\{|\phi_n\rangle\}$ given in

$$\left\{ \phi_n(x) = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} : n = 1, 2, \dots; x \in (0, L) \right\}$$

and an electron in the well has a wave function given by

$$\Psi(x, t) = \frac{1}{\sqrt{L}} \sin\left(\frac{\pi}{L}x\right) e^{-itE_1/\hbar} + \frac{1}{\sqrt{L}} \sin\left(\frac{2\pi}{L}x\right) e^{-itE_2/\hbar} = \frac{1}{\sqrt{2}} \phi_1 e^{-itE_1/\hbar} + \frac{1}{\sqrt{2}} \phi_2 e^{-itE_2/\hbar}$$

with $E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{\hbar^2 \pi^2}{2mL^2} n^2$.

a. By explicit calculation, find $\langle x \rangle$.

b. By explicit calculation, find σ_x^2 .

5.26 Find the general solution for a particle in a square 2-D well.

5.27 Find the general solution for a particle in a 3-D well.

5.28 Normalize the lowest order energy eigenfunction for the finitely deep well.

5.29 In Section 5.3, k_m is defined $k_m^2 = 2mV_b/\hbar^2$. Explain why this represents the maximum value of k to keep the electron in the finitely deep well.

5.30 In Section 5.3.3, draw the finite quantum well and place the energy levels in the well showing the correct relative placements for $k_m L = 15$. Determine or choose reasonable values for L , V_b , k_m and therefore reasonable values for k and E .

5.31 For the finitely deep well discussed in Section 5.3.3, show the following table

$z_m = k_m L$	$z = kL$
1	0.819
2	1.25
3	1.54
4	1.75, 3.67
5	1.89, 4.01

5.32 For the finitely deep well, find the normalization constants for the case of $z_m = 2$ using the results shown in Problem 5.82. What is the probability of finding the particle in the region $x < 0$?

5.33 Compare the energy levels for the infinitely deep and finitely deep wells for $z_m = 1, 2, \dots, 5$ (see Problem 5.31). Form the ratio of $E_{\text{finite}}/E_{\text{infinite}}$ and explain the any trends that you notice.

- 5.34** A quantum well has infinitely large potential at $x = 0$. The well has height V at $x = L$. Similar to Section 5.3, derive expressions for the energy and energy eigenfunctions.
- 5.35** Show $\hat{N}[\hat{a}^+|n\rangle] = (n + 1)[\hat{a}^+|n\rangle]$ for the Harmonic oscillator.
- 5.36** Show only integers n represent the eigenvalues for the Harmonic oscillator. *Hint:* Consider the lowering operator and a value between 0, 1.
- 5.37** Prove the classic integral relation

$$\frac{\hbar^2}{m} \int_{-\infty}^{\infty} dx u_b^* \frac{\partial u_a}{\partial x} = (E_a - E_b) \int_{-\infty}^{\infty} dx u_b^* x u_a$$

where

$$\begin{aligned}\hat{H} u_a &= E_a u_a \\ \hat{H} u_b &= E_b u_b \\ \hat{H} &= \frac{\hat{p}^2}{2m} + V(x)\end{aligned}$$

Use the following steps.

- Show $[\hat{H}, \hat{x}] = -\frac{i\hbar}{m} \hat{p}$.
- Use the results of part a to show

$$\frac{i\hbar}{m} \langle u_b | \hat{p} | u_a \rangle = -(E_b - E_a) \langle u_b | \hat{x} | u_a \rangle$$

Show why $\langle u_b | \hat{H} = \langle u_b | E_b$.

- Use the results of part b to finally prove the relation stated at the start of this exercise.
- 5.38** For the harmonic oscillator, calculate the second eigenfunction $u_2(x)$ using $\hat{a}^+$ and

$$u_1(x) = \left(\frac{\alpha}{2\sqrt{\pi}} \right)^{\frac{1}{2}} 2\alpha x e^{-\frac{\alpha^2 x^2}{2}}$$

where

$$\alpha^2 = \frac{m\omega_0}{\hbar}$$

- 5.39** Calculate $\left\langle \frac{\hat{p}^2}{2m} \right\rangle$ for a harmonic oscillator in the eigenstate $|u_n\rangle$.
Hint: Write the momentum operator in terms of the raising and lowering operators.
- 5.40** An engineering student discovers how to make a coherent electron trap. The device appears in Figure P5.40. An electron moves along a path that splits into two paths $|1\rangle, |2\rangle$ where it stays. The vectors $|1\rangle, |2\rangle$ representing the two paths are approximately orthonormal $\langle m|n\rangle = \delta_{mn}$. The position y of the path approximately obeys the relations $\hat{y}|1\rangle = 1|1\rangle, \hat{y}|2\rangle = 2|2\rangle$

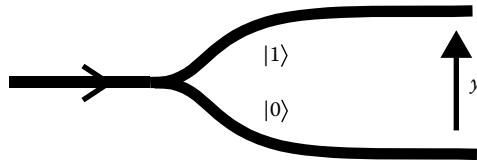


FIGURE P5.40 The coherent electron trap.

The position of path $|1\rangle$ is $y = 1$ and the position of path $|2\rangle$ is $y = 2$. We will find the average position using the density operator. There is only one wave function in the ensemble.

$$|\psi\rangle = \sqrt{\frac{3}{4}}|1\rangle + \sqrt{\frac{1}{4}}|2\rangle$$

Find the average position of the electron using the results using the density operator and the trace formula for the average.

- 5.41** An electron moves along a path located at a height $y = 0$ (Figure P.41a). The path is along the x -direction as shown in the top figure. Near $x = 0$ the electron wave divides among three separate paths at heights $y = 1$, $y = 2$, $y = 3$. Suppose each path represents a possible state for the electron. Denote the states by $|0\rangle, |1\rangle, |2\rangle, |3\rangle$ so that the position operator $\hat{y}$ has the eigenvalue equations

$$\hat{y}|n\rangle = n|n\rangle$$

The set of $|n\rangle$ forms a discrete basis. Assume that the full Hamiltonian has the form

$$\hat{H} = \frac{\hat{p}_x^2}{2m} + \hat{V} \quad \text{where } \hat{V} = mg\hat{y}$$

Further assume $\hat{p}_x|n\rangle = p_n|n\rangle$ for $x \gg 0$ or $x \ll 0$.

- a.** Use the following probabilities (at time $t = 0$) for finding the particles on the paths $x \gg 0$

$$P_1 = \frac{1}{4} \quad P_2 = \frac{1}{2} \quad P_3 = \frac{1}{4}$$

to find suitable choices for the β_n in

$$|\psi(0)\rangle = \sum_{n=1}^3 \beta_n |n\rangle$$

for the three paths $x \gg 0$. Neglect any phase factors (Figure P5.41b).

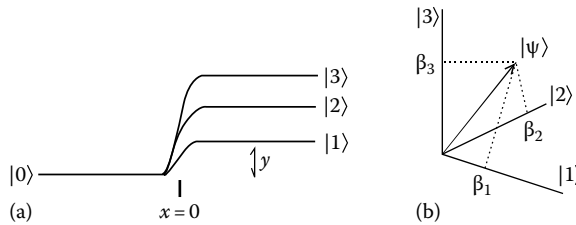


FIGURE P5.41 Electron wave divides among three paths on the right-hand side. The initial wave function.

- b.** Find the average $\langle \hat{V} \rangle = \langle \psi(0) | \hat{V} | \psi(0) \rangle$ for $x \gg 0$.
c. For $x \ll 0$, find $\langle \hat{H} \rangle$.
d. For $x \gg 0$, find $\langle \hat{H} \rangle$ in terms of n and p_n for $n = 1, 2, 3$.
e. Using the evolution operator $\hat{u}(t) = \exp \hat{H}t/(i\hbar)$, find $|\psi(t)\rangle$ for $x \gg 0$. Write the final answer in terms of n and p_n for $n = 1, 2, 3$.

- 5.42** A small perturbation is added to an infinitely deep well as shown. The bump has a small height of $\varepsilon/2$, width 2ε , and it is centered at $x=A$. Calculate the correction to energy E_1 and the original eigenvector X_1 . Keep only the lowest order terms $W_1 \cong E_1 + \langle 1|V|1 \rangle$ and $X'_1 = X_1 + H_{k1}/(E_1 - E_k)$ for $k=2$ (Figure P5.42).

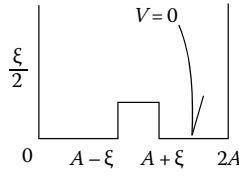


FIGURE P5.42 A well with a small bump.

- 5.43** Suppose a well replaces the bump in Figure P5.42. Find the lowest order eigenvectors and eigenvalues.
- 5.44** Consider a simple model of a heterostructure under DC bias (Figure P5.44). For the finitely deep well, suppose a voltage is applied to the well that adds the linear potential $V_{\text{Add}} = -ax$ where $a > 0$ across the entire well. To lowest order, find the new energy eigenfunctions and eigenvalues.

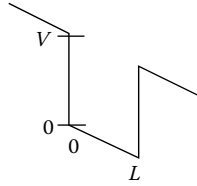


FIGURE P5.44 A linearly decreasing potential applied to the finite well.

- 5.45** Repeat the demonstrations for linear momentum based on that for the angular momentum in Section 5.6. That is, show that if the Hamiltonian is invariant with respect to translations along x , that the corresponding linear momentum p_x must be conserved.
- 5.46** Show the commutation relation for angular momentum $[\hat{L}_i, \hat{L}_j] = i\hbar\varepsilon_{ijk}\hat{L}_k$ (sum convention).
- 5.47** Show the commutation relation for angular momentum $[\hat{L}_i, \hat{L}^2] = 0$.
- 5.48** Show the commutation relation $[\hat{L}_i, \hat{r}_j] = i\hbar\varepsilon_{ijk}\hat{r}_k$.
- 5.49** Show the commutation relation $[\hat{L}_i, \hat{p}_j] = i\hbar\varepsilon_{ijk}\hat{p}_k$.
- 5.50** Show the relations for the angular momentum raising and lowering operators

$$\hat{L}_+\hat{L}_- = \hat{L}^2 - \hat{L}_z^2 + \hbar\hat{L}_z, \quad \hat{L}_-\hat{L}_+ = \hat{L}^2 - \hat{L}_z^2 - \hbar\hat{L}_z, \quad [\hat{L}_+, \hat{L}_-] = 2\hbar\hat{L}_z$$

- 5.51** Show the commutation relations for the angular momentum raising and lower operators

$$[\hat{L}_z, \hat{L}_-] = -\hbar\hat{L}_-, \quad [\hat{L}_z, \hat{L}_+] = \hbar\hat{L}_+$$

- 5.52** Show $\hat{L}_+|l, m\rangle = c_+|l, m+1\rangle$ and $c_- = \hbar\sqrt{l(l-1) - m(m+1)}$ from the chapter.
- 5.53** Show $0 \leq \langle l, m|\hat{L}_-\hat{L}_+|l, m\rangle = \hbar^2[l(l+1) - m^2 - m] \rightarrow (l + \frac{1}{2})^2 \geq (m + \frac{1}{2})^2$.
- 5.54** Start with the spherical harmonic $Y_{l=1}^{m=0}$ and find the spherical harmonics for $Y_{l=1}^{m=1}, Y_{l=1}^{m=-1}$ using the coordinate representation of the ladder operators for the z -component of angular momentum.

5.55 Show the relations for the Pauli spin operators

$$[\hat{\sigma}_i, \hat{\sigma}_j] = 2i\epsilon_{ijk}\hat{\sigma}_k, \quad \sum_i \hat{\sigma}_i^2 = 3\hat{1}$$

You might find it easiest to work with the matrices.

5.56 In a manner similar to finding the Pauli matrices for the y and z -components, derive the Pauli matrix for the x -component

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

5.57 Show for a two-particle system $[\hat{J}^2, \hat{J}_2^2] = 0$ and $[\hat{J}^2, \hat{M}] = 0$.

5.58 Does the set $\{\underline{\sigma}_x, \underline{\sigma}_y, \underline{\sigma}_z, \hat{L}\}$ form a complete set of matrices for 2×2 Hermitian matrices? Prove your answer.

5.59 Show in detail

$$|e_x\rangle \rightarrow e^{-i\theta\sigma_y/2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cos \frac{\theta}{2} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \sin \frac{\theta}{2} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

5.60 Although electrons are point particles, they still have spin angular momentum as if they rotate about their center axis. Suppose we represent spin up (i.e., along the positive z -axis) and spin down (i.e., along the negative z -axis), respectively, by the vectors/column vectors

$$\text{spin up} = |1\rangle \leftrightarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{spin down} = |2\rangle \leftrightarrow \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

(Figure P5.60). Suppose the spin exists in a superposition state

$$|\psi\rangle = |1\rangle \cos \theta + |2\rangle \sin \theta$$

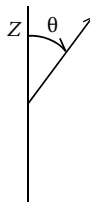


FIGURE P5.60 Spin vector at angle with respect to z -axis.

For $\theta \in (0, 90)$, what angle θ does the average spin vector make with respect to the z -axis.

5.61 For the previous problem, what is the probability of finding the electron to have spin up.

5.62 For $|\psi\rangle = \frac{1}{\sqrt{2}}[|1\rangle - i|2\rangle]$ calculate $\langle \hat{S} \rangle$ where S represents the spin.

5.63 A laboratory prepares an electron gun to produce large numbers of electrons. Assume the electrons travel along the z -axis. The electrons should have an average spin perpendicular to the direction of motion (which is the z -direction) and making an angle of 45° with respect to the x -axis. Write the wave function in complex notation.

5.64 The average spin for an electron is $\vec{S} = \frac{\hbar}{2}\hat{x}$. Find the wave function in matrix notation. Calculate $\vec{S}^2$. If the two results do not agree then explain why. Perhaps draw a picture.

- 5.65** Find the wave function that produces an average spin for an electron that makes equal angles with respect to the $+x$ -, $+y$ -, $+z$ -axes.
- 5.66** Show the spin Hamiltonian $\hat{\mathcal{H}}_s = \mu_B \vec{B} \cdot \hat{\vec{\sigma}} = \mu_B (B_x \hat{\sigma}_x + B_y \hat{\sigma}_y + B_z \hat{\sigma}_z)$ can be written in matrix notation as

$$\hat{\mathcal{H}}_s = \mu_B \begin{pmatrix} B_z & B_x - iB_y \\ B_x + iB_y & -B_z \end{pmatrix}$$

- 5.67** For a magnetic field along the z -axis $\vec{B} = B_0 \hat{z}$, find the spin wave function as a function of time assuming the spin starts in the state

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Describe the physical motion of the spin.

- 5.68** For a magnetic field along the z -axis $\vec{B} = B_0 \hat{z}$, find the spin wave function as a function of time assuming the spin starts in the state

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Describe the physical motion of the spin.

- 5.69** Show
- $\nabla^2 e^{i\vec{k} \cdot \vec{r}} = -k^2 e^{i\vec{k} \cdot \vec{r}}$ where $k^2 = k_x^2 + k_y^2 + k_z^2$.
 - $\nabla \times \hat{v} e^{i\vec{k} \cdot \vec{r}} = i\vec{k} \times \hat{v} e^{i\vec{k} \cdot \vec{r}}$ where $\hat{v}$ is a constant unit vector. Hint: Commute the operator and the unit vector.
- 5.70** Suppose $\hat{\mathcal{H}} = \frac{\hat{p}_x^2}{2m} + c$, find the Heisenberg representation of the momentum operator $\hat{p}_x$ in the x -direction where the symbol c denotes a real constant.
- 5.71** An engineering student prepares a two-level atomic system. The student does not know the exact wave function $|\psi\rangle$. After many attempts the student finds the following probability table.

$ \psi\rangle$ at $t=0$	P_ψ	where
$0.98 u_1\rangle + 0.19 u_2\rangle$	$2/3$	$\hat{H} u_1\rangle = E_1 u_1\rangle$
$0.90 u_1\rangle + 0.43 u_2\rangle$	$1/3$	$\hat{H} u_2\rangle = E_2 u_2\rangle$

- Write the density operator $\hat{\rho}(t=0)$ in a basis vector expansion.
 - What is the matrix of $\hat{\rho}(0)$?
 - What is the average energy $\langle\langle\hat{H}\rangle\rangle = \langle\hat{H}\rangle$?
- 5.72** A student is playing with a high-voltage distributor coil (30 kV) from an old car. The student is trying to make a “shock box” for a demonstration. Another student has a demonstration nearby that has excited gas molecules enclosed in a glass jar; most of the atoms have electrons in the $n=2$ excited state. The first student powers up the shock box and it emits a HUGE spark. The student notices that the nearby gas emits a photon. Assume the spark produces an electromagnetic field of the form

$$\vec{E} = \frac{E_0}{\sqrt{2\pi\sigma}} \exp\left[-\frac{t^2}{2\sigma^2}\right]$$

at the position of the atoms. The perturbation potential is then

$$\hat{V} = \hat{\mu} \cdot \vec{E} \quad \text{or} \quad \underline{V} = \underline{\mu} \vec{E} \quad \text{with} \quad V_{12} = \mu_{12} \frac{E_0}{\sqrt{2\pi\sigma}} \exp\left[\frac{t^2}{2\sigma^2}\right]$$

Find: The approximate probability of transition from state #2 to #1 given by

$$P_{2 \rightarrow 1} = |\langle u_1 | \psi(\infty) \rangle|^2$$

Hints:

a. Substitute V_{21} into $\langle u_1 | \psi(\infty) \rangle$.

b. Integrate using $\exp\left\{-\frac{\tau^2}{2\sigma^2} + i\omega_{12}\tau\right\} = \exp\left\{-\frac{\sigma^2}{2}\omega_{12}^2\right\} \exp\left\{-\frac{1}{2\sigma^2}(\tau + i\sigma^2\omega_{12})^2\right\}$

$$\int_{-\infty}^{\infty} d\tau \frac{1}{\sqrt{2\pi}b} \exp\left[-\frac{(\tau \pm a)^2}{2b^2}\right] = 1$$

You should find

$$\langle u_1 | \psi(\infty) \rangle \sim E_0 \left[\frac{\mu_{12}}{i\hbar} \right] \exp\left[-\frac{\sigma^2}{2}\omega_{12}^2\right]$$

5.73 Show the relation

$$e^{\tau \hat{A}} \hat{B} e^{-\tau \hat{A}} = \hat{B} + \tau [\hat{A}, \hat{B}] + \frac{\tau^2}{2!} [A, [\hat{A}, \hat{B}]] + \dots$$

by expanding the exponentials in a Taylor series.

5.74 Find the Heisenberg representation of the momentum operator $\hat{p}_x$ in the x -direction for the Schrödinger Hamiltonian of the form

$$\hat{H} = \frac{\hat{p}_x^2}{2m} + \hat{x}$$

5.75 Demonstrate the relation $e^{\xi \hat{A}} e^{\xi \hat{B}} = e^{\xi(\hat{A}+\hat{B})} e^{\xi^2[\hat{A}, \hat{B}]/2}$ holds so long as $[\hat{A}, [\hat{A}, \hat{B}]] = 0$, etc.

5.76 Show that the number operator is Hermitian.

5.77 By Taylor expanding the exponentials, show $e^{\xi \hat{A}} e^{\xi \hat{B}} = e^{\xi(\hat{A}+\hat{B})}$ when $[\hat{A}, \hat{B}] = 0$.

5.78 Rederive the probability of transition (to first-order approximation) using C_n in the wave function

$$|\psi(t)\rangle = \sum_n C_n(t) e^{-i\omega_n t} |n\rangle$$

Note that the C differs from the β in the chapter by the exponential.

5.79 Using the interaction Hamiltonian $\hat{\mathcal{V}} = e^{i\mathcal{V}_0}$, the adiabatic approximation, find the probability of an electron making a transition from state #1 to state #f.

5.80 Consider a two-level atom. Suppose the electron starts in the state

$$|\psi(0)\rangle = \frac{\sqrt{2}}{2} |1\rangle + \frac{\sqrt{2}}{2} |2\rangle$$

Apply an electromagnetic perturbation as given in the chapter. Determine the probability of finding the electron in state $|2\rangle$ for small times.

5.81 Rework the solutions for the probability amplitude in the case of time-dependent perturbation theory when the particle starts in states “a” and “b” equally.

5.82 The chapter discusses time-dependent perturbation theory. Using the Schrödinger representation, derive the first-order correction to β as follows.

a. Suppose $\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{V}}$ and $\hat{H}_0|n\rangle = E_n|n\rangle$. Substitute $|\psi\rangle = \sum_n \beta_n(t)|n\rangle$ into the SWE $\hat{\mathcal{H}}|\psi\rangle = i\hbar \frac{\partial}{\partial t}|\psi\rangle$ to show

$$\dot{\beta}_k - \frac{E_k}{i\hbar}\beta_k = -\frac{i}{\hbar} \sum_n \beta_n V_{kn}$$

b. For small perturbation V (i.e., make the replacement $\mathcal{V} \rightarrow 0$) to show

$\beta_k^{(0)}(t) = a_k^{(0)} e^{E_k t/(i\hbar)}$ where $a_k^{(0)}$ represents a constant of integration (independent of time). Given that the particle starts in state $|i\rangle$ at $t=0$, conclude

$$\beta_k^{(0)}(0) = \delta_{ki} \quad a_k^{(0)} = \delta_{ki} \quad \text{and} \quad \beta_k^{(0)}(t) = \delta_{ki} e^{E_k t/(i\hbar)}$$

c. Use the results of part a and the integrating factor $\mu = e^{-E_k t/(i\hbar)}$ to conclude

$$\beta_k(t) = e^{E_k t/(i\hbar)} \left\{ \beta_k(0) - \frac{i}{\hbar} \sum_n \int_0^t dt' \beta_k(t') \mathcal{V}_{kn}(t') e^{-E_k t'/(i\hbar)} \right\}$$

assuming the perturbation starts at $t=0$

d. Use the result of parts a and b to find the first-order correction

$$\beta_k^{(1)}(t) = e^{E_k t/(i\hbar)} \left\{ \delta_{ki} - \frac{i}{\hbar} \int_0^t dt' \mathcal{V}_{ki}(t') e^{E_{ik} t'/(i\hbar)} \right\}$$

e. Compare the results of part d with the results for $a_k^{(1)}$ derived in the chapter.

5.83 Show that the components of the average wave function $\text{Ave}\{|\psi\rangle\} = \sum_s P_s |\psi^{(s)}\rangle$ do not necessarily sum to 1. Consider the simplest case: Assume that each wave function lives in a 2-D Hilbert space $|\psi^{(s)}\rangle = \beta_1^{(s)}|1\rangle + \beta_2^{(s)}|2\rangle$. Consider only two wave functions for $s=1,2$. Assume all coefficients $\beta_n^{(s)}$ are real. To make the problem simpler, consider the case of $\beta_1^{(2)} = (1 + \varepsilon_1)\beta_1^{(1)}$ and $\beta_2^{(2)} = (1 + \varepsilon_2)\beta_2^{(1)}$.

a. Show that the sum of the square of the components equals 1 if and only if $\varepsilon_1 = 0 = \varepsilon_2$. Hint: Sum the squares of the coefficients of $\text{Ave}\{|\psi\rangle\}$ in the usual application of Pythagorean's theorem, collect the squared terms of P_1^2 and P_2^2 , and add terms to 1 where appropriate. You should find a result similar to

$$1 + 2P_1 P_2 \{\beta_1^{(1)2} \varepsilon_1 + \beta_2^{(1)2} \varepsilon_2\}$$

b. Explain why the diagonal components of the density operator add to 1 but the sum of the square of the components of the average wave function do not.

5.84 Consider a two-electron system with overlapping wave functions. However, assume that the lowest order Hamiltonian for each system has the form $\hat{H} = E_1|1\rangle\langle 1| + E_2|2\rangle\langle 2|$.

a. Show that the following state is normalized

$$\frac{1}{\sqrt{2}}(|1\rangle_1|2\rangle_2 - |2\rangle_1|1\rangle_2)$$

b. Show that the following two states are orthogonal

$$\frac{1}{\sqrt{2}}(|1\rangle_1|2\rangle_2 - |2\rangle_1|1\rangle_2) \quad \frac{1}{\sqrt{2}}(|1\rangle_1|2\rangle_2 + |2\rangle_1|1\rangle_2)$$

c. Find the average energy for each state.

5.85 Show that the state $|1, 0, 2, 0, 0, \dots\rangle = \frac{1}{\sqrt{3}}\{|\phi_1\rangle_1|\phi_3\rangle_2|\phi_3\rangle_3 + |\phi_1\rangle_2|\phi_3\rangle_1|\phi_3\rangle_3 + |\phi_1\rangle_3|\phi_3\rangle_2|\phi_3\rangle_1\}$ is correctly normalized. If all permutations are included, show the correct normalization must be $1/\sqrt{6}$.

5.86 Suppose the single-particle states $\{\phi_1, \phi_2, \phi_3\}$ correspond to energy $E_n = n^2$ for $n = 1, 2, 3$. Assume a system of two bosons. Find the basis states for the two-particle system.

5.87 Repeat Problem 76 for two fermions.

5.88 Starting with the fermion field commutators

$$\{\hat{\psi}(\vec{r}), \hat{\psi}^+(\vec{r}')\} = \delta(\vec{r} - \vec{r}') \quad \{\hat{\psi}(\vec{r}), \hat{\psi}(\vec{r}')\} = 0 = \{\hat{\psi}^+(\vec{r}), \hat{\psi}^+(\vec{r}')\}$$

show the commutation relations for the fermion creation and annihilation operators

$$\{\hat{f}_m(t), \hat{f}_n^+(t)\} = \delta_{mm} \quad \{\hat{f}_m(t), \hat{f}_n(t)\} = 0 = \{\hat{f}_m^+(t), \hat{f}_n^+(t)\}$$

5.89 Show

$$[N_{\Delta V}, \Psi^+(r, t)] = \begin{cases} \Psi^+(r, t) & \vec{r} \in \Delta V \\ 0 & \vec{r} \notin \Delta V \end{cases}$$

5.90 Show that the wave function

$$|\psi(1, 2, \dots, N)\rangle = \sqrt{\frac{N_a!N_b!\dots N_n!}{N!}}\{(|a\rangle_1|b\rangle_2\dots|n\rangle_N) + (|b\rangle_1|a\rangle_2\dots|n\rangle_N) + \dots\}$$

in Equation 5.322a has the correct normalization.

5.91 Show that the wave function

$$|\psi(1, 2, \dots, N)\rangle = \sqrt{\frac{1}{N!}}\{+(|a\rangle_1|b\rangle_2\dots|n\rangle_N) - (|b\rangle_1|a\rangle_2\dots|n\rangle_N) + \dots\}$$

in Equation 5.323a has the correct normalization.

5.92 Explain in detail why, in relation to the Pauli exclusion principle, bosons and fermions obey commutation and anticommutation relations, respectively. Explain in detail, the relation to the Fock states and creation–annihilation operators.

5.93 Show that the two-particle wave function for fermions has the correct normalization

$$|1, 1\rangle \sim \frac{1}{\sqrt{2}}\{\phi_{E_a}(x_1)\phi_{E_b}(x_2) - \phi_{E_a}(x_2)\phi_{E_b}(x_1)\}$$

5.94 Suppose G is a green function satisfying

$$\hat{L}G(x, t) = \delta(x)\delta(t)$$

Determine if in fact a solution to $\hat{L}\psi(x, t) = f(x, t)$ is given by

$$\psi(x, t) = \int dx' dt' G(x - x', t - t') f(x', t')$$

If it is not, what is wrong? If it is, discuss the appropriate limits of integration and whether or not this last equation is a particular or general solution.

5.95 Assume a free electron is in a plane wave state with momentum p given by $|p\rangle$. The coordinate form of a plane wave state at $t=0$ is $\langle x|p\rangle = \frac{e^{ipx/\hbar}}{\sqrt{2\pi}}$. Show $1 = \int \frac{dp}{\hbar} |p\rangle \langle p|$. *Hint:* Integrate $\langle x|p\rangle \langle p|x'\rangle$ over all p and use properties of the Dirac delta function.

5.96 Given a wave function $\psi(x, 0)$ show $\psi(x, t)$ can be written as

$$|\psi(t)\rangle = \left\{ \int_{-\infty}^{\infty} \frac{dp}{\hbar} |p\rangle \frac{e^{\frac{ipx}{\hbar} - \frac{E_p t}{\hbar}}}{\sqrt{2\pi}} \langle p| \right\} |\psi(0)\rangle = \int_{-\infty}^{\infty} \frac{dp}{\hbar} |p\rangle \frac{e^{\frac{ipx}{\hbar} - \frac{E_p t}{\hbar}}}{\sqrt{2\pi}} \langle p|\psi(0)\rangle$$

5.97 Determine if the Bell basis set is orthonormal (refer to Section 5.19)

$$\begin{aligned} |\psi^A\rangle &= \frac{1}{\sqrt{2}} \{|0\rangle_1 |1\rangle_2 - |1\rangle_1 |0\rangle_2\} & |\psi^B\rangle &= \frac{1}{\sqrt{2}} \{|0\rangle_1 |1\rangle_2 + |1\rangle_1 |0\rangle_2\} \\ |\psi^C\rangle &= \frac{1}{\sqrt{2}} \{|1\rangle_1 |1\rangle_2 - |0\rangle_1 |0\rangle_2\} & |\psi^D\rangle &= \frac{1}{\sqrt{2}} \{|1\rangle_1 |1\rangle_2 + |0\rangle_1 |0\rangle_2\} \end{aligned}$$

REFERENCES AND FURTHER READING

Quantum Theory

1. Eisberg R. and Resnick R., *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*, John Wiley & Sons, New York (1974). Comment: A lot of explanation on the wave aspects of quantum theory.
2. Park D., *Introduction to the Quantum Theory*, 2nd ed., McGraw-Hill Book Company, New York (1974). Has a discussion of the operator approach.
3. Tang C.L., *Fundamentals of Quantum Mechanics for Solid State Electronics and Optics*, Cambridge University Press, Cambridge, U.K. (2005).
4. Liboff R.L., *Introductory Quantum Mechanics*, 3rd ed., Addison-Wesley Publishing Company, Reading, MA (1997). One of the best introductory books with the operator approach.
5. Shankar R., *Principles of Quantum Mechanics*, 2nd ed., Kluwer Academic/Plenum Publishers, New York (1994). Quantum theory at the level of the present text.
6. Elbaz E., *Quantum: The Quantum Theory of Particles, Fields, and Cosmology*, Springer, Berlin (1998). Covers most topics.
7. Baym G., *Lectures on Quantum Mechanics*, Addison-Wesley, Reading, MA (1990).
8. Messiah A., *Quantum Mechanics*, Dover Publications, Mineola (1999). This is a must-have.
9. Dirac P.A.M., *The Principles of Quantum Mechanics*, Oxford at the Clarendon Press, Oxford, U.K. (1978). A great classic.
10. Sakurai J.J., *Advanced Quantum Mechanics*, Addison-Wesley Publishing Company, Reading, MA (1980).
11. Thaller B., *Visual Quantum Mechanics*, Springer-Verlag, New York (2000).
12. Pauling L. and Wilson, E.B. Jr., *Introduction to Quantum Mechanics with Applications to Chemistry*, Dover Publications Inc., New York (1963).

Density Operator

13. Blum K., *Density Matrix Theory and Applications*, 2nd ed., Plenum Press, New York (1996).

Concepts, Interpretation and Philosophy of Quantum Theory

14. Herbert N., *Quantum Reality: Beyond the New Physics*, Anchor Books, New York (1987). Easy bedtime reading.
15. Baggott J., *The Meaning of Quantum Theory*, Oxford University Press, New York (1992).
16. Albert D.Z., *Quantum Mechanics and Experience*, Harvard University Press, Cambridge, MA (1992).
17. Hughes R.I.G., *The Structure and Interpretation of Quantum Mechanics*, Harvard University Press, Cambridge, MA (1989).
18. Auyang S.Y., *How is Quantum Field Theory Possible?* Oxford University Press, New York (1995).
19. Teller P., *An Interpretive Introduction to Quantum Field Theory*, Princeton University Press, Princeton, NJ (1995).
20. Omnes R., *The Interpretation of Quantum Mechanics*, Princeton University Press, Princeton, NJ (1994).
21. Prigogine I. *From Being to Becoming: Time and Complexity in the Physical Sciences*, W.H. Freeman & Company, New York (1980).
22. Treiman S., *The Odd Quantum*, Princeton University Press, Princeton, NJ (1999).
23. Feynman R.P., QED. The Strange Theory of Light and Matter, Princeton University Press, Princeton, NJ (1985).

Optoelectronics and Quantum Optics

24. Parker M.A., *Physics of Optoelectronics*, Taylor & Francis Group, CRC Press, Boca Raton, FL (2005).
25. Coldren L.A. and Corzine S.W., *Diode Lasers and Photonic Integrated Circuits*, John Wiley & Sons, Inc., New York (1995).
26. Yariv A., *Quantum Electronics*, 3rd ed., John Wiley & Sons, New York (1989).
27. Saleh B.E.A. and Teich M.C., *Fundamentals of Photonics*, John Wiley & Sons, Inc., New York (1991).
28. Chuang S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York (1995).
29. Milonni P.W., *The Quantum Vacuum: An Introduction to Quantum Electrodynamics*, Academic Press, Inc., Boston, MA (1994). This is a great book and one of my favorites, but not in the “fast” reading category.
30. Mandel L. and Wolf E., *Optical Coherence and Quantum Optics*, Cambridge University Press, Cambridge, MA (1995). Excellent, one of the best.
31. Bachor H.A. and Ralph T.C., *A Guide to Experiments in Quantum Optics*, Wiley-VCH, Weinheim (2004).

Applications

32. Milburn G.J., *The Feynman Processor: Quantum Entanglement and the Computing Revolution*, Purseus Books, New York (1998). Easy reading.
33. Bugh G.J., *Spin Wave Technology*, Vasant Corporation (www.vasantcorporation.com), Fort Worth (2002).
34. Nielsen M.A. and Chuang I.L., *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, MA (2000). A good in-depth book accessible to readers of the present text.
35. Williams C.P. and Clearwater S.H., *Explorations in Quantum Computing*, Springer-Verlag, New York (1998).
36. Hirvensalo M., *Quantum Computing*, Springer-Verlag, Berlin (2001).

Standing Problems in Physics

37. Veltman M., *Facts and Mysteries in Elementary Particle Physics*, World Scientific, New Jersey (2003).
38. Smolin L., *The Trouble with Physics: The Rise of String Theory, the Fall of a Science, and What Comes Next*, Houghton Mifflin Company, Boston, MA (2006).

Miscellaneous

39. Mattuck R.D., *A Guide to Feynman Diagrams in the Many-Body Problem*, 2nd ed., Dover Publications, New York (1992). An easy book to read and details many of the concepts of Quantum Field Theory.

40. Fetter A.L. and Walecka J.D., *Quantum Theory of Many-Particle Systems*, McGraw-Hill, Inc., New York (1971). Check for the Dover Version.
41. Mahan G.D., *Many-Particle Physics*, 2nd ed., Plenum Press, New York (1990). A massive book.
42. Aitchison I.J.R. and Hey A.J.G., *Gauge Theories in Particle Physics*, Adam Hilger LTD, Bristol (1982). ISBN: 0-85274-534-6. This is a “must-have” book for those in more advanced Physics study especially Particle Physics.
43. Goldstein R., *Incompleteness: The Proof and Paradox of Kurt Godel*, W.W. Norton Company, New York (2002).
44. Nagel E. and Newman J.R., *Godels Proof*, New York University Press, New York (2001).

6 Solid-State: Structure and Phonons

A study of the solid-state form of matter provides the foundations for many diverse fields. It explores a wide range of concepts and tools for modern science and engineering. This chapter introduces concepts necessary for understanding and engineering state-of-the-art electronic and optoelectronic devices. The invention and development of new devices requires not only a clear understanding of present engineering and physics practice, but also sufficient theoretical background to understand new discoveries in a variety of fields.

The book has divided the effects of the regular arrays of atoms and molecules into two parts. One part as described in the present chapter relates to the mechanical effects of the arrayed atoms and molecules. In particular, this chapter discusses the vibrational motion of the array which produces phonons. The next chapter focuses on the effects of the periodic array on the conduction of electrical current. Perhaps the most technologically important effect consists of the formation of electronic bands and the associated effective mass.

Technology makes use of all forms of matter including gasses, liquids, and solids. The solids can have structure ranging from crystalline to amorphous. The study briefly describes how bonding occurs and shows how it produces a periodic structure, and the resulting concepts of Bravais lattice, reciprocal lattice, the phonon dispersion curves, and phonon distributions and specific heat.

6.1 ORIGIN OF CRYSTALS

A crystal consists of a periodic array of atoms. The bonding results from the interplay of the electronic wave functions between the atoms. These orbital wave functions also produce the periodicity of the crystal. The study of bonds underlies the field of chemistry. This section therefore specializes to the tetrahedrally bonded semiconductors such as gallium arsenide (GaAs) and silicon (Si). This section can be omitted on first reading without loss of continuity; interested readers can refer to the Tang book or the Coulson book listed in the chapter references.

6.1.1 ORBITALS AND SPHERICAL HARMONICS

The bonding and crystal structure of many technologically important materials derive their properties from the s and p orbitals. Four of the most important semiconductors have valence electrons as follows:

Si	Core + $3s^2 3p^2$
Ge	Core + $4s^2 4p^2$
Ga	Core + $4s^2 4p^1$
As	Core + $4s^2 4p^3$

The symbol mp^n refers to p orbitals corresponding to energy level $\#m$ for the radial wave function and having n electrons in that p orbital. An orbital (state) corresponds to energy for the electron. A number of mechanisms contribute to the electron energy including coulomb attraction to the

nucleus, electron orbital angular momentum, and the interaction between the magnetic field due to electron spin and the magnetic field produced from the electron orbiting the nucleus (the $\vec{L} \cdot \vec{S}$ interaction).

The s and p orbitals refer to the angular momentum states of the electron. Recall from Section 5.5

$$\begin{aligned}\hat{L}^2|\ell, \ell_z\rangle &= \hbar^2 \ell(\ell+1)|\ell, \ell_z\rangle & \ell &= 0, 1, \dots \\ \hat{L}_z|\ell, \ell_z\rangle &= \hbar \ell_z|\ell, \ell_z\rangle & \ell_z &= -\ell, -(\ell-1), \dots, (\ell-1), \ell\end{aligned}$$

where

$\hat{L}^2$ represents the squared magnitude of the angular momentum

L_z represents the z-component

The s orbital refers to the state without any orbital angular momentum $\hat{L}^2|s\rangle = 0|s\rangle$. Therefore, the s orbital does not have any z-component of angular momentum $\hat{L}_z|s\rangle = 0|s\rangle$. However the electron still has spin and hence nonzero “total” angular momentum. The $|s\rangle$ state has spherical symmetry and can be related to the uncoupled or coupled basis set for angular momentum according to

$$|s\rangle = \left| \ell = 0, \ell_z = 0, s = \frac{1}{2}, s_z = \pm \frac{1}{2} \right\rangle = \left| j = \frac{1}{2}, j_z = \pm \frac{1}{2}, s = \frac{1}{2}, s_z = \pm \frac{1}{2} \right\rangle \quad (6.1a)$$

The spherical harmonic corresponding to this state is

$$Y_{\ell, \ell_z}(\theta, \phi) = \frac{1}{\sqrt{4\pi}} \quad (6.1b)$$

This spherical harmonic appears in Figure 6.1. The radial part of the wave function provides the spherical “boundary.” In the s state, the wave function does not have any angular variation. It obviously has even parity $\hat{p}|s\rangle = +1|s\rangle$.

The p states correspond to the lowest nonzero orbital angular momentum states. The “p” does not refer to linear momentum. We use the following definitions for p_x (or X), etc.

$$|p_z\rangle = |Z\rangle = |\ell = 1, \ell_z = 0\rangle \sim Y_{10}(\theta, \phi) = \sqrt{\frac{3}{4\pi}} \cos \theta = \sqrt{\frac{3}{4\pi}} \frac{z}{r} \quad (6.2a)$$

$$|p_y\rangle = |Y\rangle = \frac{i}{\sqrt{2}} \{|\ell = 1, \ell_z = 1\rangle + |\ell = 1, \ell_z = -1\rangle\} \sim \frac{i}{\sqrt{2}} \{Y_{1,1} + Y_{1,-1}\} = \sqrt{\frac{3}{4\pi}} \frac{y}{r} \quad (6.2b)$$

$$|p_x\rangle = |X\rangle = \frac{1}{\sqrt{2}} \{|\ell = 1, \ell_z = 1\rangle - |\ell = 1, \ell_z = -1\rangle\} \sim \frac{1}{\sqrt{2}} \{Y_{1,1} - Y_{1,-1}\} = \sqrt{\frac{3}{4\pi}} \frac{x}{r} \quad (6.2c)$$

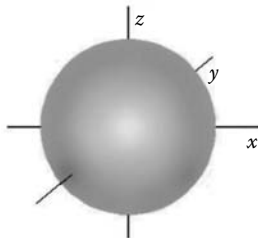


FIGURE 6.1 The spherically symmetrical s orbital. The radial wave function provides the spherical boundary.

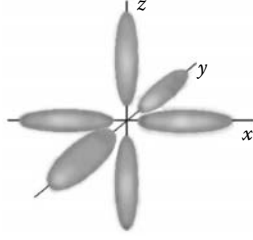


FIGURE 6.2 Each p orbital has two lobes; one along the positive and one along the negative axis. The elongation has been exaggerated for illustration purposes.

where Y_{lm} represents a spherical harmonic. We use the capital letters such as “Z” to refer to a specific angular momentum state rather than “ p_z ” to avoid confusing the orbitals with the linear momentum. Notice that the labels X, Y, Z match the x, y, z in the resulting expression. It is easy to see that the states in Equations 6.1 and 6.2 are orthonormal.

Define $\hat{P}_x$ to be the parity operator that replaces x with $-x$ (etc.). The symbols X, Y, Z denoting the orthogonal states provides useful notation for two reasons. First they refer to the direction of odd parity such as, for example, $\hat{P}_x|X\rangle = -|X\rangle$ (etc). Second they refer to the direction of the lobes illustrated Figure 6.2. The parity operator makes it easy to calculate some averages such as for the x -momentum

$$\langle X|\hat{p}_x|X\rangle = \{\langle X|-1\rangle\hat{p}_x\{-1|X\rangle\} = \{\langle X|\hat{P}_x^+\}\hat{p}_x\{\hat{P}_x|X\rangle\} = \langle X|\{\hat{P}_x^+\hat{p}_x\hat{P}_x\}|X\rangle = -\langle X|\hat{p}_x|X\rangle$$

Therefore the expectation value must be $\langle X|\hat{p}_x|X\rangle = 0$.

6.1.2 HYBRID ORBITAL

The appearance of the bonds differs from that for the unbonded atoms. First consider the unbonded case. Figure 6.1 shows the spherically symmetric s orbitals and Figure 6.2 shows the “dumbbell” shaped p orbitals before bonding. The orbitals correspond to different values of the total angular momentum and spin. The four states

$$|S\rangle, |X\rangle, |Y\rangle, |Z\rangle \quad (6.3)$$

are orthonormal. As shown in Figure 6.2, these states form 90° spatial angles with respect to each other. If these were the bonding states, then the adjacent atoms would need to have 90° angles between them rather than the approximately 110° commonly found for tetrahedral bonding.

The bonding orbitals are not the same as the unbonded ones. The bonded orbitals are linear combinations of those in Equation 6.3. Combining angular momentum states in this manner produces the sp^3 hybridization. In particular,

$$\begin{aligned} |\psi_1\rangle &= \frac{1}{2}(|S\rangle + |X\rangle + |Y\rangle + |Z\rangle) & |\psi_2\rangle &= \frac{1}{2}(|S\rangle + |X\rangle - |Y\rangle - |Z\rangle) \\ |\psi_3\rangle &= \frac{1}{2}(|S\rangle - |X\rangle + |Y\rangle + |Z\rangle) & |\psi_4\rangle &= \frac{1}{2}(|S\rangle - |X\rangle - |Y\rangle + |Z\rangle) \end{aligned} \quad (6.4)$$

The functions in Equation 6.4 are orthonormal. The spatial plot of the functions in Equation 6.4 form a tetrahedron with the bonds separated by approximately 109.5° . The hybrid orbitals in Equation 6.4 produce the face-centered cubic (FCC) unit cell with a two atom basis.

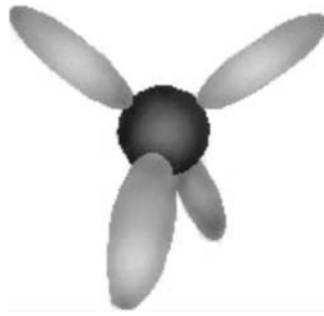


FIGURE 6.3 Combining s and p orbitals.

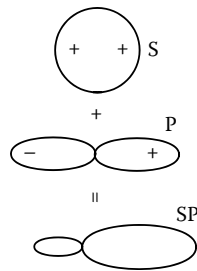


FIGURE 6.4 Hybrid orbitals. They have been artificially elongated to clarify the picture.

We can see how the hybrid states make the required angles and why we add the s orbital. First consider the angles. The states X, Y, Z have their positive lobes along the directions $\hat{x}, \hat{y}, \hat{z}$, respectively. We therefore expect the linear combination for $|\psi_1\rangle$, for example, to produce the positive lobe in the direction $\hat{x} + \hat{y} + \hat{z}$. Similarly, $|\psi_2\rangle$ must have the positive lobe along the direction $\hat{x} - \hat{y} - \hat{z}$. The dot product gives the angle between these two vectors as approximately 109.5° .

Now we can see the s orbital increases the probability of finding the electron between bonded atoms which decreases the energy of the system consisting of the electrons and atoms. These produce the bonding levels. Consider the s and P orbitals shown in Figure 6.4. The S wave function has a positive value everywhere. The P orbital has a positive and negative lobe consistent with the parity. Adding S and P together produces an oblong orbital SP. Those orbitals with the large lobe between two bonded atoms increase the stability by lowering the system energy. Notice the s state does not change the direction of the positive lobe.

The states shown in Figure 6.4 then overlap with similar states from neighboring atoms and produce the tetrahedral structure common for silicon and GaAs.

6.2 CRYSTAL, LATTICE, ATOMIC BASIS, AND MILLER NOTATION

A crystal consists of a collection of “individual atoms” or “groups of atoms” arranged as a periodic array. The “lattice” as a mathematical construct describes the periodicity and symmetry of the array. Bravais lattices have special symmetry properties. Attaching an atom or a group of atoms to each lattice point produces the crystal.

6.2.1 LATTICE

A physical crystal receives its periodic structure from a lattice, which is a mathematical object. One imagines the lattice as an infinite collection of points (in the sense of Euclidean geometry) with a specific arrangement. We now supply two equivalent definitions for the lattice.

Definition 6.1: Given three noncoplanar vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$, we define the lattice to be the collection of points given by

$$\vec{r} = n_1 \vec{a}_1 + n_2 \vec{a}_2 + n_3 \vec{a}_3 \quad (6.5)$$

for all integers n_i (positive and negative). The lattice consists of a set of points $\vec{r}$ —not atoms! The primitive vectors are the shortest vectors $\vec{a}_i$ such that every linear combination in Equation 6.5 produces the lattice. The primitive vectors span the lattice in the sense of Equation 6.5. The “primitive” vectors $\vec{a}_i$ generate (span) the lattice but a given lattice does not uniquely determine the primitive vectors. For example, the vector $\vec{a}_1$ can be replaced by $-\vec{a}_1$ and still generate the same lattice. As an important note, the vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$ in the definition need be neither unit vectors nor orthogonal. For this reason, the primitive vectors should not be called “basis vectors” in order to avoid any possible confusion (although some authors do call them basis vectors).

Definition 6.2: If there exist three noncoplanar vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$ such that the point $\vec{r} = \vec{r}' + n_1 \vec{a}_1 + n_2 \vec{a}_2 + n_3 \vec{a}_3$ is equivalent to the point $\vec{r}'$ for all integers n_1, n_2, n_3 , then the array of points forms a lattice. “Equivalent” means that the arrangement of atoms looks the same from point $\vec{r}$ as it does from point $\vec{r}'$. The points $\vec{r}, \vec{r}'$ do not necessarily coincide with lattice points or with atoms. The two definitions can be easily seen to be equivalent by noting that $\Delta \vec{r} = \vec{r} - \vec{r}' = n_1 \vec{a}_1 + n_2 \vec{a}_2 + n_3 \vec{a}_3$ must be the lattice points themselves.

Example 6.1

The primitive vectors for Figure 6.5 can be written as

$$\vec{a}_1 = 2\hat{x} \quad \text{and} \quad \vec{a}_2 = \hat{x} + \hat{y}$$

where $\hat{x}, \hat{y}$ represent orthogonal unit vectors.

Example 6.2

A “finite” array of points does not form a lattice because you can find integers such that points entirely surround $\vec{r}'$ but not $\vec{r}$. As a result, the “view” from the two points differ and Definition 6.2 cannot be applied.

6.2.2 TRANSLATION OPERATOR

We can translate the lattice by any combination of primitive vectors and still end up with the same lattice. In other words, if the symbol $\vec{R}$ represents a specific vector in the lattice (i.e., there exist

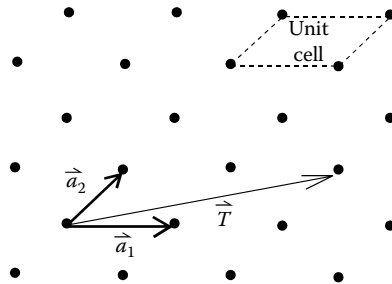


FIGURE 6.5 A 2-D lattice with primitive vectors $\vec{a}_1, \vec{a}_2$. The horizontal points along the x -direction are spaced by two unit vectors and the vertical points along the y -direction are spaced by two unit vectors.

integers n_1, n_2, n_3 such that $\vec{R} = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3$ is a lattice point) then translations by $\vec{R}$ leave the lattice unchanged. Equivalently stated, translations through lattice vectors leave the lattice invariant. We can define one translation operator $\hat{T}_{\vec{R}}$ by

$$\hat{T}_{\vec{R}}\vec{V} = \vec{V} + \vec{R} \quad (6.6a)$$

where $\vec{V}$ represents an arbitrary vector that does not necessarily correspond to a lattice point. The translation operator represents vector addition. Notice that the subscript on the operator $\hat{T}_{\vec{R}}$ gives the vector $\vec{R}$ through which all other vectors must be translated. We will generally use the equivalent ($\vec{R} \rightarrow -\vec{R}$) but more convenient definition for the translation as

$$\hat{T}_{\vec{R}}\vec{V} = \vec{V} - \vec{R} \quad (6.6b)$$

The definition given in Equation 6.6b is a special case of a more general one. The translation operator can be defined for functions. Let $f(\vec{r})$ be a function of the position vector $\vec{r}$. We define the translation operator $\hat{T}_{\vec{R}}$ by

$$\hat{T}_{\vec{R}}f(\vec{r}) = f(\vec{r} - \vec{R}) \quad (6.7)$$

Notice the use of the minus sign to match the convention in Section 3.15. The translation operator assigns a new value to $f(\vec{r})$, namely the value it would have at the position $\vec{r} - \vec{R}$. A moment's thought in connection with Figure 6.6 shows the translation in Equation 6.7 moves the function to the right.

Example 6.3

Consider a one-dimensional (1-D) crystal of atoms with spacing a as shown in Figure 6.7. Let $f(\vec{r})$ be the electrostatic potential as illustrated in the figure. We expect the electrostatic potential to be periodic along the chain of atoms. Let $\vec{R}$ be one of the Bravais lattice vectors given by $\vec{R} = n_1\vec{a}_1$

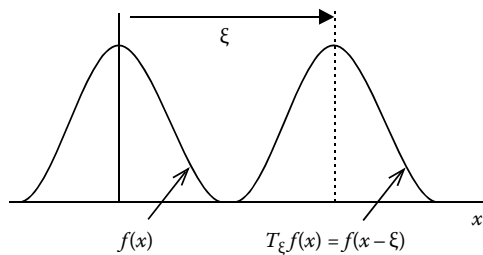


FIGURE 6.6 Translation of function f through X .

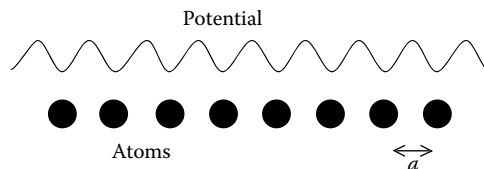


FIGURE 6.7 The electrostatic potential is periodic.

where $\vec{a}_1$ denotes the primitive vector with magnitude a and n_1 is an arbitrary integer. The translation operator $\hat{T}(\vec{R})$ produces

$$\hat{T}_{\vec{R}}f(\vec{r}) = f(\vec{r} - \vec{R})$$

We know that $f(\vec{r} - \vec{R}) = f(\vec{r})$ because the electrostatic potential $f(\vec{r})$ must be periodic in the lattice. So in this case, $\hat{T}_{\vec{R}}f(\vec{r}) = f(\vec{r})$ and the function f must be invariant under translations by a lattice vector $\vec{R}$.

Example 6.4

Show that the definition of the translation operator in Equation 6.7 leads to the definition in Equation 6.6b.

SOLUTION

Let $f(\vec{V}) = \vec{V}$ then $\hat{T}_{\vec{R}}\vec{V} = \hat{T}_{\vec{R}}f(\vec{V}) = f(\vec{V} - \vec{R}) = \vec{V} - \vec{R}$ as required.

6.2.3 ATOMIC BASIS

The crystal consists of an atomic basis (or atomic cluster) attached to the lattice points. The “basis” can be a single atom or a group of atoms attached to each lattice point. Each lattice point receives an identical basis (or cluster). The (infinite) crystal consists of the collection of these regularly arranged clusters.

6.2.4 UNIT CELLS

Unit cells consist of small regions of space that, when duplicated, can be translated to fill the entire volume of the crystal. We briefly consider the primitive unit cell and the conventional unit cell. The primitive unit cell contains exactly a single lattice point and a single cluster. The primitive cell has boundaries made of the primitive vectors, which are the shortest vectors that span the lattice. Therefore, translating the primitive unit cell through every possible integer combination of primitive vectors covers the entire crystal. Figure 6.8 shows two equally valid primitive unit cells. In both cases, the two vectors $\vec{a}_1, \vec{a}_2$ span a region of space that contains only one lattice point or atomic cluster as is obviously true for the primitive cell in the bottom of the figure. The upper unit cell contains exactly one point and one unit cluster since the sides of the parallelepiped cut through the points and clusters in such a way that the sum of the pieces adds up to a single unit. Although

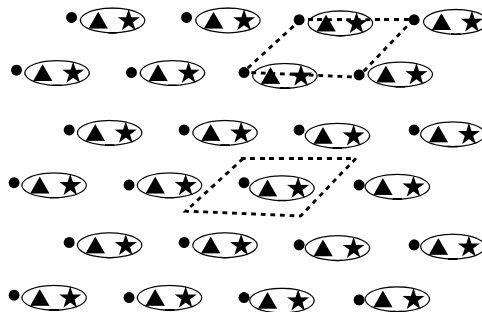


FIGURE 6.8 A crystal is a lattice with an attached atomic basis.

primitive unit cells might appear to be the simplest for calculation purposes, it is sometimes more convenient to work with nonprimitive unit cells.

The conventional unit cell does not necessarily contain exactly one lattice point and one atomic cluster. For calculational convenience, we usually choose orthogonal spanning vectors to define this unit cell. Translating the conventional unit cell by all integer combinations of spanning vectors covers the entire crystal. The next section lists the most typical examples of conventional unit cells.

Example 6.5

Consider the vectors described in the previous Example 6.1

$$\vec{a}_1 = 2\tilde{x} \quad \text{and} \quad \vec{a}_2 = \tilde{x} + \tilde{y}$$

A nonprimitive unit cell can be defined by

$$\vec{a}_1 = 2\tilde{x} \quad \text{and} \quad \vec{b}'_2 = 2\vec{a}_2 - \vec{a}_1 = 2\tilde{y}$$

The spanned volume contains two points.

6.2.5 MILLER INDICES

The points of intersection of a plane with the primitive or nonprimitive spanning vectors can be used to specify a crystal plane. Figure 6.9 shows an example of an infinite plane intersecting the axes defined by three spanning vectors. In this example, the intersection points are $(6, 0, 0)$, $(0, 4, 0)$, and $(0, 0, 4)$.

Miller indices specify a particular plane in the crystal; however, all parallel planes have the same indices. The Miller indices for the plane can be found as follows:

1. Combine the numbers into a single set of parenthesis as $(6, 4, 4)$.
2. Take the reciprocal of these numbers $(\frac{1}{6}, \frac{1}{4}, \frac{1}{4})$.
3. Find three integers having the same ratio $(4, 6, 6)$. This can be accomplished by finding a common denominator and applying it to each number.
4. Convert to the smallest such integers $(2, 3, 3)$.

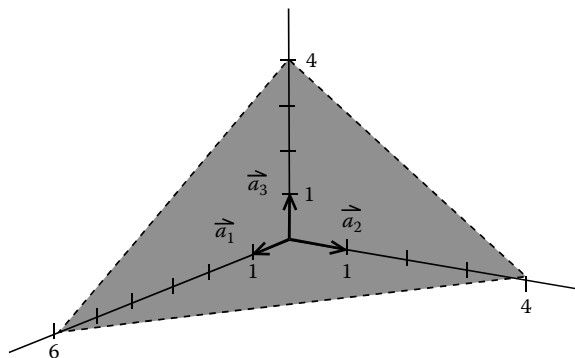


FIGURE 6.9 A crystal plane intersects the three axes.

Two more rules for Miller indices:

1. Intercepts with a negative axis must be indicated with a “bar” over the number.
For example $(\bar{1}, 1, 2)$ indicates planes that intersect the negative $\vec{a}_1$ axis. These indices $(\bar{1}, 1, 2)$ indicate that the plane intersects the axes at $(-2, 0, 0)$, $(0, 2, 0)$, and $(0, 0, 1)$. We can see this since for the single set of coordinates $(-2, 2, 1)$, the reciprocal numbers must be $(-\frac{1}{2}, \frac{1}{2}, \frac{1}{1})$ and so the indices must be $(\bar{1}, 1, 2)$.
2. If a plane is parallel to an axis then the corresponding index must be zero. For example, a crystal plane parallel to the $\vec{a}_2, \vec{a}_3$ axes that passes through the point $+1$ on the $\vec{a}_1$ axis has indices $(1, 0, 0)$. One might reason that the zero occurs since an intercept for a plane parallel to a given axis could be taken as ∞ so that Step 2 above provides $1/\infty = 0$.

The literature often has additional notation in connection with Miller indices.

1. Numbers in braces $\{h, k, l\}$ indicate a set of planes. These planes all have the same Miller indices (h, k, l) and must therefore be parallel to each other.
2. Numbers in brackets $[h, k, l]$ indicate direction. In general, these are not the same numbers as for Miller indices. The direction of $[h, k, l]$ is parallel to $\vec{R} = h\vec{a}_1 + k\vec{a}_2 + l\vec{a}_3$ where $\vec{a}_1, \vec{a}_2, \vec{a}_3$ are the spanning vectors. A direction parallel to $\vec{a}_1$, for example, is $[1, 0, 0]$. For cubic crystals, the direction specifies a vector perpendicular to the corresponding crystal plane.

6.3 SPECIAL UNIT CELLS

A great deal of technology uses semiconductor materials having a diamond-like structure. The crystal has a FCC lattice with a two atom basis. The atomic basis consists of the same atom for silicon while it has two different atoms for GaAs. Subsequent sections will show the corresponding reciprocal lattice forms a body-centered cubic (BCC) lattice. This section covers a number of common lattice types.

6.3.1 BODY-CENTERED CUBIC LATTICE

The body-centered cubic (BCC) cell takes its name from the topology of its conventional unit cell. Conventional unit cells typically have orthogonal spanning vectors and can contain more than a single lattice point unlike the primitive cells. The BCC conventional cell encloses a total of two points (and two clusters when assigned basis atoms) with one of them located at the center of the cube as shown in Figure 6.10. The volume of the BCC conventional cell has twice the volume of the primitive cell for this lattice.

The figure shows three BCC unit cells. The cells have length a along all sides. The conventional spanning vectors can be written as

$$\vec{a}_1^c = a\tilde{x} \quad \vec{a}_2^c = a\tilde{y} \quad \vec{a}_3^c = a\tilde{z} \quad (6.8)$$

where the c superscript indicates “conventional.” The spanning vectors should not be confused with the basis vectors that span the three-dimensional (3-D) vector space or with the primitive vectors. The “primitive” vectors for the BCC lattice can be written in terms of the unit vectors

$$\vec{a}_1 = \frac{a}{2}(\tilde{x} + \tilde{y} - \tilde{z}) \quad \vec{a}_2 = \frac{a}{2}(-\tilde{x} + \tilde{y} + \tilde{z}) \quad \vec{a}_3 = \frac{a}{2}(\tilde{x} - \tilde{y} + \tilde{z}) \quad (6.9)$$

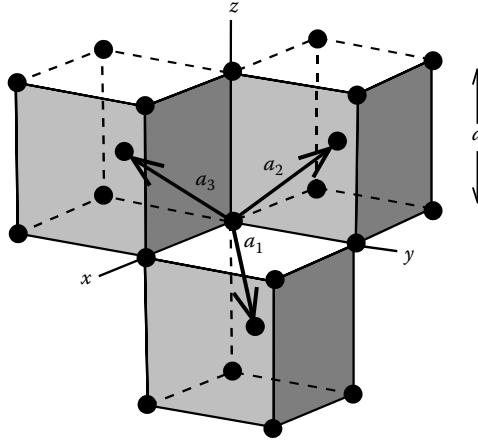


FIGURE 6.10 Three conventional unit cells for the BCC lattice. (From Kittel, C., *Introduction to Solid State Physics*, 5th Edn., John Wiley & Sons, New York, 1976. With permission.)

We can verify that the BCC conventional cell has twice the volume of the primitive cell. The volume enclosed by arbitrary vectors $\vec{a}, \vec{b}, \vec{c}$ has the form

$$V = \vec{a} \cdot (\vec{b} \times \vec{c}) \quad (6.10)$$

We want $V = \vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)$. Calculating $\vec{a}_2 \times \vec{a}_3$

$$\vec{a}_2 \times \vec{a}_3 = \begin{vmatrix} \tilde{x} & \tilde{y} & \tilde{z} \\ -\frac{a}{2} & \frac{a}{2} & \frac{a}{2} \\ \frac{a}{2} & -\frac{a}{2} & \frac{a}{2} \end{vmatrix} = \left(\frac{a}{2}\right)^2 \begin{vmatrix} \tilde{x} & \tilde{y} & \tilde{z} \\ -1 & 1 & 1 \\ 1 & -1 & 1 \end{vmatrix} = \left(\frac{a}{2}\right)^2 [2\tilde{x} - (-2)\tilde{y} + 0\tilde{z}] = \frac{a^2}{2}(\tilde{x} + \tilde{y})$$

The enclosed volume must be

$$V = \vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3) = \frac{a^3}{2}$$

which is half the volume of the conventional cell a^3 .

6.3.2 FACE-CENTERED CUBIC LATTICE

The FCC cell has primitive vectors given by

$$\vec{a}_1 = \frac{a}{2}(\tilde{x} + \tilde{y}) \quad \vec{a}_2 = \frac{a}{2}(\tilde{y} + \tilde{z}) \quad \vec{a}_3 = \frac{a}{2}(\tilde{z} + \tilde{x}) \quad (6.11)$$

An atom or a cluster of atoms occupies the eight corners of the cube and the center of each of the six faces. Keep in mind that the cell contains only 1/8 of each corner point and 1/2 of each face point. Therefore, as indicated as in Figure 6.11, the conventional cell contains exactly four lattice points or four atomic clusters for the crystal.

6.3.3 WIGNER–SEITZ PRIMITIVE CELL

The Wigner–Seitz primitive cell encloses an entire lattice point (or atomic basis) along with all of the volume closest to that point as shown in Figure 6.12. Besides being conceptually convenient, it

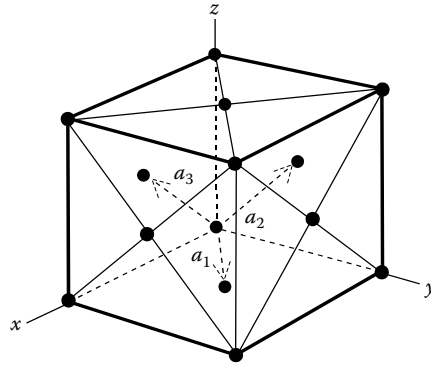


FIGURE 6.11 The FCC lattice.

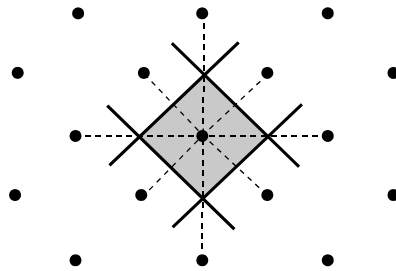


FIGURE 6.12 The Wigner–Seitz cell.

is important for the reciprocal lattice (i.e., the lattice of the Fourier transform variable such as $\vec{k}$). The Wigner–Seitz primitive cell surrounding a given lattice point can be found by

1. Drawing a dotted line from the central point to all other points in the lattice (usually nearest neighbor points as shown in the figure).
2. Draw planes (solid lines) perpendicular to the dotted lines (i.e., planes with a normal that is parallel to the line).
3. Collect all of the space within the interior of the volume formed by the planes. Figure 6.12 shows an example for the two-dimensional (2-D) lattice with the shaded region depicting the Wigner–Seitz cell.

6.3.4 DIAMOND AND ZINC BLENDE LATTICE

The diamond and zinc blende structures have an underlying FCC lattice with a two-atom basis. The diamond structure has all identical atoms (such as carbon). Zinc blende differs from the diamond structure only because the basis contains two different atoms such as gallium and arsenic. It bears repeating that both the diamond and zinc blende structures require a two-atom basis. For diamond, both atoms are identical whereas for zinc blende, the two are different. For clarity, we discuss the zinc blende structure for GaAs.

The left-hand side of Figure 6.13 shows an FCC lattice with atoms in an atomic basis connected by arrows. The structure can be viewed if desired as two FCC lattices with one shifted along the body diagonal by a fraction of the lattice constant of $\sqrt{3}/4$. The basic structure can most easily be seen in the right-hand side of the figure (the distances are distorted for clarity). Technically, GaAs has the zinc blende (or sometimes called the cubic zinc sulfide) structure. The zinc blende lattice is identical to that of the diamond except one of the carbon atoms in the atomic cluster is replaced by

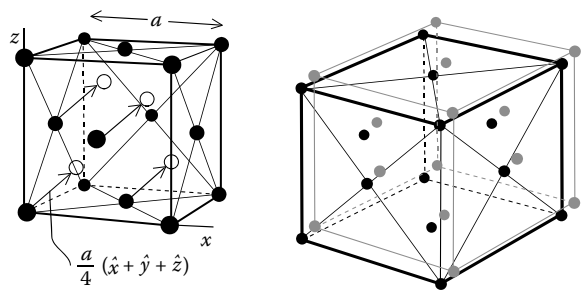


FIGURE 6.13 Two representations of the zinc blende structure.

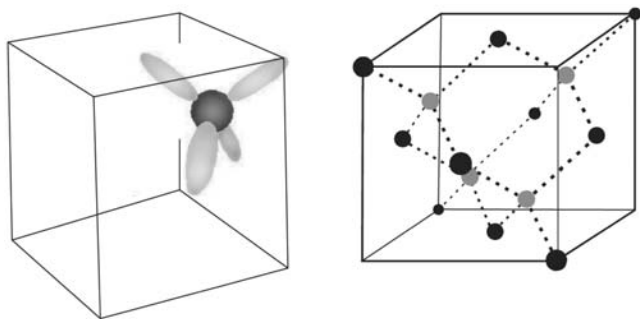


FIGURE 6.14 Left: How the basic set of hybridized orbitals fit into the FCC conventional cell. Right: Slightly rotated view with added atoms.

gallium and the other is replaced with arsenic. For example, the dark atoms in the right-hand side of the figure might be the gallium while the lighter ones might be arsenic. The table below provides examples of the zinc blende structure with the lattice spacing a in angstroms.

Crystal	a	Crystal	a	Crystal	a
AlP	5.45	GaP	5.45	InP	5.87
AlAs	5.62	GaAs	5.65	InAs	6.04
AlSb	6.13	GaSb	6.12	InSb	6.48

6.3.5 TETRAHEDRAL BONDING AND THE DIAMOND STRUCTURE

One can see how the hybrid orbitals discussed in Section 6.1 combine to form the Diamond-like structures. The left-hand side of Figure 6.14 shows how the hybridized orbitals fit into the FCC lattice. Notice how the end of one bond sits at a vertex while the other three produce the faces of the conventional cell. By continuing the construction, one obtains the FCC crystal. The right-hand side of Figure 6.14 shows the construction and how it fits into the FCC conventional unit cell. For some crystals, such as silicon, all of the atoms are the same whereas the figure shows two different types such as for GaAs.

6.4 RECIPROCAL LATTICE

The spatial Fourier expansions of functions having the periodicity of the lattice use special k -vectors. Quantities with the periodicity of the lattice must be made of sines and cosines with wavelengths equal or smaller than the separation between the direct lattice points. The k -vectors corresponding to

these wavelengths reside in the so-called reciprocal lattice and are customarily denoted by G . The reciprocal lattice vectors define the Brillouin zones for phonon and electron band diagrams.

The first topic provides easy-to-use formulas for the reciprocal lattice vectors. The second topic shows that they must be related to the Fourier expansion.

6.4.1 PRIMITIVE RECIPROCAL LATTICE VECTORS

We define the primitive reciprocal lattice vectors by

$$\vec{b}_1 = 2\pi \frac{\vec{a}_2 \times \vec{a}_3}{\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)} \quad \vec{b}_2 = 2\pi \frac{\vec{a}_3 \times \vec{a}_1}{\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)} \quad \vec{b}_3 = 2\pi \frac{\vec{a}_1 \times \vec{a}_2}{\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)} \quad (6.12)$$

where $\vec{a}_i$ denote the primitive vectors for the “direct” lattice. The vectors $\vec{b}_1, \vec{b}_2, \vec{b}_3$ span the reciprocal lattice. Recall from the previous section, the denominator in Equation 6.12, namely $\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)$, gives the volume of the unit cell in the direct lattice. The denominators are numbers (not vectors) that normalize the reciprocal primitive vectors. The set of all vectors $\{\vec{G} = h\vec{b}_1 + k\vec{b}_2 + l\vec{b}_3\}$ for integers h, k, l defines the reciprocal lattice. These reciprocal lattice vectors appear in a Fourier expansion of a function with the periodicity of the lattice. Sometimes people imagine that the reciprocal lattice exists in a separate physical space from the direct lattice. Not true. The cross-product vectors $\vec{a}_i \times \vec{a}_j$ in Equation 6.12 represent a third vector albeit with different purpose. We must deal with distinct lattices of points in a single 3-D space.

The primitive reciprocal lattice vectors must be perpendicular to their corresponding primitive direct lattice vectors. Elementary studies in vector analysis indicate that a cross product $\vec{v} \times \vec{w}$ must always be perpendicular to both of the vectors $\vec{v}, \vec{w}$. Therefore the primitive and reciprocal lattice vectors satisfy the relation

$$\vec{a}_i \cdot \vec{b}_j = 2\pi \delta_{ij} \quad (6.13)$$

where δ_{ij} represents the Kronecker delta function

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \quad (6.14)$$

It is easy to show the following relations between the direct and reciprocal lattices: (1) cubic cells become cubic cells, (2) BCC becomes FCC, and (3) FCC becomes BCC. Of course, reverse transforming the reciprocal lattice again gives the direct lattice.

The next example shows that a cubic direct lattice transforms into a cubic reciprocal lattice. Notice that if the atoms are separated by roughly $3 \text{ \AA}$ (0.3 nm) then the reciprocal vectors have lengths larger than $k \sim 2\pi/\lambda = 2\pi/0.3 \text{ nm} \sim 20/\text{nm}$, which is quite large compared with the typical optical vectors having a length of $2\pi/628 \text{ nm} = 0.01/\text{nm}$ for red light.

Example 6.6

Find the reciprocal lattice for the simple cubic (SC) direct lattice.

SOLUTION

The primitive vectors for the SC lattice can be written as

$$\vec{a}_1 = a\vec{x} \quad \vec{a}_2 = a\vec{y} \quad \vec{a}_3 = a\vec{z}$$

We find that the reciprocal lattice is also SC by calculating the primitive reciprocal lattice vectors

$$\vec{b}_1 = 2\pi \frac{\vec{a}_2 \times \vec{a}_3}{\vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)} = 2\pi \frac{a^2 \tilde{x}}{a^3} = \frac{2\pi}{a} \tilde{x} \quad \vec{b}_2 = \frac{2\pi}{a} \tilde{y} \quad \vec{b}_3 = \frac{2\pi}{a} \tilde{z}$$

6.4.2 DISCUSSION OF RECIPROCAL LATTICE VECTOR IN THE FOURIER SERIES

We now show the importance of the reciprocal lattice for the Fourier series. The previous section defines the reciprocal lattice and demonstrates how the simple cubic (SC) direct lattice produces a SC reciprocal lattice. For simplicity, we work with a 1-D line of atoms with spacing a . For this case, the primitive vector for the direct lattice is

$$\vec{a}_1 = a\tilde{x} \quad (6.15)$$

The lattice points must be given by ma where m denotes an integer and we drop the vector notation for simplicity. For the 1-D case, the operator that translates a function through a lattice vector of ma has the definition

$$\hat{T}_{ma}f(x) = f(x - ma) \quad (6.16)$$

Of particular importance, the result of Example 6.6 shows that the reciprocal lattice must be given by

$$G_n = \frac{2\pi n}{a} \quad (6.17)$$

where n is an integer.

The reciprocal lattice consists the collection of wave vectors for the Fourier series that has the periodicity of the direct lattice (see Figure 6.15). One can see this most simply by working with a 1-D Fourier series as might be appropriate for the potential function shown in Figure 6.15.

The Fourier series representation of the function $f(x)$ can be written as either

$$f(x) = A_0 + \sum_{n=1}^{\infty} [A_n \cos(k_n x) + B_n \sin(k_n x)] \quad (6.18a)$$

or using an alternate basis set as

$$f(x) = \sum_{n=0}^{\pm\infty} c_n e^{ik_n x} \quad (6.18b)$$

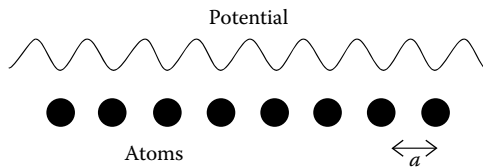


FIGURE 6.15 A function with the periodicity of the lattice.

where the value of k_n must be determined so as to ensure the series has the same periodicity as $f(x)$. The value of k_n can be determined by requiring the series to be invariant under lattice translations ma where a is the length of the lattice vector and m is an integer. However, we must use the smallest lattice translation ($m = 1$) since the function must repeat from one unit cell to the next. A value of $m = 2$ (or larger) would require the function to repeat after 2 (or more) unit cells and the function would not be periodic in the lattice. Therefore, we require $\hat{T}_a f(x) = f(x - a) = f(x)$ which produces the following string of equalities.

$$\sum_n c_n e^{ik_n x} = f(x) = \hat{T}_a f(x) = \sum_n c_n e^{ik_n(x-a)} \quad (6.19)$$

This last equation shows $e^{ika} = 1$ or $k = 2\pi n/a$. Notice that this matches the value produced in Example 6.6. Substituting these values into Equation 6.18 provides the Fourier series

$$f(x) = A_0 + \sum_{n=1}^{\infty} \left[A_n \cos\left(\frac{2n\pi x}{a}\right) + B_n \sin\left(\frac{2n\pi x}{a}\right) \right] \quad (6.20a)$$

or

$$f(x) = \sum_{n=0}^{\pm\infty} c_n e^{ik_n x} \quad \text{with} \quad k_n = \frac{2n\pi}{a} \quad (6.20b)$$

Each integer n provides a value for the wave vector. The reciprocal lattice is defined to be the collection of these wave vectors.

6.4.3 FOURIER SERIES AND GENERAL LATTICE TRANSLATIONS

The reciprocal lattice corresponds to the wave vectors used in the Fourier expansion of functions periodic in the lattice. Such functions must be invariant with respect to displacement through an arbitrary lattice vector $\vec{R} = \sum n_i \vec{a}_i$ where, as before, the $\vec{a}_i$ represent the primitive vectors and the n_i are arbitrary integers. The invariance has the form $\hat{T}_{\vec{R}} f(\vec{r}) = f(\vec{r} - \vec{R}) = f(\vec{r})$ for three spatial dimensions. A general function not periodic in the lattice would use the Fourier transform with the corresponding continuous set of wave vectors. However, our interest at the moment centers on the periodic functions.

Consider the function of the three dimensions $f(\vec{r})$

$$f(\vec{r}) = \sum_{\vec{k}} A_{\vec{k}} e^{i\vec{k} \cdot \vec{r}} \quad (6.21)$$

where $A_{\vec{k}}$ represent the Fourier coefficients for $f(\vec{r})$. The periodicity of $f(\vec{r})$ requires

$$f(\vec{r} - \vec{R}) = f(\vec{r}) = \sum_{\vec{k}} A_{\vec{k}} e^{i\vec{k} \cdot (\vec{r} - \vec{R})} \quad (6.22)$$

to be the same as the function in Equation 6.21 for any arbitrary lattice vector

$$\vec{R} = \sum_i n_i \vec{a}_i \quad (6.23)$$

Because $\vec{R}$ is arbitrary, $A_{\vec{k}}$ cannot be required to satisfy a special relation in order to satisfy Equation 6.22. Instead require

$$e^{i\vec{k} \cdot \vec{R}} = 1 \quad (6.24a)$$

Therefore, only certain $\vec{k}$ are allowed. This last relation can be equivalently written as

$$\vec{k} \cdot \vec{R} = 2\pi N \quad (6.24b)$$

where N represents an integer.

We can first see that the reciprocal lattice vectors satisfy this last relation. Using

$$\vec{R} = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3 \quad \text{and} \quad \vec{G} = m_1\vec{b}_1 + m_2\vec{b}_2 + m_3\vec{b}_3$$

for m_i integer, one finds

$$\vec{G} \cdot \vec{R} = (m_1\vec{b}_1 + m_2\vec{b}_2 + m_3\vec{b}_3) \cdot (n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3)$$

Using the “orthogonality” between primitive vectors $\vec{a}_i \cdot \vec{b}_j = 2\pi\delta_{ij}$, one finds

$$\vec{G} \cdot \vec{R} = 2\pi(n_1m_1 + n_2m_2 + n_3m_3) = 2\pi N$$

where $N = n_1m_1 + n_2m_2 + n_3m_3$ must be an integer.

Now to show that $\vec{k} \cdot \vec{R} = 2\pi N$ (Equation 6.24b) requires $\{\vec{k}\}$ to be the set of reciprocal lattice vectors $\{\vec{k}\} = \{\vec{G}\}$. Let $\vec{k} = \sum_j c_j\vec{b}_j$ where we will show that the c_j must be integers and hence $\{\vec{k}\}$ must be reciprocal lattice vectors. Equations 6.24b and 6.13 produce

$$2N\pi = \vec{k} \cdot \vec{R} = \sum_i 2\pi c_i n_i \quad \text{or equivalently} \quad \sum_i c_i n_i = N \quad (6.25)$$

where N must be an integer (but unspecified). The vector $\vec{R}$ is arbitrary, which means all of the integers n_i are arbitrary, and therefore if c_i is a fraction, then the sum $\sum c_i n_i$ might be an integer for one set of n_i but not for another set of n_i . This observation therefore requires all c_i to be integer.

6.4.4 APPLICATION TO X-RAY DIFFRACTION

The vectors corresponding to the reciprocal lattice are the wave vectors appearing in the Fourier expansion of functions with the periodicity of the lattice. In particular, these functions satisfy the relation $\hat{T}_{\vec{R}} f(\vec{r}) = f(\vec{r} - \vec{R}) = f(\vec{r})$ where $\vec{R} = \sum n_i\vec{a}_i$ represents an arbitrary lattice vector, $\vec{a}_i$ represents the primitive vectors and the n_i are arbitrary integers.

One most common application in solid-state books concerns x-ray and electron diffraction from crystals. Consider x-rays for example, since both applications develop in similar fashion. Assume a wave incident on a scattering center has the form (Figure 6.16)

$$W_{\text{in}} = A e^{i\vec{k}_0 \cdot \vec{r}} \quad (6.26)$$

where $\vec{k}_0$ represents the wave vector of the incoming monochromatic wave. The diffracted wave has the form

$$W_{\text{out}} = e^{i\vec{k} \cdot (\vec{r} - \vec{\xi})} f(\vec{\xi}) W_{\text{in}}(\vec{\xi}) \quad (6.27)$$

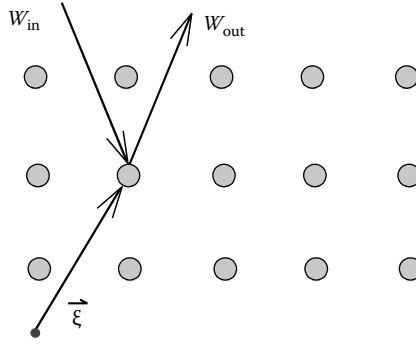


FIGURE 6.16 An example periodic structure with input and diffracted waves W_{in} and W_{out} , respectively.

This last equation has the interpretation that the “strength” of the scattering center represented by $f(\vec{\xi})$ changes the direction of the incident wave and is proportional to the magnitude of the incident wave at the position of the scattering center $\vec{\xi}$, and reradiates the wave in the direction of $\vec{k}$ as if the wave originated at $\vec{\xi}$ through the argument $\vec{r} - \vec{\xi}$. The total diffracted wave can be found by integrating over all the scattering centers as follows

$$W_{Total} = \int d^3\xi e^{i\vec{k} \cdot (\vec{r} - \vec{\xi})} f(\vec{\xi}) W_{in}(\vec{\xi}) \quad (6.28)$$

For x-rays, the function f can be interpreted as the electron density which has the periodicity of the crystal. Regardless of the origin of the scattering, assume f has the periodicity of the crystal so that it can be expanded in a Fourier series with the reciprocal lattice vectors as wave vectors.

$$f(\vec{\xi}) = \sum_{\vec{G}} f_G e^{i\vec{G} \cdot \vec{\xi}} \quad (6.29)$$

All of the possible reciprocal lattice vectors for this situation appear in the summation. Substituting Equation 6.29 into Equation 6.28 produces

$$W_{Total} = \sum_{\vec{G}} f_G e^{i\vec{k} \cdot \vec{r}} \int d^3\xi e^{i(\vec{G} - \Delta\vec{k}) \cdot \vec{\xi}} \quad (6.30)$$

where $\Delta\vec{k} = \vec{k} - \vec{k}_0$. The integral produces zero unless $\Delta\vec{k} = \vec{G}$ where $\vec{G}$ must be one of the reciprocal lattice vectors. This can easily be seen by either considering the integral to be the inner product of two Fourier series basis vectors or by Figure 6.17. For the figure, when $\Delta\vec{k} \neq \vec{G}$, the factor $e^{i(\vec{G} - \Delta\vec{k}) \cdot \vec{\xi}}$ has unit length in the complex plane and arbitrary angles depending on $\vec{\xi}$. Note that the angle between the vector and the real axis (the horizontal axis) is the exponent (without the i) of the exponential function. The figure shows only eight of the possible factors but the integral will reference an infinite number of them. However when $\Delta\vec{k} = \vec{G}$, the exponential will have the value of one and the integral will give the volume of the crystal.

The condition that a reciprocal lattice vector must be equal to the difference between the wave vectors for the diffracted and incident wave vectors $\Delta\vec{k} = \vec{G}$ has applications to material studies including the “powder method” and the “Laue” method for diffraction. Similar consideration applies to photonic band-gap materials.

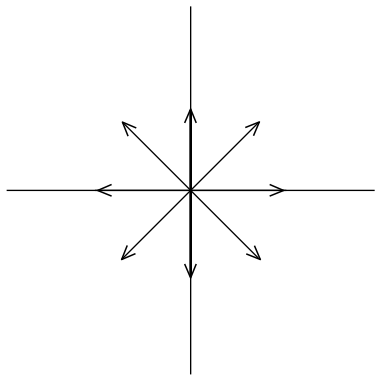


FIGURE 6.17 Plot of an exponential function with arbitrary phase in the complex plane.

6.4.5 COMMENT ON BAND DIAGRAMS AND DISPERSION CURVES

We will have primary interest in the application of the reciprocal lattice vectors to electron and phonon bands. Consider the electronic case for a simple cubic (SC) lattice with interatomic spacing of a . The electron moving in the crystal rarely has a wave vector $\vec{k}$ equal to a reciprocal lattice vector. The electron can have wavelengths as long as the crystal L (i.e., assuming periodic boundary conditions-more on this later) so that (Figure 6.18)

$$\lambda_e = L/m \quad \text{and} \quad k_m = 2\pi/\lambda_e = 2\pi m/L \tag{6.31}$$

where the first term refers to the wavelength of the electron. The wavelength obtains by assuming a multiple number of electron wavelengths must fit within the length L of the crystal.

The wave vectors define the allowed states for the electrons in the semiconductor bands as will be discussed in Chapter 7. The atomic spacing a leads to the reciprocal lattice vectors with magnitude given by

$$G_n = 2\pi n/a \tag{6.32}$$

An estimate of the magnitude can be calculated assuming an atomic spacing of $5 \text{ \AA}$ and a crystal having size of $5000 \text{ \AA}$ (a very-small crystal). For $n = m = 1$, we find $k_1 = G_1/5000$ where is very small and definitely not equal to even the smallest reciprocal lattice vector. One should note that the reciprocal lattice vectors lead to spatial wavelengths smaller than the atomic spacing (in order that the associated functions be periodic on the lattice) where as the electron wavelengths are most often much longer than the atomic spacing (but with wavelengths a sub-multiple of the crystal length).

Figure 6.19 shows a typical band diagram (direct band gap) with the allowed states k_m represented by circles. Each band has states. The first Brillouin zone (FBZ) corresponds to the

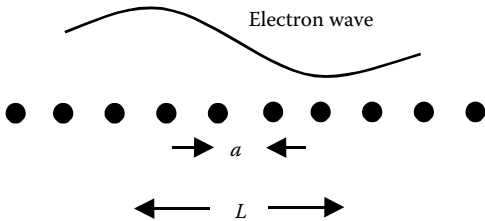


FIGURE 6.18 The crystal has length L while the atoms have spacing a .

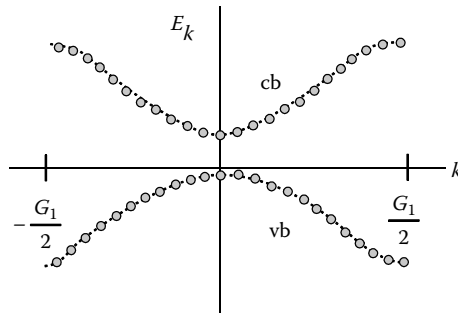


FIGURE 6.19 Electron band diagram showing allowed k and the FBZ.

Wigner–Seitz cell discussed in Section 6.3 but in the reciprocal lattice rather than in the direct lattice. The bands repeat from one zone to the next and therefore contain redundant information. Figure 6.19 represents the reduced zone scheme for representing semiconductor bands. As will be seen in the next chapter, high-energy electrons with $k = \pm G_1/2$ undergo very strong reflections from the crystal atoms and thereby form standing waves which consists of high-speed waves moving in either direction.

6.5 COMMENTS ON CRYSTAL SYMMETRIES

Crystal symmetries play an important role in determining the properties of the solid. We already know one type of symmetry operation consisting of translation through a lattice vector. Latter chapters show this symmetry operation produces the Bloch wave function and the band structure. The reader is encouraged to refer to the books by Yu or Yariv or Ascroft and Mermin for more information on crystal symmetry.

6.5.1 SPACE AND POINT GROUPS

Symmetry operations transform a lattice or crystal into itself. A symmetry of the crystal must take into account the symmetry of the atomic basis. In this case, the basis includes the collection of atoms along with their bonds. Sometimes, we assume that the basis (i.e., cluster) has the same symmetry as the lattice for convenience. Let us consider the symmetry operations on the Bravais lattice.

If the operator $\hat{O}$ transforms the lattice vectors $\{\vec{R}\}$ into the set $\{\vec{R}'\}$ then the operator $\hat{O}$ represents a symmetry operation for the lattice when $\{\vec{R}\} = \{\vec{R}'\}$. Equality between sets just requires that both sets have exactly the same points. The equality between sets can also be stated as the set $\{\vec{R}'\}$ must be contained in the set $\{\vec{R}\}$, written as $\{\vec{R}'\} \subseteq \{\vec{R}\}$, and also vice versa $\{\vec{R}\} \subseteq \{\vec{R}'\}$. As far as concerns the crystal, these operations must be equivalent to the identity operator.

We can list some typical operations. These operations can be used to generate lattices.

1. Translations:

A translation through the lattice vector $\vec{R}$ can be written as $\hat{T}_{\vec{R}}f(\vec{r}) = f(\vec{r} - \vec{R})$.

2. Reflection through a plane:

Figure 6.20 shows the reflection through a plane where each solid arrow produces the reflected image represented by the dotted arrow.

3. Rotation about an axis with angle $360/n$:

$$\hat{R}_nf(\theta) = f\left(\theta + \frac{2\pi}{n}\right)$$

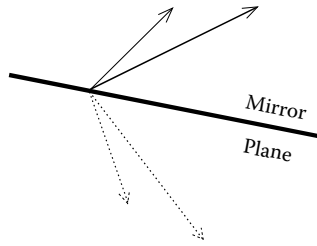


FIGURE 6.20 Solid vectors representing lattice points are reflected in the mirror plane to produce the dotted lines.

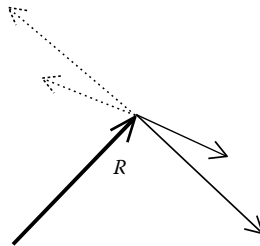


FIGURE 6.21 Inversion through a point.

4. Inversion through a point $\vec{R}$:

$$\hat{I} f(\vec{R} + \vec{r}) = f(\vec{R} - \vec{r})$$

which can be written for the origin as $\hat{I} f(\vec{r}) = f(-\vec{r})$. See Figure 6.21.

5. Glide = reflection + translation (translation through $\frac{1}{2}$ primitive lattice vectors).

6. Screw = rotation + translation.

7. Compound operations consist of two of those listed in 1–4 above.

The following definitions are important for the study of symmetry since symmetry can be applied to either the lattice or to the crystal.

1. The space group consists of the collection of all symmetry operations including translations in 3-D space.
2. The point group consists of the collection of all symmetry operations except translations. These operations leave at least one point fixed in space. The point group sometime refers to the lattice and sometimes to the atomic basis. If applied to the crystal, both the lattice and the basis must be invariant.
3. The plane group consists of all symmetry operations for a 2-D crystal (i.e., all atoms in a plane).

Example 6.7

All Bravais lattices have inversion symmetry (Figure 6.21). The lattice vector has inversion symmetry when $\hat{I} F(\vec{r}) = F(-\vec{r})$ and $F(-\vec{r}) = F(\vec{r})$. We can see that the lattice has inversion symmetry as follows. Let $\vec{v} = m\vec{a} + n\vec{b} + p\vec{c}$ be a lattice vector where m , n , and p are integers (positive, negative, and zero) so that $\vec{v}$ must be in the lattice defined by $\vec{v} \in \{\vec{R} = m\vec{a} + n\vec{b} + p\vec{c}; m, n, \text{ and } p \text{ are integers}\}$. The inversion operator provides

$$\hat{I} \vec{v} = -\vec{v} = -m\vec{a} - n\vec{b} - p\vec{c}$$

Defining new integers m', n', p' produces

$$\hat{I} \vec{v} = m'\vec{a} + n'\vec{b} + p'\vec{c} \in \{\vec{R}\}$$

Therefore we find $\{\vec{R}\} \subseteq \{\vec{R}'\}$. The case for $\{\vec{R}'\} \subseteq \{\vec{R}\}$ can be similarly demonstrated. Therefore we conclude that inversion must be a symmetry operation for all Bravais lattices.

Example 6.8

Show reflections of a square 2-D lattice through the 45° mirror plane produces the same lattice (see Figure 6.22). Note that we use the notation $|1\rangle$, $|2\rangle$, and $|3\rangle$ to represent the unit vectors along the x -, y -, and z -axis, respectively.

SOLUTION

The mirror operator in this case has the effect

$$\hat{M}|1\rangle = -|2\rangle \quad \hat{M}|2\rangle = -|1\rangle$$

which by Chapter 3, completely defines the transformation.

Let $\vec{v}$ be in the square lattice so that

$$\vec{v} = m|1\rangle + n|2\rangle$$

where m and n must be integers. We find

$$\vec{v}' = \hat{M}\vec{v} = m\hat{M}|1\rangle + n\hat{M}|2\rangle = -m|2\rangle - n|1\rangle$$

We can define new integers $m' = -n$, $n' = -m$ so that

$$\vec{v}' = m'|1\rangle + n'|2\rangle$$

which must also be a vector in the original lattice.

6.5.2 ROTATIONS

The operations listed above can be used to generate lattices. We have already discussed how the translation vectors generate a lattice. There exist only 5 types of 2-D lattices and 14 types of 3-D lattices. The operations must be consistent with the translation symmetry. This consistency requires rotations to have angles of $360/n$ where n takes on only the values $n = 1, 2, 3, 4, 6$.

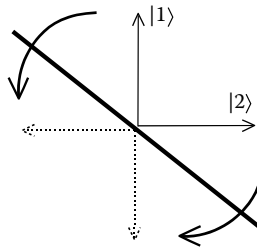


FIGURE 6.22 Reflection of unit vectors in the mirror plane.

One can easily show this last assertion that rotations describing Bravais lattices can have no other angles than $360/n$ where $n = 1, 2, 3, 4, 6$. The proof proceeds by defining a rotation between primitive vectors and then comparing with the traditional rotation for an orthonormal basis set. One will find that taking the trace of the two types of rotation matrices produces the desired relation of $360/n$. To start, one must first define the vectors and operations.

Let $|v\rangle$ be a lattice vector described in terms of the primitive vectors as

$$|v\rangle = \sum_j v_j^{(a)} |a_j\rangle \quad (6.33)$$

(here, $|v\rangle$ is used rather than $\vec{r}$ as in Section 6.2 in order to prevent confusion with the coordinate ket) where each n_i represents an integer and $|a_j\rangle$ represents a primitive vector. The (a) superscript indicates the coefficients refer to the primitive vectors. For $|v\rangle$ to be a lattice vector, the coefficients $v_j^{(a)}$ must be integers $v_j^{(a)} = n_j$. The same lattice vector $|v\rangle$ can be written in terms of the orthonormal basis vectors $|i\rangle$ (representing $\hat{x}, \hat{y}, \hat{z}$) by specifying $|a_j\rangle$ in terms of $|i\rangle$.

$$|a_j\rangle = \sum_i S_{i,j} |i\rangle \quad (6.34)$$

Combining Equations 6.33 and 6.34 produces

$$|v\rangle = \sum_{i,j} S_{i,j} v_j^{(a)} |i\rangle \quad (6.35)$$

Comparing this last result with the usual expression for the vector components in an orthonormal set $|v\rangle = \sum_i v_i |i\rangle$ shows the components must be given by

$$v_i = \sum_j S_{i,j} v_j^{(a)} \quad (6.36a)$$

In matrix notation, we then have

$$\underline{v} = \underline{S} \underline{v}^{(a)} \quad (6.36b)$$

One can show that $\underline{S}^{-1}$ must exist.

Next consider a rotation $\hat{R}$ that maps the lattice into itself (a symmetry operation). For the rotation to be consistent with the translational symmetry, one requires $|v'\rangle = \hat{R}|v\rangle$ also be a lattice vector with an expansion

$$|v'\rangle = \sum_j v_j'^{(a)} |a_j\rangle \quad (6.37)$$

where again, the coefficients $v_j'^{(a)}$ must be integers m_j . Use the notation that $\underline{R}^{(a)}$ refers to an array that operates on the column vectors formed from $v_j^{(a)} = n_j$

$$\begin{pmatrix} v_1^{(a)} \\ v_2^{(a)} \\ v_3^{(a)} \end{pmatrix} = \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} \quad (6.38)$$

One can see the matrix $\underline{R}^{(a)}$ can only have integer elements $R_{ij}^{(a)}$ by considering its effect on each unit column vector

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (6.39)$$

which are formed by the coefficients of the primitive vector expansion. For example, consider the rotation of the first column vector, then

$$\begin{pmatrix} R_{11}^{(a)} & R_{12}^{(a)} & R_{13}^{(a)} \\ R_{21}^{(a)} & R_{22}^{(a)} & R_{23}^{(a)} \\ R_{31}^{(a)} & R_{32}^{(a)} & R_{33}^{(a)} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} R_{11}^{(a)} \\ R_{21}^{(a)} \\ R_{31}^{(a)} \end{pmatrix} \quad (6.40)$$

Therefore, because the components of the resultant vector must be the integers m_i , it follows that $R_{ij}^{(a)}$ must be these integers. The argument can be repeated for the other unit column vectors. Given that the matrix consists of integers then requires the trace of the matrix to be an integer N .

$$\text{Trace}(\underline{R}^{(a)}) = N \quad (6.41)$$

Now, one needs to relate the rotation to an angle by using the orthonormal basis $|i\rangle$. In such a case, the rotation can be expressed as

$$\underline{R} = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (6.42)$$

where the third axis is the axis of rotation.

It remains to relate the two rotation matrices. Consider the following sequence using Equation 6.36b, namely $\underline{v} = \underline{S}\underline{v}^{(a)}$

$$\underline{v}^{(a)} = \underline{R}^{(a)}\underline{v}^{(a)} \rightarrow \underline{S}\underline{v}^{(a)} = \underline{S}\underline{R}^{(a)}\underline{S}^{-1}\underline{S}\underline{v}^{(a)} \rightarrow \underline{v}' = \underline{S}\underline{R}^{(a)}\underline{S}^{-1}\underline{v} \quad (6.43a)$$

Comparing this last result with the matrix equation $\underline{v}' = \underline{R}\underline{v}$ shows the rotation matrices must be related by a similarity transformation.

$$\underline{R} = \underline{S}\underline{R}^{(a)}\underline{S}^{-1} \quad (6.43b)$$

Finally now, the Trace properties provide the relation

$$\text{Trace}(\underline{R}) = \text{Trace}(\underline{R}^{(a)}) = N \quad (6.43c)$$

where use has been made of Equation 6.41 and N must be an integer. The trace of the $\underline{R}$ matrix becomes

$$1 + 2 \cos \theta = N \quad (6.44)$$

Given that $-1 \leq \cos \theta \leq 1$, one can only have that $N = -1, 0, 1, 2, 3$. Solving for the angle provides the values

$$\theta = 360/n \quad n = 1, 2, 3, 4, 6 \quad (6.45)$$

as required.

6.5.3 DEFECTS

A crystal defect occurs by altering a perfect crystal in such a way that the original crystal is not reproduced. Generally, defects can be classified as point defects, line and surface defects, and dislocations. As mentioned in the introductory chapter, defects can produce band gap states. Some defects have beneficial effects such as for doping.

Defects that produce states near the middle of the band gap tend to function as recombination centers. If conduction carriers are lost to these recombination centers then the carrier population must decrease and the conductivity must be lower than without the defects. Perhaps paradoxically, these defects decrease the response time to sudden changes in the carrier population as would be required for high-speed modulation for example. The tradeoff between the conductivity and modulation rate leads to the concept of the gain-bandwidth product. For example, if light momentarily shines on a semiconductor, without recombination centers, it might take the electrons and holes a long time to recombine. In this case, a current will flow in response to an applied voltage for a long time after extinguishing the light. However, with recombination centers, the excess carriers will be rapidly removed from their respective conduction bands after extinguishing the light, and the current will rapidly drop even though a voltage might still be applied. Therefore defects can “speedup” the carrier response time. The same recombination centers also reduce the number of carriers during the time that the beam illuminates the semiconductor. In effect, it also reduces the gain as required by the gain-bandwidth product.

Point defects can be subdivided into impurity and native point defects. Impurities refer to the random placement of foreign atoms into the crystal. Some impurities, such as dopants, have useful effects. However, other impurities and native point defects tend to reduce conduction and emission efficiency. Point defects can extend over several atomic lattice sites. For example, a missing atom causes nearby atoms to relax. Vacancies refer to an atom missing from a periodic array. Sometimes, we require the missing atom to appear elsewhere within the crystal in order to maintain a constant number of atoms in the sample. Some authors require the atom to appear on the surface; however, a surface is a huge lattice defect.

Line and surface defects can extend across millions of atoms. Cleaving a solid in two parts necessarily produces surface defects since the surface interrupts the periodicity of the crystal. Generally this type of defect produces many dangling bonds (surface states) that appear as states within the band gap.

An example of the edge and screw dislocations appear in Figure 6.23 (taken from Blakemore Figure 1-55).

6.5.4 INTRODUCTION TO SYMMETRIES IN QUANTUM MECHANICS

Our studies of quantum mechanics shows that electron wave functions satisfy the Schrödinger wave equation (SWE) having the form

$$\hat{H}\Psi = i\hbar \frac{\partial \Psi}{\partial t} - \frac{\hbar^2}{2m} \nabla^2 \Psi + V\Psi = i\hbar \frac{\partial \Psi}{\partial t} \quad (6.46)$$

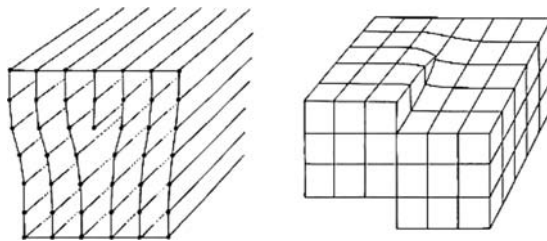


FIGURE 6.23 The edge (left) and screw (right) dislocations after Blakemore. (From Blakemore, J.S., *Solid State Physics*, 2nd Edn., W.B. Saunders Company, Philadelphia, PA, 1974. With permission.)

where V denotes the potential energy. The symbol $\hat{H}$ refers to the Hamiltonian that represents the kinetic and potential energy of the system. The Hamiltonian $\hat{H}$ must be an operator according to

$$\hat{H} = -\frac{\hbar^2}{2m} \nabla^2 + V \quad (6.47)$$

The solution to Schrödinger's equation provides the wave function Ψ ; these wave functions can have the form of a traveling plane wave. As discussed in Chapter 5 on quantum mechanics, the solution to Equation 6.46 can be written as a product of two terms (using separation of variables)

$$\Psi(\vec{r}, t) \sim \psi(\vec{r})e^{iEt/\hbar} \quad (6.48)$$

where the time-independent wave function satisfies the time-independent Schrödinger's equation

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V\psi = E\psi \quad \text{or} \quad \hat{H}\psi = E\psi \quad (6.49)$$

The second of Equations 6.49 has the form of an eigenfunction equation. The parameter E gives the energy of the energy level (or orbital). Recall that the wave function ψ leads to a probability. In many cases, we only need to consider one wave function ψ for each energy E . However, Equation 6.49 sometimes has many solutions ψ for each energy E . In this case, many different electron configurations have the same energy. For example, two traps in a material might hold electrons with identical energy but obviously different wave functions (since the traps must be spatially separated—just think of the wave function as an ordinary function). Another example would be the p orbitals in a silicon atom, for example, without regard to spin. In this case, the p_x , p_y , and p_z states have the same energy.

The symmetries of the Hamiltonian lead to multiple distinct eigenfunctions for the same energy eigenvalue (degenerate eigenvalues). Consider an operator in the group G of symmetries of the Hamiltonian $\hat{O} \in G$. Then starting with

$$\hat{H}\psi = E\psi \quad (6.50a)$$

and applying the operator $\hat{O}$ to both sides, we find

$$\hat{O}\hat{H}\psi = \hat{O}E\psi \quad (6.50b)$$

Using the fact that every group element has an inverse $\hat{O}^{-1}\hat{O} = 1$, we can write

$$\hat{O}\hat{H}\hat{O}^{-1} \hat{O}\psi = E \hat{O}\psi \quad (6.51)$$

The operator $\hat{O}$ represents a symmetry of the Hamiltonian when $\hat{H} = \hat{H}' = \hat{O}\hat{H}\hat{O}^{-1}$. Therefore, the operator $\hat{O}$ leaves the Hamiltonian invariant when $\hat{H}\hat{O} - \hat{O}\hat{H} = 0$. The expression $\hat{H}\hat{O} - \hat{O}\hat{H}$ refers to the commutator $[\hat{H}, \hat{O}] = \hat{H}\hat{O} - \hat{O}\hat{H}$. Therefore $\hat{O}$ represents a symmetry of the Hamiltonian when it commutes with the Hamiltonian $[\hat{H}, \hat{O}] = \hat{H}\hat{O} - \hat{O}\hat{H} = 0$.

Continuing with Equation 6.46,

$$\hat{O}\hat{H}\hat{O}^{-1} \hat{O}\psi = E \hat{O}\psi \quad \rightarrow \quad \hat{H}'\hat{O}\psi = E \hat{O}\psi \quad (6.52)$$

Taking the operation represented by $\hat{O}$ to be a symmetry of the Hamiltonian $\hat{H} = \hat{H}' = \hat{O}\hat{H}\hat{O}^{-1}$, we now find the two results

$$\hat{H}\psi = E\psi \quad \hat{H}\hat{O}\psi = E\hat{O}\psi \quad (6.53)$$

Therefore, both ψ and $\hat{O}\psi$ must be eigenfunctions of the Hamiltonian corresponding to the single eigenvalue E . The eigenfunction ψ and $\hat{O}\psi$ might be distinct from one another in which case applications of the operator determines a vector space of eigenfunctions corresponding to the single eigenvalue E (degenerate eigenvalues). For example, suppose $\hat{O}$ represents the rotation operators $e = R(0)$, $f = R(120)$, $g = R(240)$. In this case we might expect to find three solutions of the form

$$\psi, \hat{f}\psi, \hat{g}\psi$$

all giving the same energy. In such a case, linear combinations produce suitable wave functions.

Another possibility takes ψ and $\hat{O}\psi$ as being essentially the same. That is

$$\hat{O}\psi = C\psi \quad (6.54)$$

where C is a constant. Therefore, we find that eigenvectors of the Hamiltonian $\hat{H}$ must also be eigenvectors of the operator $\hat{O}$. As a matter of fact, as discussed in previous chapters, we can always find simultaneous eigenvectors of commuting Hermitian operators. If $\hat{O}$ represents translations through lattice vectors, then translating the system through a lattice vector leaves the total energy $\hat{H}$ invariant. Equivalently the translated function $\hat{O}\psi$, must be the same as the original function $\hat{O}\psi = C\psi$. We have seen similar statements for the translation of function through a lattice vector $\hat{T}_{\vec{R}}f(\vec{r}) = f(\vec{r} + \vec{R}) = f(\vec{r})$. We will see later how translational symmetry gives rise to the Bloch wave function that describes the motion of electrons and holes in the conduction and valence bands.

6.6 PHONON DISPERSION CURVES FOR MONATOMIC CRYSTAL

The phonon is a particle of energy for the movement of atoms making up a material. The properties of the phonon must be intimately related to the physical structure of the crystal. Any disturbance causing the atoms to move produce phonons. Sound consists of the motion of phonons through the material although it is more common to discuss the wave nature of sound. Heat produces phonons quite naturally since the thermal energy can be stored as the movement of atoms in the material.

The phonons have importance for conduction and optical processes. Collisions with phonons limit the mobility of electrons and holes. Higher temperatures imply larger numbers of phonons and therefore lower electron and hole mobility. Optical processes are also sensitive to the density of phonons; often the phonons reduce the efficiency for the production of photons. For example, phonons enable Auger recombination as well as the transitions for indirect band gaps.

This section presents the equations for the atomic motion starting with Newton's laws. We deduce the band structure (i.e., dispersion curves) for the transverse and longitudinal modes for the acoustic branches. The reciprocal lattice vector has an important role in defining the domain of the bands in k -space. Afterward, we discuss the 3-D crystal and Young's modulus. The results can be compared with the Lagrangian approach in Section 4.6 which treats the material as a continuous medium. Subsequent sections discuss the group velocity, density of states, the probability distribution for phonons, and the case for diatomic crystals. The reader might want to start with Section 6.6.2 and then return to Section 6.6.1 for more information on the normal modes.

6.6.1 INTRODUCTION TO NORMAL MODES FOR MONATOMIC LINEAR CRYSTAL

Energy propagates in crystals through the “wave motion” of the atoms. The phonon is the smallest quantum of energy for the wave motion and pertains to the amplitude of the wave. However, consider two types of oscillatory behavior describe the atomic motion. One type of motion applies to the individual atom oscillating about its equilibrium point and consists of the Fourier sum of multiple frequency components. On the other hand, the normal coordinates describe a collective motion with a single frequency. The focus shifts from a single atom to a spatially extended sinusoidal wave on the crystal. Each atom participating in the oscillation has the same oscillation frequency as every other. The normal modes can be Fourier summed to provide the general wave in the crystal. The phonon normally refers to the smallest quantum of energy for the amplitude of the normal mode. In this sense, the phonon energy must be distributed across all of the atoms participating in the collective motion to form the normal modes; that is, the phonon is not associated with just one atom. The present section illustrates the difference between the motion of single atoms and those participating in the collective motion for the normal modes.

The main issue consists of finding the appropriate equations of motion for the atoms in a simple crystal (1-D). Starting with the potential functions, one can find the forces and thereby deduce the oscillation frequencies and the normal modes (for more information on the normal modes, see Section 4.5). In general, for a linear monatomic crystal with many atoms in the 1-D array, atom $\#n$ exists in an electrostatic “potential well” V created by its immediate neighbors. One often assumes only nearest neighbor atoms, namely $\#(n-1)$ and $\#(n+1)$, directly exert forces on atom $\#n$ through the electrostatic potential. The *displacement* of atom $\#n$ from equilibrium is represented by u_n as shown in Figure 6.24. We denote the *equilibrium position* of atom $\#n$ by x_n . The coordinates x_n serve as indices rather than functions of time. The function $u_n = u(x_n)$ describes the displacement of atom $\#n$ from equilibrium. Notice how a displacement is associated with a particular x -coordinate; in general, different atoms n will be moved different amounts u_n from their equilibrium points x_n . Further, the displacement from equilibrium of atom n (as with any of the atoms) must vary with time and can therefore be represented as $u_n = u(x_n, t)$.

The forces arise from potential energy V . Assuming x_n represents equilibrium and considered to be an index rather than a variable, the potential energy for atom $\#n$ at its location $x_n + u_n$ has the Taylor expansion

$$V(u_n + x_n) \cong V(x_n) + \left. \frac{dV}{du_n} \right|_{x_n} u_n + \frac{1}{2} \left. \frac{d^2V}{du_n^2} \right|_{x_n} u_n^2 + \cdots$$

The equilibrium point x_n corresponds to zero slope and therefore the term with the first derivative in the Taylor expansion must be zero. The quadratic term has the form $\beta x^2/2$ which arises from the linear force of the form “ $F = -\beta x$ ” similar to Hook’s law for springs but with x replaced by u_n and the parameter β as the spring constant. Therefore we identify the spring constant as $\beta \sim \left. \frac{d^2V}{du_n^2} \right|_{x_n}$ which arises from the quadratic approximation for the electrostatic potential. The use of springs simplifies the diagrams (and the math).

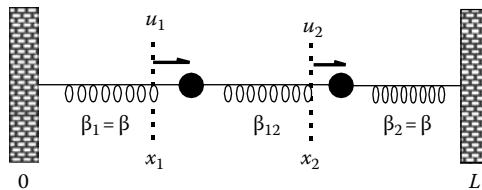


FIGURE 6.24 Longitudinal vibration of masses m coupled by springs.

The simplest demonstration of normal modes uses two atoms as shown in Figure 6.24. The equations of motion can be found from Newton's second law by determining the forces exerted on each mass m by the springs when the masses move from equilibrium. Figure 6.24 shows the "amount of stretch from equilibrium" for the *spring* between atoms #1 and #2 must be given by $u_2 - u_1$. Therefore, the forces exerted on masses #1 and #2, respectively must be

$$\begin{aligned} F_1 &= +\beta_{12}[u_2 - u_1] - \beta u_1 \\ F_2 &= -\beta_{12}[u_2 - u_1] + \beta u_1 \end{aligned} \quad (6.55)$$

The acceleration of each mass has the form $\ddot{u}_n$ which then provides the equations of motion from Equation 6.55.

$$\begin{aligned} m\ddot{u}_1 + (\beta + \beta_{12})u_1 - \beta_{12}u_2 &= 0 \\ m\ddot{u}_2 + (\beta + \beta_{12})u_2 - \beta_{12}u_1 &= 0 \end{aligned} \quad (6.56)$$

We already know the masses will execute harmonic motion and so assume solutions of the form

$$\begin{aligned} u_1(t) &= B_1 e^{i\omega t} \\ u_2(t) &= B_2 e^{i\omega t} \end{aligned} \quad (6.57)$$

Substitute and collect terms to write the matrix equation

$$\begin{bmatrix} \beta + \beta_{12} - m\omega^2 & -\beta_{12} \\ -\beta_{12} & \beta + \beta_{12} - m\omega^2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \underline{0} \quad (6.58)$$

If the matrix has an inverse then we would find that $B_1 = 0 = B_2$ and the atoms would not move from equilibrium. Such a solution does not describe wave motion. Therefore, we must require the matrix to be noninvertible by requiring its determinant to be zero. If the matrix on the left side M could be inverted then we would find $\underline{B} = \underline{M}^{-1}\underline{0} = 0$ where $\underline{B}$ represents the column vector with entries B_1 and B_2 . Such a trivial solution requires the motions to have zero amplitude and therefore the wave does not exist! Instead, one must assume the inverse of the matrix does not exist.

As a point of interest, we will find the frequencies and not the amplitudes B_1 and B_2 . In order to find the amplitudes, one must have further information on the driving force behind the motion. For example, if someone taps the crystal, then the displacement will be related to the energy transferred and we have not specified this amount. If the motion is due to thermal energy, we likewise would need to specify the temperature. So the best we can do without further information consists of finding frequencies and we can normalize the amplitudes. In such a case, once further information is available, the amplitudes might then be specified.

Taking the determinant of the 2×2 matrix equal to zero so as to assure the matrix has not any inverse, provides an equation with ω^4 . Solving for the frequency provides four roots. Define the positive angular frequencies

$$\omega_1 = \sqrt{\frac{\beta + 2\beta_{12}}{m}} \quad \text{and} \quad \omega_2 = \sqrt{\frac{\beta}{m}} \quad (6.59)$$

so that all four solutions will be $\pm\omega_1, \pm\omega_2$.

Before continuing, two observations can be made. (1) If one mass were held in place, and the equations solved for the other mass, the oscillation frequency would be $\omega_0 = \sqrt{(\beta + \beta_{12})/m}$. Therefore, the coupling for the two masses "splits" the oscillation frequency according to

$\omega_2 < \omega_o < \omega_1$. Including N (an even number) particles of mass m in the linear chain produce $N/2$ frequencies above and $N/2$ frequencies below ω_o , while for N an odd number produces $(N-1)/2$ above, $(N-1)/2$ below, and one equal to ω_o . Consequently, the number of modes (positive frequencies) must be the same as the number of masses. Also notice the number of degrees of freedom (DOF) for the atoms matches the number of allowed frequencies. (2) Substituting Equation 6.59 into the matrix Equation 6.58 produces the two solutions $B_1 = -B_2$ and $B_1 = B_2$, respectively. This shows that angular frequencies define the modes for the masses to move “180° out-of-phase” or “completely in phase” (i.e., the displacement between them does not change and they oscillate together).

The solutions u_1 and u_2 must be a linear combination of complex exponentials in time having the four possible frequencies listed below Equation 6.59. For $\pm\omega_1$, result #2 in the previous paragraph shows that u_1 and u_2 have terms that are negative of each other, while for $\pm\omega_2$, u_1 and u_2 have equal terms

$$\begin{aligned} u_1(t) &= +ae^{i\omega_1 t} + be^{-i\omega_1 t} + ce^{i\omega_2 t} + de^{-i\omega_2 t} \\ u_2(t) &= -ae^{i\omega_1 t} - be^{-i\omega_1 t} + ce^{i\omega_2 t} + de^{-i\omega_2 t} \end{aligned} \quad (6.60)$$

where a , b , c , and d are constants (see Section 4.5 for an alternate treatment). The important point here is that the motion of either atom has quite complicated time dependence being a mixture of two different Fourier components. The complexity arises because we focus on the individual atoms (i.e., u_n represents the coordinate of atom # n) rather than a simpler wave motion as described by the “normal coordinates” for which one focuses on specific collective motions of all the atoms as described next. The normal modes appear as sinusoidal waves in space similar to $\sin(k_x x)$ for fixed-endpoint conditions and oscillate in time. These fundamental modes can be Fourier superposed to describe the more complicated motions of each atom.

As mentioned, normal modes represent a simpler (and perhaps more intuitive) motion of the atoms (cf., Figure 6.25). The coordinates for the normal modes obtain from a linear combination of the atomic coordinates and resemble the coordinates for the motion of the center of mass and the group of atoms with respect to the center of mass. Define the following new coordinates

$$\begin{aligned} u_1 &= v_1 + v_2 \\ u_2 &= v_1 - v_2 \end{aligned} \quad \text{or equivalently} \quad \begin{aligned} v_1 &= (u_1 + u_2)/2 \\ v_2 &= (u_1 - u_2)/2 \end{aligned} \quad (6.61)$$

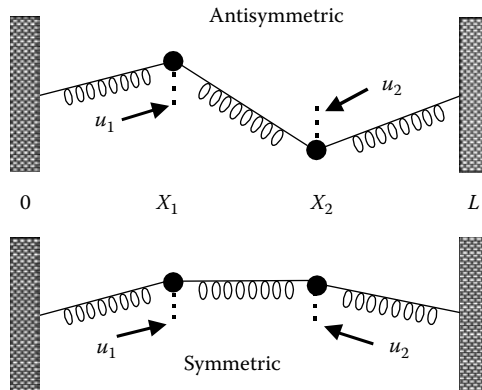


FIGURE 6.25 The two normal modes for transverse oscillations on a spring system with two masses confined to the single-transverse motion.

Substitute into Equation 6.56 and separate variables to find

$$\begin{aligned} m\ddot{v}_1 + (\beta + 2\beta_{12})v_1 &= 0 \\ m\ddot{v}_2 + \beta v_2 &= 0 \end{aligned} \quad (6.62)$$

The uncoupled solutions can be written as

$$\begin{aligned} v_1(t) &= a'e^{i\omega_1 t} + b'e^{-i\omega_1 t} \\ v_2(t) &= c'e^{i\omega_2 t} + d'e^{-i\omega_2 t} \end{aligned} \quad (6.63)$$

where a', b', c', d' are constants. Note that $v_1(t)$ corresponds to the larger frequency. The motion can be easily visualized for the specific initial conditions given in Table 6.1.

The first set of initial conditions corresponding to ω_1 provide a stationary center of mass and the two atoms oscillate 180° out-of-phase. The second set corresponding to ω_2 shows both atoms oscillate in phase which likewise gives the center-of-mass a sinusoidal time dependence.

Instead of the longitudinal waves shown in Figure 6.24, consider the transverse waves shown in Figure 6.25 where it is easy to see the antisymmetric character for ω_1 and the symmetric character for ω_2 . Notice the shape of the normal modes along the x -axis approximates a sine wave with wavelength either $\lambda = L$ or $\lambda = 2L$ which provides a wave vector of either $k = 2\pi/L$ or $k = \pi/L$. Notice further, the number of normal modes, frequencies and wave numbers k coincide with the number of degrees of freedom of 2 for the system. The number of degrees of freedom equals the number of dimensions that the particles can independently move. Each atom can move in one direction in this case but including the two atoms provides the 2 degrees of freedom. The “modes” of a system can refer to the frequencies, wave numbers, polarization, or shapes depending on how the term appears in context. For shape, one refers to the time-independent shape as the mode (a time-independent sinusoid in this case) but more exactly refers to the time-independent eigenfunctions of the wave equation.

The normal modes could have been found from Equation 6.56 and Figure 6.24 (longitudinal motion) right from the start by assuming a solution of the form

$$u_n = u(x_n, t) = A_k e^{ikx_n - i\omega_k t}$$

where

A represents the amplitude

x_n the equilibrium position has the value $x_n = na$ where a provides the atomic spacing at equilibrium

In this case (as demonstrated in the next topic), the time derivatives provide a relation between ω and k while the boundary conditions determine the allowed values of k (and hence, the allowed ω).

TABLE 6.1
Specific Examples for the Normal Modes

Initial Conditions	Solutions
$u_1(0) = -u_2(0)$	$v_1(t) = 0 \rightarrow u_1(t) = -u_2(t)$
$\dot{u}_1(0) = -\dot{u}_2(0)$	$\omega_1 = \sqrt{\frac{\beta + 2\beta_{12}}{m}}$
$u_1(0) = u_2(0)$	$v_2(t) = 0 \rightarrow u_1(t) = u_2(t)$
$\dot{u}_1(0) = \dot{u}_2(0)$	$\omega_2 = \sqrt{\frac{\beta}{m}}$

Section 4.5 discusses the theoretical basis for normal modes of coupled oscillators with attention to wave motion of a linear array of N masses coupled by quadratic potentials (i.e., springs). That section first focuses on the motion of each individual mass with coordinate u_n . That section shows there results an $N \times N$ determinant equation that must be solved for the fundamental frequencies (i.e., the frequencies of the normal modes). However, the following section shows that finding the solutions to the equations of motion do not require the $N \times N$ determinant equation as long as one starts with the normal modes. For the diatomic crystal, a 2×2 determinant will appear but it corresponds to the two atoms per basis and does not have the full size of say $2N$ -atoms.

6.6.2 EQUATIONS OF MOTION

Now we find the dispersion relation $\omega(k)$ for a monatomic crystal with atoms of mass m and lattice constant a as shown in Figure 6.26. As before, we denote the “equilibrium position” of atom $\#n$ by x_n . The coordinates x_n serve as indices rather than functions of time. The function $u_n = u(x_n)$ describes the displacement of atom $\#n$ from equilibrium where the time dependence of the oscillation has been suppressed for convenience; that is $u_n(x_n) \equiv u_n(x_n, t)$. We assume the atoms oscillate back and forth parallel to the direction of wave propagation. That is, Figure 6.26 represents a longitudinal wave since the atom displacement parallels the wave vector $\vec{k}$ for the wave. Figure 6.26 shows the “amount of stretch from equilibrium” for the spring between atoms $\#n$ and $\#(n+1)$ must be given by $u_{n+1} - u_n$. The two bonds on either side of atom $\#n$ produce two forces. Further, assume a single coupling constant β due to the symmetry of the crystal. Therefore, we can write the total force on atom $\#n$ as

$$F_{\text{tot}} = \beta[u_{n+1} - u_n] - \beta[u_n - u_{n-1}] \quad (6.64)$$

Atom $\#n$ obeys Newton’s second law

$$m \frac{d^2 u(x_n)}{dt^2} = \beta[u_{n+1} - u_n] - \beta[u_n - u_{n-1}] = \beta[u_{n+1} + u_{n-1} - 2u_n] \quad (6.65)$$

We already know that atoms execute simple harmonic motion. A solution to the differential equation has the form of a plane wave.

$$u_n = A \exp(ikx_n - i\omega_k t) = A \exp(inka - i\omega_k t) \quad (6.66)$$

where position index x_n can be written in terms of the lattice constants as $x_n = na$. We assume that the angular frequency is positive. A general solution would have the form of a Fourier sum over the plane waves.

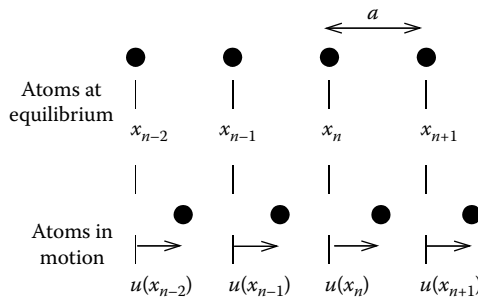


FIGURE 6.26 Top: Atoms at their equilibrium positions. Bottom: Atoms displaced from their equilibrium positions.

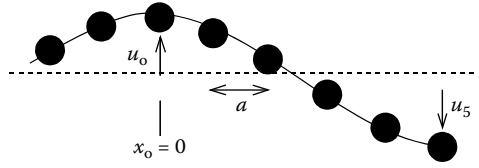


FIGURE 6.27 Transverse wave motion where atoms oscillate in a direction perpendicular to the direction of the wave vector.

Momentarily consider the transverse modes. We assume that these modes produce similar solutions to Newton's differential equation as the longitudinal modes in Equation 6.66. The transverse modes displace the atoms in a direction perpendicular to the direction of motion of the wave as shown in Figure 6.27. For example, we can set $n=0$ and watch atom #0 oscillate about its equilibrium position according to

$$u_n = A \exp(-i\omega_k t)$$

Or we can set $t=0$, and look at the collection of points

$$\{u_n = u(x_n) = A \exp(inka)\}$$

We know that the real part provides the wave depicted in Figure 6.27. The propagation wave vector $\vec{k}$ is related to the wavelength and not necessarily to the lattice spacing constant a .

The dispersion relation for phonons in the crystal lattice obtains by substituting Equation 6.66 into Equation 6.65

$$-m\omega^2 = \beta \{ \exp(ika) + \exp(-ika) - 2 \} = \beta \left\{ \exp\left(i\frac{ka}{2}\right) - \exp\left(-i\frac{ka}{2}\right) \right\}^2 = -4\beta \sin^2\left(\frac{ka}{2}\right).$$

Using the trigonometric angle doubling formula $\cos(ka) = 1 - 2\sin^2(ka/2)$, the last equation becomes

$$\omega_k = 2\sqrt{\frac{\beta}{m}} \left| \sin\left(\frac{ka}{2}\right) \right| = \sqrt{\frac{\beta}{m}} \sqrt{2[1 - \cos(ka)]} \quad (6.67)$$

Notice that this last equation uses the positive square root to keep the angular frequency positive.

Equation 6.67 represents the dispersion curve plotted in Figure 6.28. The dispersion curve repeats itself every $2\pi/a$ in k -space where a represents the equilibrium spacing between neighboring atoms. The region $(-\frac{\pi}{a}, \frac{\pi}{a})$ comprises the 'first Brillouin zone' (FBZ) (it is the Wigner-Seitz cell for

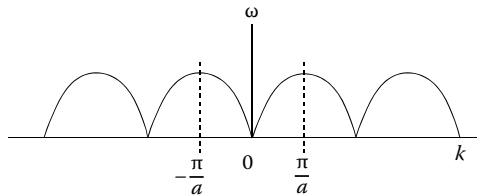


FIGURE 6.28 The dispersion curve for the monatomic crystal with 1-D motion.

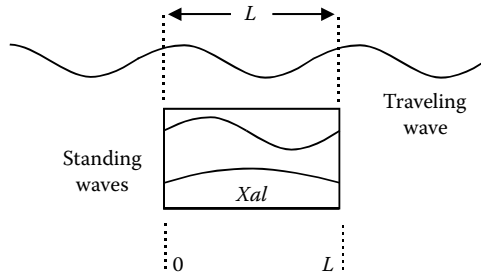


FIGURE 6.29 The phonon dispersion curves limited to the FBZ.

the reciprocal lattice). Closer atomic spacing (i.e., smaller unit cells) produces wider FBZs. Recall, the lowest order reciprocal lattice vector has the magnitude $G_1 = 2\pi/a$. We usually limit the dispersion curve to the FBZ as in Figure 6.29 since any other point can be reached by translation through a reciprocal lattice vector.

The curves in the FBZ repeat every reciprocal lattice vector $G_1 = 2\pi/a$. What does the zone boundary mean physically? First observe that the wave vector k must be distinct from the reciprocal lattice vectors and that the wave vectors must be related to the wavelength of the mode according to $k = 2\pi/\lambda$. Apparently $G_1/2$ represents the smallest phonon wavelength. As we will see, the smallest wavelength must be $2a$ which corresponds to adjacent atoms oscillating 180° out-of-phase. This gives a wave vector of $k = 2\pi/\lambda = \pi/a$ which exactly matches the value for the zone boundary. Given that the atoms oscillate 180° out-of-phase, the wave does not move when it has a wave vector at the zone boundary.

How small can we make the wave vector k ? Again using the relation $k = 2\pi/\lambda$, the smallest wave vector must correspond to the longest wavelength. For a finite size solid of length L , the longest wavelength can be no longer than $2L$ as shown in the bottom portion of Figure 6.30. These wavelengths form standing waves within the crystal and do not show the propagation of energy from say one end ($x = 0$) to the other end ($x = L$). However, traveling plane waves would show the propagation of energy. Usually models for wave motion in crystals use the traveling waves (plane waves) with periodic boundary conditions so that the wavelength cannot be larger than L as shown in the top portion of Figure 6.30. The periodic boundary conditions consider the finite crystal to be infinite and require the phonon wave function to repeat every distance L . Usually L is taken as the finite length of the physical crystal. Elementary studies of Fourier series show that these functions periodic in L must be made of those sinusoidal functions with wavelength a submultiple of L as in L/n . For either the fixed-endpoint or for the periodic boundary conditions, the longest phonon wavelength (and hence the shortest wave vector) must be determined by the physical size of the solid. At the other extreme, the smallest physical size of the inter-atomic spacing a determines the shortest wavelength and the largest wave vector.

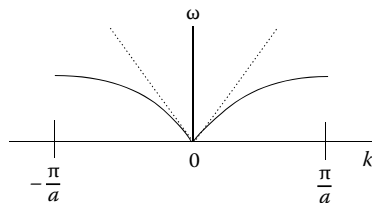


FIGURE 6.30 Longest wavelength for the periodic boundary conditions (top) is L and for the fixed-endpoint boundary condition (bottom) is $2L$.

6.6.3 PHONON GROUP VELOCITY FOR MONATOMIC CRYSTAL

The phase and group velocities, respectively, have the form

$$v = \frac{\omega(k)}{k} \quad \text{and} \quad v_g = \frac{\partial \omega}{\partial k} \quad (6.68)$$

as reviewed in Appendix F. The slope of the dispersion curve gives the group velocity. Near the origin where $k=0$, the phase and group velocity must be the same (refer to the dotted line in Figure 6.31). The group velocity refers to the motion of a wave packet (i.e., the Fourier sum of plane waves) and describes the speed with which energy (or “mass” in the case of quantum mechanical particles and wave functions) can be transferred. In particular, the wave packet consists of plane waves with various wave vectors that might, for example, superpose to produce a shape reminiscent of a Gaussian. With respect to the dispersion curve in Figure 6.31, the group velocity differs from the phase velocity away from the origin $k=0$. In particular, the two velocities differ near the ends of the FBZ (i.e., at $\pm\pi/a$). The group velocity has the value of 0 which means that a wave packet with average wave vector $k \approx \pi/a$ cannot propagate. The maximum angular frequency occurs at the edges of the FBZ. Substituting $k = \pi/a$ into the dispersion relation in Equation 6.67, specifically

$$\omega_k = \sqrt{\frac{\beta}{m}} \sqrt{2[1 - \cos(ka)]} \quad (6.69)$$

provides

$$\omega_{\max} = 2\sqrt{\frac{\beta}{m}} \quad (6.70)$$

The edges of the FBZ correspond to neighboring atoms moving in opposite directions as shown in Figure 6.32. The energy cannot propagate along the crystal. This is easy to see from Equation 6.66, specifically

$$u_n = A \exp(ikx_n - i\omega_k t) = A \exp(inka - i\omega_k t) \quad (6.71)$$

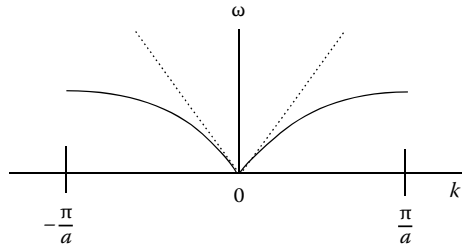


FIGURE 6.31 A phonon dispersion curve in the FBZ. The parameter a represents the spacing between atoms.



FIGURE 6.32 Motion of atoms at the FBZ boundaries.

by substituting $k = \pi/a$ to get

$$u_n = A \exp(inka - i\omega_k t) = A \exp(in\pi - i\omega_k t) \sim (-1)^n e^{i\omega t} \quad (6.72)$$

therefore $\frac{u_{n+1}}{u_n} = -1$ which indicates opposite motion (180° phase shift).

The group velocity for the monatomic chain can be calculated from the dispersion relation given in Equation 6.69 to find

$$\frac{\partial \omega_k}{\partial k} = \sqrt{\frac{\beta}{m}} a \cos\left(\frac{ka}{2}\right) \quad (6.73)$$

Clearly the group velocity is 0 at the edges of the FBZ. Near $k=0$, the group velocity must be “ $a \sqrt{\beta/m}$.” We can compare this with the phase velocity near $k=0$ using Equation 6.69. Taylor expanding the cosine to second order provides $\omega_k = \sqrt{\beta/m} \sqrt{(ka)^2} = ka \sqrt{\beta/m}$ and therefore the phase velocity must be $a \sqrt{\beta/m}$. As expected, the group and phase velocity agree near the point $k=0$.

A line of atoms can exhibit both transverse and longitudinal motion as indicated in Figure 6.33. There exist two transverse modes consisting of displacements along the x - or y -axes when the wave propagates along z . For the longitudinal modes, the atoms move parallel to the wave vector. The “spring constant” β can be different for each of the three motions. We might expect the same value β for the two transverse modes and a different β for the longitudinal mode. The spring constants β can be found by measuring the propagation speed near $k=0$, which is the speed of sound in the crystal, using

$$v_g|_{k=0} = \left. \frac{\partial \omega_k}{\partial k} \right|_{k=0} = \left. \sqrt{\frac{\beta}{m}} a \cos\left(\frac{ka}{2}\right) \right|_{k=0} = \sqrt{\frac{\beta}{m}} a$$

Figure 6.34 shows the dispersion relation with two transverse and one longitudinal modes.

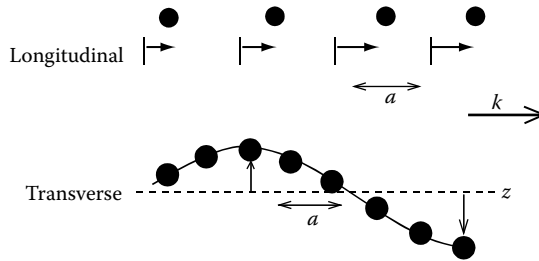


FIGURE 6.33 Top: Longitudinal motion of the atoms. Bottom: Transverse motion of the atoms.

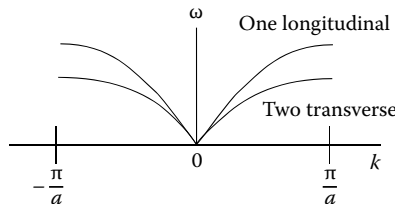


FIGURE 6.34 Dispersion diagram for 3-D motion along a 1-D crystal.

6.6.4 THREE-DIMENSIONAL MONATOMIC CRYSTALS

A 3-D crystal can have waves that propagate along any of the three orthogonal spatial directions. Each direction can support atoms oscillating transverse (perpendicular) or parallel to the direction of propagation. We have already discussed the parallel case—the longitudinal wave. For a symmetric crystal, we expect the spring constants in either of the two transverse directions to be similar. Therefore, as shown in Figure 6.34, the dispersion curves can be identical. The spring constant along the longitudinal direction can be different from the transverse direction and therefore produces a different dispersion curve.

6.6.5 LONGITUDINAL VIBRATION OF A ROD AND YOUNG'S MODULUS

The analysis of the previous section can be used to find a wave equation for the longitudinal vibrations of a solid rod rather than using the Lagrangian approach in Section 4.6. We consider a long thin rod free to vibrate only along the x -direction, which is also the direction for the propagation of the wave (Figure 6.35). The equation of motion (Equation 6.65) can be modified by replacing m by Δm

$$\Delta m \frac{d^2 u(x)}{dt^2} = \beta[u(x + \Delta x) - u(x)] - \beta[u(x) - u(x - \Delta x)] \quad (6.74)$$

where $u(x)$ is the horizontal displacement of the little bit of mass Δm that has its equilibrium position at point x . The variable x replaces an index i used to label each mass. We define the volume density of the rod by

$$\rho = \frac{\Delta m}{A \Delta x}$$

where A denotes the cross-sectional area of the rod. Substituting the volume density into Equation 6.74 provides the equation

$$\rho A \frac{d^2 u(x)}{dt^2} = \beta \left[\frac{u(x + \Delta x) - u(x)}{\Delta x} \right] - \beta \left[\frac{u(x) - u(x - \Delta x)}{\Delta x} \right]$$

Taking the limit $\Delta x \rightarrow 0$ of both sides and using the definition of derivative, we can write

$$\rho A \frac{d^2 u(x)}{dt^2} = \beta[u_x(x) - u_x(x - \Delta x)]$$

where the subscript x stands for the partial derivative with respect to x . Let us multiply and divide the right-hand side by Δx to find

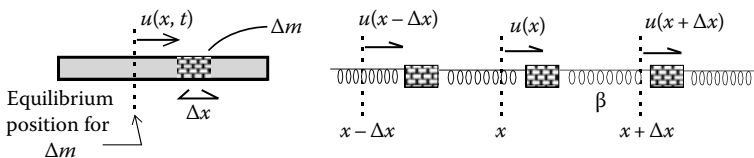


FIGURE 6.35 A long thin rod divided into small masses can be represented by a spring model.

$$\rho A \frac{d^2 u(x)}{dt^2} = \beta \Delta x \left[\frac{u_x(x) - u_x(x - \Delta x)}{\Delta x} \right]$$

Again taking the limit $\Delta x \rightarrow 0$ of both sides provides

$$\rho A \frac{d^2 u(x)}{dt^2} = \left(\lim_{\Delta x \rightarrow 0} \beta \Delta x \right) u_{xx}(x)$$

or equivalently

$$\frac{\partial^2 u(x, t)}{\partial x^2} = \frac{1}{E/\rho} \frac{\partial^2 u(x, t)}{\partial t^2}$$

where we define the speed of the wave must be

$$v = \sqrt{E/\rho}$$

and Young's modulus is defined to be

$$E = \frac{1}{A} \lim_{\Delta x \rightarrow 0} \beta \Delta x \quad (6.75)$$

which is essentially the spring constant per unit cross-sectional area.

At first, it might seem that Young's modulus should be zero ($E = 0$) because of the limit and the fact that we term β as a spring “constant” (emphasis on the word “constant”). We can correct this reasoning in the following manner. Consider a large spring stretched from equilibrium by a distance Δz where L represents the equilibrium length of the spring (Figure 6.36). The force on the mass at the end can be written as

$$F = -\beta \Delta z = -\frac{\beta'}{L} \Delta z \quad (6.76)$$

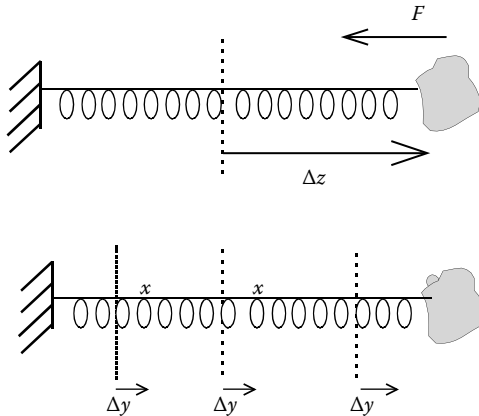


FIGURE 6.36 Dividing a long spring into smaller ones.

where the spring constant has been redefined as $\beta' = \beta/L$. Now consider the original spring to be made of N smaller springs as shown. Each little spring must be stretched by an amount Δy such that

$$\Delta z = N\Delta y$$

Now the force in Equation 6.76 can be written as

$$F = -\frac{\beta'}{L}\Delta z = -\frac{\beta'}{L}N\Delta y = -\frac{\beta'}{L}(N\Delta x)\frac{\Delta y}{\Delta x} \quad (6.77)$$

where Δx is the equilibrium length of each individual spring. Therefore, substituting the total length of the long spring $L = N\Delta x$, the force equation becomes

$$F = -\frac{\beta'}{L}L\frac{\Delta y}{\Delta x} = -\beta'\frac{\partial y}{\partial x} \quad (6.78)$$

Thus setting $\beta' = EA$ shows how Young's modulus E must be related to the force that the stretched bar applies to the mass at the end.

6.7 CLASSICAL PHONONS IN DIATOMIC LINEAR CRYSTAL

This section finds the dispersion curves for phonons in a cubic lattice with a two-atom cluster at each lattice point. The spacing between each cluster is $2a$ and the separation between each adjacent atom is a . The diatomic crystal supports optical and acoustic phonon dispersion curves. The width of the FBZ decreases by a factor of 2 since the lattice spacing increases by a factor of 2 as compared with monatomic crystal.

6.7.1 THE DISPERSION CURVES

The diatomic crystal appears in Figure 6.37 with large mass M and small mass m occupying alternating sites. Assume that the large atoms with mass M occupy the “even” numbered sites x_{2n} while the small atoms with mass m occupy the “odd” numbered ones x_{2n+1} . The integer n labels the lattice points. We consider longitudinal motion along the x -axis. Newton's law can be applied to each mass in a manner similar to the previous sections for the monatomic crystal. We consider the coordinates x_{2n} and x_{2n+1} as indices; they give the x -coordinate of the atomic “equilibrium” position. The symbol u represents the displacement of the atom from equilibrium. Focusing on the forces for the large atom at x_{2n} we find

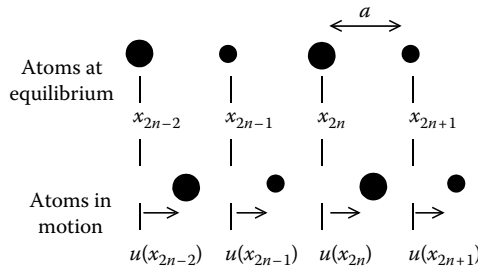


FIGURE 6.37 Top: Atoms at equilibrium. Bottom: Symbol u denotes the displacement from equilibrium. The x serves as an index.

$$\begin{aligned}
 M \frac{d^2}{dt^2} u(x_{2n}, t) &= \beta[u(x_{2n+1}, t) - u(x_{2n}, t)] - \beta[u(x_{2n}, t) - u(x_{2n-1}, t)] \\
 &= \beta[u(x_{2n+1}, t) + u(x_{2n-1}, t) - 2u(x_{2n}, t)]
 \end{aligned} \tag{6.79}$$

Similarly, the odd numbered atoms have the equation of motion of

$$m \frac{d^2}{dt^2} u(x_{2n+1}, t) = \beta[u(x_{2n+2}, t) + u(x_{2n}, t) - 2u(x_{2n+1}, t)] \tag{6.80}$$

Equations 6.79 and 6.80 have plane wave solutions. We can expect the atoms with different mass to oscillate with different amplitude. Assume the solutions have the form

$$u_k(x_{2n}, t) = \xi_k e^{ikx_{2n} - i\omega_k t} \quad u_k(x_{2n+1}, t) = \eta_k e^{ikx_{2n+1} - i\omega_k t} \tag{6.81}$$

where ξ_k, η_k denote complex amplitudes for the large and small atoms, respectively; they can contain phase information. Substituting for the indices

$$x_{2n} = 2na \quad x_{2n+1} = (2n+1)a$$

we find

$$u_k(x_{2n}, t) = \xi_k e^{i2nka - i\omega_k t} \quad u_k(x_{2n+1}, t) = \eta_k e^{i(2n+1)ka - i\omega_k t} \tag{6.82}$$

Substituting these solutions into Equations 6.79 and 6.80 and canceling common terms such as $e^{i2nka - i\omega_k t}$, we find

$$\begin{aligned}
 -M\xi_k \omega_k^2 &= \beta[\eta_k e^{ika} + \eta_k e^{-ika} - 2\xi_k] \\
 -m\eta_k \omega_k^2 &= \beta[\xi_k e^{ika} + \xi_k e^{-ika} - 2\eta_k]
 \end{aligned}$$

These equations can be rearranged as

$$\begin{bmatrix} 2\beta - M\omega_k^2 & -\beta(e^{ika} + e^{-ika}) \\ \beta(e^{ika} + e^{-ika}) & -(2\beta - m\omega_k^2) \end{bmatrix} \begin{bmatrix} \xi_k \\ \eta_k \end{bmatrix} = 0$$

If the 2×2 matrix has an inverse then the complex amplitudes ξ_k, η_k must necessarily be zero. In such a case, there would not be any wave motion at all contrary to common sense. Therefore, we require the 2×2 matrix to be singular in the sense that it does not have an inverse. This can only be accomplished by requiring the determinant of the 2×2 matrix to be zero. We find the following equation

$$(2\beta - M\omega_k^2)(2\beta - m\omega_k^2) - \beta(e^{ika} + e^{-ika})\beta(e^{ika} + e^{-ika}) = 0$$

Using $2 \cos(ka) = e^{ika} + e^{-ika}$, we solve for the angular frequency of the k th phonon mode to find

$$\omega^2(k) = \beta \left(\frac{1}{m} + \frac{1}{M} \right) \pm \beta \sqrt{\left(\frac{1}{m} + \frac{1}{M} \right)^2 - \frac{4 \sin^2(ka)}{mM}} \tag{6.83}$$

Equation 6.83 gives rise to two dispersion curves depending on the chosen sign. The “optical” phonons have the larger frequency (plus sign) compared with the “acoustic” phonons (minus sign).

Therefore we have longitudinal optical (LO) phonons and longitudinal acoustic (LA) phonons which differ in frequency for the same wave vector $\vec{k}$. As a note, one cannot fully specify the complex amplitudes ξ_n, η_n since further information would be required as to the driving force or thermal distribution.

6.7.2 APPROXIMATION FOR SMALL WAVE VECTOR

Finding the functional form of the dispersion curves “near the origin of k -space” (small k) reveals two distinct dispersion curves.

Starting with Equation 6.83, factoring out

$$\left(\frac{1}{m} + \frac{1}{M}\right)$$

from the radical, and using $\sqrt{1-x} \approx 1 - x/2 + \dots$ (small x) produces

$$\omega^2(k) \cong \beta \left(\frac{1}{m} + \frac{1}{M}\right) \pm \beta \left(\frac{1}{m} + \frac{1}{M}\right) \left[1 - \frac{2 \sin^2(ka)}{mM \left(\frac{1}{m} + \frac{1}{M}\right)^2}\right] \quad (6.84a)$$

Next, approximating $\sin(ka) \approx ka + \dots$ shows that the dispersion curves have the following form near $k = 0$.

$$\omega^2(k) \cong \beta \left(\frac{1}{m} + \frac{1}{M}\right) \pm \beta \left(\frac{1}{m} + \frac{1}{M}\right) \left[1 - \frac{2(ka)^2}{mM \left(\frac{1}{m} + \frac{1}{M}\right)^2}\right] \quad (6.84b)$$

Therefore the plus sign provides

$$\omega_{LO}(k) \cong \sqrt{2\beta \left(\frac{1}{m} + \frac{1}{M}\right) - \frac{2\beta(ka)^2}{(m+M)}} \cong \sqrt{2\beta \left(\frac{1}{m} + \frac{1}{M}\right)} \quad (6.85a)$$

and the minus sign gives us

$$\omega_{LA}(k) \cong ka \sqrt{\frac{2\beta}{m+M}} \quad (6.85b)$$

Clearly as $k \rightarrow 0$, the last equation shows $\omega \rightarrow 0$ while that in Equation 6.85a shows $\omega > 0$.

6.7.3 DISCUSSION

Figure 6.38 shows the two branches for the phonons in the diatomic crystal. The dispersion curve for the LA phonons gives the speed of sound in the crystal (the slope for small k)

$$v_{\text{phase}} = a \sqrt{\frac{2\beta}{m+M}} ka \cong 0 \quad (6.86)$$

The group velocity for the LA phonon is seen to be identical with the phase velocity near the origin of k -space but the group velocity is zero near the edges of the FBZ.

$$v_{\text{group}} = \frac{\partial \omega}{\partial k} \quad (6.87)$$

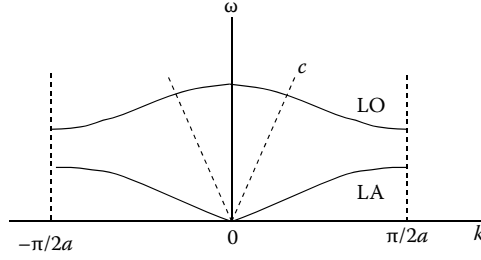


FIGURE 6.38 Dispersion curves for LA and LO waves.

The LA phonons represent motion of the whole unit cell consisting of two atoms with the unit cells separated by a distance of $2a$. On the contrary, the optical phonons have high-frequency waves due to the motion of neighboring atoms in opposite directions (especially for $k=0$); the unit cells as a whole remain fixed in space. If the two atoms in the unit cell carry opposite net charge such as for an ionic crystal like salt (i.e., maybe m is positive while M is negative), then an oscillating electromagnetic field with a frequency equivalent to the LO frequency can excite the LO mode. This is because, for the LO mode, the adjacent atoms move in opposite directions to form an electric dipole which can interact with the optical EM field. The line c represents the speed of the EM wave in the crystal; it does not intersect the LA branch. Notice also that the first Brillouin zone (FBZ) is half the size of that for the monatomic crystal (but now there are two branches). However, the size of the unit cell is $2a$ for our atom spacing of a and so the FBZ has a width of $2\pi/2a = \pi/a$.

A few comments should be made regarding the (1) oscillation frequency ω_k for optical and acoustic phonons and (2) the phase velocity for the acoustic branches. For simplicity, imagine transverse waves. One might wonder why the acoustic mode has decreased frequency for smaller k and why the phase velocity approaches a constant. And yet, the optical modes have the larger oscillation frequencies. The answer concerns the “spring constant” for the collection of atoms. Recall from the previous section, that the spring constant really depends on the length of the spring $\beta = \beta_o/L$, where L represents the equilibrium length of the spring and β_o is a constant. For the acoustic modes, the atoms tend to vibrate in phase. If m represents the mass of a single atom and if the length L of atoms vibrate together, then the displaced mass m_d will be of the form $m_d \sim mL$. The oscillation frequency has the usual definition of $\omega = \sqrt{\beta/m_d}$. Combining the relations provides

$$\omega = \sqrt{\beta/m_d} \sim \sqrt{\beta_o/m}/L = \omega_o/L \quad (6.88a)$$

One can see that longer lines of oscillating atoms should decrease the frequency. The longer lines correspond to smaller

$$k = 2\pi/\lambda \sim 2\pi/L \quad \text{or} \quad k = k_o/L \quad (6.88b)$$

where k_o represents a constant of proportionality. This last expression says the wave vector decreases with L as it was set up to do. Comparing these last two equations shows that for acoustic phonons, the frequency can be expected to decrease as k decreases. Furthermore, one can see the phase velocity approaches a constant since

$$v = \omega/k \sim (\omega_o/L)/(k_o/L) = \omega_o/k_o \quad (6.89)$$

The optical mode has large frequency since the adjacent atoms move out-of-phase. As a result, L takes on a very small value which requires the “springs” to be very stiff. Then Equation 6.88a shows the frequency will be quite large.

Finally as a comment for 3-D diatomic crystals, there are three acoustic and three optical branches. Often the three curves are termed “polarizations” since they describe the direction of motion of the atoms with respect to the direction of propagation. A group of three curves describing the polarization must consist of two curves for the two transverse wave motions and one for the longitudinal motion. For each wave vector $\vec{k}$, there will be three polarizations. The number of phonon branches (i.e., the number of groups of three polarizations) will equal the number of atoms in the lattice cell.

6.8 PHONONS AND MODES

Phonons represent the elementary unit of energy for the wave motion of atoms in a crystal. These sinusoidal waves defined by the wave vector $\vec{k}$ and the angular frequency ω_k represent the modes (or states) for the phonons. The number of phonons (with wave vector $\vec{k}$ and frequency ω_k) determine the amplitude of the sinusoidal wave with that frequency and wave vector. A given mode can have any number of phonons. The boundary conditions placed on the wave motion determine the allowed wave vectors $\vec{k}$ (and hence allowed phonon states) within the FBZ. The phonon modes for dispersion curves can be specified by both ω_k and $\vec{k}$. In many cases, we are interested in the number of phonons within a given range of frequencies. The easiest method to count them consists of finding the number of modes falling within the range of frequency and then multiplying by the number of phonons in each mode. The number of phonons per mode describes the amplitude of the wave corresponding to the mode and having parameters ω_k and $\vec{k}$. The section concludes with an introduction to the particle aspects of the phonon.

6.8.1 MODES IN MONATOMIC 1-D FINITE CRYSTAL WITH 1-D MOTION AND FIXED-ENDPOINT BOUNDARY CONDITIONS

The number of phonons in a given energy range can be calculated by multiplying the number of states in that range by the number of phonons in each state as will be discussed in subsequent sections. To perform the calculation, one must first know the allowed states and how to describe them with wave vectors $\vec{k}$. For now, we will not distinguish between the terms “mode” and “state” but perhaps “mode” more frequently refers to wave motion while “state” concerns the phonon aspects as a quantum of energy. The mode or state corresponds to the allowed wave vector $\vec{k}$ with a corresponding frequency ω_k or energy $E_k = \hbar\omega_k$ as represented, for example, by the open circles in Figure 6.39. A single phonon “occupies” a particular mode ($\vec{k}, \omega_k$) when the corresponding sinusoidal wave has the minimum oscillation amplitude (neglecting the zero-point motion). Adding a second phonon to the state further increases the amplitude but leaves the wave vector and frequency unaltered. Adding phonons to the state loosely correspond to thinking of the open circles in Figure 6.39 as “buckets” and adding the phonons to the buckets. Given the states in Figure 6.39

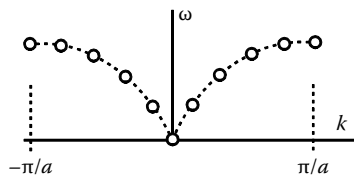


FIGURE 6.39 Allowed states ($\vec{k}, \omega_k$) for the monatomic linear crystal with atomic spacing a .

represent a normal mode (specific-oscillation frequency and wave vector), “adding a phonon to a state” corresponds to increasing the amplitude of the corresponding sinusoidal wave that can extend across the entire crystal. That is, the oscillation of all atoms share the single phonon. The problem of specifying the modes reduces to finding the allowed wave vectors $\vec{k}$.

The exact description of the mode depends on the boundary conditions applied to the atomic wave motion. Consider for example, a crystal of length L comprised of $N + 1$ atoms. We require the wave represented by $u(x, t)$ to satisfy either “fixed end point” or “periodic” boundary conditions. The fixed end point conditions typically take the form $u(0) = 0 = u(L)$ where $x = 0$ and $x = L$ represent the two ends of the crystal. The end atoms are fixed in place although it is not clear how one could do this since for a free-standing crystal, nothing clamps the motion of the end atoms. These fixed-endpoint boundary conditions give rise to “standing” sinusoidal waves. Conceptually, standing waves do not propagate and therefore do not transport energy. The standing waves have the form $\sqrt{2/L} \sin(kx)$ which are normalized to “one” in the sense of Chapter 2. Notice that only the “shape” of the sine wave defines the mode (i.e., the value of k , and hence ω_k) and not the amplitude. The standing waves consist of oppositely propagating traveling waves.

As an exercise, consider the case of the fixed-endpoint boundary conditions that should only very skeptically be applied to atomic systems since one cannot guarantee that the boundary conditions are satisfied. We want to calculate the number of different sine waves (i.e., the number of modes) that fit in the length L . The number of possible modes must be finite since (1) the sine wave must fit within a finite length (boundary conditions) and (2) wavelengths can only be as small as $2a$ (twice the lattice constant). The second requirement comes from the fact that when $\lambda = 2a$ when adjacent atoms move 180° out-of-phase and the wave does not propagate. The wave vector k then has a value of π/a at the edge of the first Brillouin zone (FBZ).

Assume a very long line of $N + 1$ atoms (i.e., N large) with the first one and last one fixed in place. The longest possible wavelength for the system appears in Figure 6.40. The atoms move together in unison—a collective motion—as expected for normal modes as discussed in Section 4.5. The figure suggests that full and half wavelengths must fit in the length L according to

$$\lambda = \frac{2L}{1}, \frac{2L}{2}, \dots, \frac{2L}{n}, \dots \quad (6.90a)$$

where n must be finite. These wavelengths correspond to wave vectors

$$k = \frac{2\pi}{\lambda} = \frac{n\pi}{L} \quad (6.90b)$$

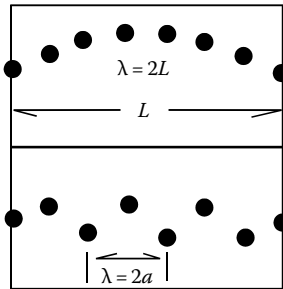


FIGURE 6.40 The maximum and minimum wavelength.

For large numbers of atoms N , the bottom panel in Figure 6.40 shows the minimum wavelength must be $\lambda = 2a$ where a represents the lattice constant. The integer n in Equation 6.90a would then be no larger than

$$n = \frac{2L}{2a} = N \quad (6.90c)$$

as found by substituting $\lambda = 2a$ into Equation 6.90a. The allowed values of k then have the form in Equation 6.90b with $n = 1, 2, \dots, N$. However, the correct expression can be no larger than the number of atoms able to move ($N - 1$ in this case). So what happened?

To resolve this issue, consider the sequence of atoms shown in Figure 6.41 and assume transverse wave motion along a single direction. First note for $N + 1 = 2$ atoms in the line that these two atoms must be fixed and wave motion cannot occur. In this case we begin to suspect the number of modes must be $N - 1$ since in this case $n = N - 1 = 0$. Next consider the case of $N + 1 = 3$. The middle atom can move up and down. The number of modes is $n = N - 1 = 1$. For $N + 1 = 4$, there must be two modes as shown in Figure 6.41. In general, the number of modes equals the number of atoms free to move so that $n = N - 1$ where the number of atom in the line is $N + 1$.

Now here is the reason that the simple formula in Equation 6.90c does not give the correct value. Consider the case of $N + 1 = 4$ shown at the bottom of Figure 6.41. The maximum and minimum displacement of the two free atoms do not line up with the maximum and minimum of the sine wave that would result from the fixed-endpoint boundary conditions since the sine wave must be zero at the position of atom #0 and atom #4. The wavelength cannot be exactly equal to $2a$ but must be slightly larger with a value of $\lambda = 3a$ in this case (three lengths a fit between the two end atoms in the bottom of Figure 6.41). Therefore, Equation 6.90a gives the number of modes as

$$n = \frac{2}{3} \frac{L}{a} = \frac{2}{3} \frac{3a}{a} = 2$$

We again recover the number of modes as $N - 1$. The correct formulas must be

$$\lambda = \frac{2L}{n} = \frac{2Na}{n} \quad \text{with } n = 1, 2, \dots, N - 1 \quad (6.91)$$

$$k = \frac{2\pi}{\lambda} = \frac{n\pi}{L} = \frac{n\pi}{Na} \quad \text{with } n = 1, 2, \dots, N - 1 \quad (6.92)$$

Two notes are in order. First, note that k has only positive values. This occurs because the standing waves only use positive values. Furthermore, Chapter 2 showed that the basis vectors of the form $\sin(kx)$ and negative values for the k do not produce new basis vectors. For density of states for 3-D crystals, one only counts the k states in the positive octant of a sphere where k_x, k_y, k_z all have positive values. Second, a normal size crystal has on the order of $N = 10^{24}$ atoms or approximately

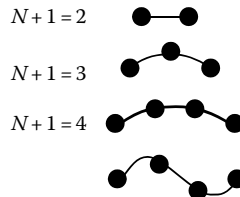


FIGURE 6.41 The four atom chain has two modes. Note that the wavelength for two atoms must be larger than $2a$.

10^8 atoms per side, which fit in approximately 1 cm. Clearly, the number of modes can be well approximated by $n = N - 1 \sim N$ as given by Equation 6.90c. For large N , the minimum wavelength in Equation 6.92 becomes $2a$.

6.8.2 PERIODIC BOUNDARY CONDITIONS

Here, we discuss the more commonly applied periodic boundary conditions requiring the wave $u(x, t)$ to be periodic over the length L according to $u(x) = u(x + L)$. Such a boundary condition places restrictions on the wave vector k in the plane waves $e^{ikx - i\omega_k t}$ comprising the Fourier series expansion of the function u . Contrary to the fixed-endpoint conditions, the wave vector can be positive, negative, or zero. Positive values produce waves traveling along the positive axis and negative ones produce waves traveling along the negative axis. The periodic boundary conditions normally apply to systems of infinite extent for which case, the length L resembles an arbitrary normalization length. However, the periodic boundary conditions can be applied to a finite crystal of length L by an artificial construction that places multiple copies of the finite crystal next to each other so as to fill all space (Figure 6.42). Ultimately, the exact form of the boundary condition does not affect the physics.

We apply the periodic boundary conditions $u(x + L) = u(x)$ to a very large crystal with a phonon wave extending over many atoms as shown in Figure 6.42. For this case, none of the atoms in the crystal need remain fixed in space as a phonon propagates. We assume that the phonon wave function $u(x + L, t) = u(x, t)$ repeats itself over the large distance L . The waves do not need to have the same phase, just the same wavelength (or smaller) in such a way that the wave repeats itself. The wave can move either right or left. Adding a phonon to a mode corresponds to adding a single quantum to the collective oscillation of the atoms (i.e., a normal mode) across the length L . A general wave periodic on the length L can be described as a Fourier sum of traveling waves

$$u(x, t) = \sum_k \xi_k e^{ikx - i\omega_k t} \quad (6.93)$$

where $k = 2\pi/\lambda$. The index $x = sa$ refers to the equilibrium position of atom $\#s$ similar to Sections 6.6 and 6.7 where a refers to the lattice spacing. In general, artificially imposing periodicity on the length L does not interfere with the physics.

Figure 6.42 shows that the maximum wavelength must be $\lambda_{\max} = L$. Other wavelengths fit in the length L according to

$$\lambda = \frac{L}{1}, \frac{L}{2}, \dots, \frac{L}{m}, \dots \quad (6.94a)$$

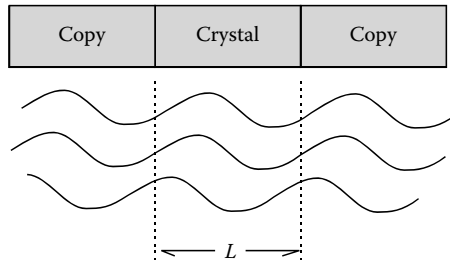


FIGURE 6.42 Longest wavelength satisfying periodic boundary conditions over the length L . Notice the waves do not need to be zero at the dotted lines.

The wave vectors must then have the form

$$k = \frac{2\pi}{\lambda} = 0, \pm \frac{2\pi}{L}, \dots, \pm \frac{2m\pi}{L}, \dots \quad (6.94b)$$

The same results can be deduced using the Fourier series in Equation 6.93 and requiring $u(x) = u(x+L)$ which then requires $e^{ikL} = 1$ and therefore reproduces Equation 6.94b. Positive values of k signify a wave propagating along the positive axis and negative values signify a wave moving along the negative direction. For the periodic boundary conditions, the minimum k can be zero (corresponding to $\lambda = \infty$) because the whole line of atoms might be displaced. The case of $k=0$ corresponding to the wave function $u(x) = c$, where $c = \text{constant}$, certainly satisfies the periodic boundary condition of $u(x+L) = c = u(x)$.

We need to find the largest integer m in Equation 6.94b. The smallest wavelength corresponds to two adjacent atoms vibrating 180° out-of-phase so that $\lambda_{\min} = 2a$. Unlike a very-small finite crystal, the infinite one can have the minimum wavelength of $2a$ because none of the atoms remain fixed in place as a wave passes through.

We assume the crystal has length L consisting of $N+1$ atoms with spacing a so that $L = Na$. However, we further use an odd number of atoms (i.e., $N+1 = \text{odd}$, $N = \text{even}$) in length L . Figure 6.43 shows that for even integers $N+1$, multiples of the smallest wavelength $\lambda_{\min} = 2a$ do not fit in the length L . In a real crystal, the number of atoms must be on the order of 10^{24} and so ± 1 makes little difference.

We can write the possible wavelengths and k -vectors. Setting

$$\lambda_{\min} = \frac{L}{n_{\max}} = 2a \quad \text{and} \quad L = Na \quad N = \text{even} \quad (6.95)$$

in Equation 6.94 provides

$$n_{\max} = \frac{L}{2a} = \frac{N}{2} \quad (6.96)$$

so that

$$\lambda = \frac{L}{n} = \infty, \frac{Na}{1}, \frac{Na}{2}, \dots, 2a \quad \Rightarrow \quad k_n = \frac{2\pi}{\lambda_n} = \frac{2n\pi}{L} = \frac{2n\pi}{Na} \quad (6.97a)$$

One issue remains concerning the number of modes and the maximum value of n . The simple monatomic crystal in this case has N -atoms in the length L capable of 1-D motion, which produces N degrees of freedom. If we were to take $n = 0, \pm 1, \pm 2, \dots, \pm N/2$ in Equation 6.97a (and also 6.97b below), there would be $N+1$ modes rather than the N required by the number of degrees of freedom. The issue can be resolved by noting the two motions due to $\pm N/2$ are not really

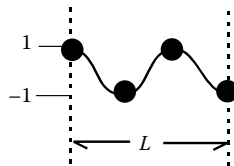


FIGURE 6.43 For L spanning an even number of atoms, the smallest waveform is not periodic on L . As shown $u(0) = 1$ and $u(L) = -1$ contrary to the requirement $u(0) = u(L)$ for periodic boundary conditions.

independent by considering $e^{iks a}$ in the Fourier expansion using $n = \pm N/2$ which gives the two values of $k_n = \pm \pi/a$. We find both values produce the same number $(-1)^s = e^{is\pi} = e^{-is\pi}$. Therefore, one should restrict the range to

$$n = 0, \pm 1, \pm 2, \dots, \pm \frac{N-1}{2}, \frac{N}{2} \quad (6.97b)$$

Each branch of a dispersion curve will therefore have the same number of allowed k -states in the FBZ as there are atoms in the linear chain in the length L . One should notice that the “spring constant” does not affect the number of modes. Rather it affects the slope of the dispersion relation (and group velocity).

In general, whether for phonons or photons, one starts with the classical description of the phenomena. This means solving a boundary value problem (i.e., often a wave equation with boundary conditions) and finding the basic modes of the system. The modes (either standing waves or traveling waves) account for the basic geometry of the system. For example, one might have metal spheres placed in a room and attempt to solve the wave equation for light. The basic modes of the system can then be superposed to find the classical solution and the general form of the wave. Now, adding a quantum of energy to a specific mode does not alter the shape of the mode because the mode is often defined independent of amplitude—only the shape of the wave counts to define the mode, the amplitude is often normalized to one. Adding a quantum of energy (i.e., a particle) does affect the amplitude of the physical wave (but not of the mode when thought of as a “bucket” to hold the quanta—buckets do not change!).

6.8.3 MODES FOR 2-D AND 3-D WAVES ON LINEAR MONATOMIC ARRAY

The 2-D and 3-D motion on a monatomic linear array increases the number of possible modes compared with 1-D motion on the same linear array. Here, by mode, we explicitly refer to the wave motion impressed on the 1-D array of atoms in the chain. The modes are characterized by the wave vector, angular frequency, and polarization (transverse or longitudinal).

One can count the total number of modes by counting the number of degrees of freedom for the N -atoms in the crystal. If one allows 1-D motion along the z -direction, for example, with the energy propagating along x , then the motion constitutes a transverse wave. There are N -atoms, which equals the number of allowed k -states in the first Brillouin zone (FBZ), and hence equals the number of possible total states for the wave. For 2-D atomic motion along for example, the y - and z -directions with the wave propagating along x (linear 1-D chain of atoms), there will be 2 degrees of freedom for each atom (motion in y and z) which produces two polarizations in the collective oscillation mode for the waves. For N -atoms, there will be a total of $2N$ degrees of freedom, or alternatively twice the number of allowed k -vectors in the FBZ as for the 1-D motion. Therefore, one can either determine the total number of possible modes by multiplying the number of degrees of freedom per atom with the number of atoms, or by multiplying the number of polarizations by the number of allowed k -vectors.

As an example, Figure 6.44 shows three acoustic polarizations for the monatomic crystal. We assume the wave can only propagate in one direction along x , which means the wave vectors have only the k_x component. The figure assumes distinct “spring constants” for each direction x, y, z with the relation $\beta_1 < \beta_2 < \beta_3$. The spring constant only affects the shape of the branch. We know that the dispersion curves must be arranged as shown because of the formula for the dispersion curves repeated here

$$\omega_k = 2\sqrt{\frac{\beta}{m}} \left| \sin\left(\frac{ka}{2}\right) \right| = \sqrt{\frac{\beta}{m}} \sqrt{2[1 - \cos(ka)]}$$

The spring constant does not affect the boundary conditions nor do we use it to account for the spacing of the atoms at equilibrium. Therefore the total number of modes for the three branches must be three times that for a single branch.

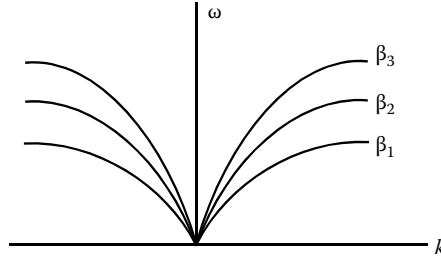


FIGURE 6.44 Three acoustic branches for 3-D motion in a linear monatomic array.

If one considers a diatomic crystal with unit cell size a but with two atoms in each cell, then the number of atoms will be $2N$ (for example, in the crystal length L), but the number of k -states for an acoustic branch (assume 1-D transverse motion) will be N (i.e., half the number of atoms). As a result, one should expect to find a transverse optical branch with N states. The total number of states must be $2N$ (half from each branch) to match the total number of degrees of freedom. The alternative calculation for this 1-D case, simply counts the number of allowed k -vectors and multiplies by the number of branches.

6.8.4 MODES FOR THE 2-D AND 3-D CRYSTAL

Previous sections show that the wave motion in 1-D crystals can be pictured as a sine wave along the x -axis for example. In the 2-D case, a rectangular array of atoms must have sine waves along the x - and y -axes as indicated in Figure 6.45. The wave has the form

$$u \sim e^{i\vec{k} \cdot \vec{r} - i\omega t} = e^{i(k_x x + k_y y) - i\omega t} \quad (6.98)$$

whereas the standing waves for fixed-endpoint boundary conditions consist of linear combinations of Equation 6.98 and have the form

$$u \sim \sin(k_x x) \sin(k_y y) \quad (6.99)$$

Periodic boundary conditions support traveling waves. The figure shows that the crystal can have different lengths L_x and L_y on the two sides.

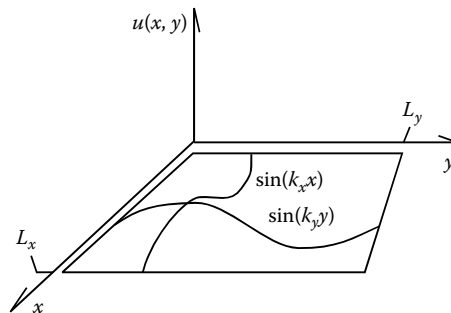


FIGURE 6.45 Wave motion on a finite 2-D crystal.

The conditions on the 2-D wave vector $\vec{k} = \tilde{x}k_x + \tilde{y}k_y$ can be found from the picture or from the usual Fourier series

$$u(\vec{r}, t) = \sum_{\vec{k}} u_{\vec{k}}(t) e^{i\vec{k} \cdot \vec{r}} = \sum_{\vec{k}} u_{\vec{k}}(t) e^{i(k_x x + k_y y)}$$

Suppose $\vec{L}$ denotes a vector representing the crystal size as $\vec{L} = \tilde{x}L_x + \tilde{y}L_y$. Then the periodic boundary conditions over either dimension of the crystal take the form

$$\sum_{\vec{k}} u_{\vec{k}}(t) e^{i\vec{k} \cdot (\vec{r} + \vec{L})} = \sum_{\vec{k}} u_{\vec{k}}(t) e^{i\vec{k} \cdot \vec{r}}$$

which then requires $e^{i\vec{k} \cdot \vec{L}} = 1$ so that

$$k_x = \frac{2\pi}{\lambda_x} = \frac{2m_x\pi}{L_x} \quad k_y = \frac{2\pi}{\lambda_y} = \frac{2m_y\pi}{L_y} \quad (6.100)$$

where $m_x, m_y = 0, \pm 1, \pm 2, \dots$ and the minimum wavelength must be larger than $2a$. The same results obtain from the figure by calculating the wavelength in the x - and y -directions as

$$\lambda_x = \frac{2L_x}{m_x} \quad \lambda_y = \frac{2L_y}{m_y}$$

If each direction has N -atoms on a side then in total, there must be N^2 allowed points in (k_x, k_y) space. Now if we allow each propagation direction to have two transverse and a longitudinal mode then the total number of states must be $3N^2$ matching the 3 degrees of freedom per atom multiplied by the N^2 atoms. For the diatomic case with N as the number of clusters consisting of two atoms, there are a total of $2N^2$ atoms and the total number of degrees of freedom must be $6N^2$. Therefore, one can expect N^2 states on each surface (ω vs. k_x, k_y) with six such surfaces for all of the longitudinal and optical modes.

6.8.5 AMPLITUDE AND PHONONS

So far we have discussed the relation of the boundary conditions to the k -vectors and these produce the allowed modes for the waves. The quantization of the wave motion results in the phonon as the quantum of energy for the wave motion. Initially we view the motion of the individual atoms about equilibrium as comprising the wave motion in the crystal. Each atomic harmonic oscillator has an integral number of quanta associated with it. Now for the wave motion, which we view as a harmonic oscillator in its entirety, a quantum of energy is associated with the collective motion of all the atoms rather than one individual atom. Adding a phonon to a mode characterized by $\vec{k}$ and ω_k (and, of course, a polarization direction) increases the amplitude of the wave across the crystal.

We briefly show how adding a phonon to a mode can increase the amplitude but leave the actual quantization procedure for a subsequent section in this chapter. We ignore the zero-point motion. The calculation will first find the total average energy of each atom consisting of the average kinetic and potential energy. Then the average energy of all atoms in the chain can be found by multiplying the average total energy per atom by the number of atoms in the chain. This will be equated to the energy of n_{Ξ} phonons with energy $n_{\Xi}\hbar\omega_k$. The case $n_{\Xi} = 1$ describes a single phonon in the mode (and ignores zero-point motion) viewed as the collective motion of all the atoms rather than one specific single atom.

Consider a monatomic linear crystal with N -atoms in length L and atomic spacing a executing motion in a single polarization direction. As before, the wave will be given by

$$u(x, t) = u_0 \sin(ksa - \omega t) \quad (6.101)$$

for atom # s at equilibrium position $x = sa$. The kinetic energy T varies in time according to

$$T(t) = \frac{mu_0^2 \omega_k^2 \sin^2(ksa - \omega t)}{2} \quad (6.102)$$

where m represents the mass of each atom. The average kinetic energy is $1/2$ times the peak and the average potential energy is the same as the average kinetic energy for a harmonic oscillator. We find the average total energy $\bar{E}_a$ of each oscillating atom must be

$$\bar{E}_a = \frac{mu_0^2 \omega_k^2}{2} \quad (6.103)$$

For N -atoms in the linear chain of length L , the total average energy of all atoms becomes $\bar{E} = N\bar{E}_a$. Equating this to the energy $n\hbar\omega_k$ of n phonons in the mode described by (k, ω_k) provides a wave amplitude proportional to $\sqrt{n\Xi}$

$$u_0 = \sqrt{\frac{2n\Xi\hbar}{m\omega_k N}} \quad \text{and} \quad u(x, t) = \sqrt{\frac{2n\Xi\hbar}{m\omega_k N}} \sin(ksa - \omega t) \quad (6.104)$$

One should realize that a transverse wave (with polarization along the z -direction) and propagating along the x -direction, for example, appears to manifest energy as displacements of atoms along z . This means that the “quantum particle—the phonon” moves along the x -direction. However, notice that the atoms only move (oscillate) along the z -direction. The particle aspects of the phonon are “imbedded” in the wave motion and not immediately discernable.

6.9 THE PHONON DENSITY OF STATES

The phonon density of states (P-DOS) provides the number of states that phonons can occupy in a given range of frequency for a unit volume of crystal. Knowing the probable number of phonons occupying each state then allows one to calculate the number of phonons in a given range of energy. The results apply to conduction, specific heat, carrier trapping, and band transition mechanisms. The section provides a method to calculate P-DOS for 1-D, 2-D, and 3-D crystals.

6.9.1 INTRODUCTORY DISCUSSION

The phonon generally occupies “extended states” in a material meaning that infinite plane waves represent the modes (especially for the periodic boundary condition for the infinite crystal). However, one can speculate on the ability of phonons to also occupy localized states which correspond to microscopic regions capable of supporting standing waves. The present section focuses on the plane waves and determines the number of available modes (phonon states) within a given frequency range (ω_1, ω_2) , which is equivalent to finding the number within a range of energy since $E = \hbar\omega$ (c.f., Figure 6.46). The density of phonon states describes the number of available states per unit energy per unit crystal volume. The definition divides out the crystal volume in order to treat the density of states as a material property independent of the size of the crystal. One might expect to find the number of phonons (per volume) in a given range of angular frequency (ω_1, ω_2) (or equivalently energy) as follows:

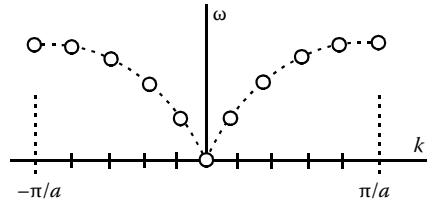


FIGURE 6.46 Phonon states in an acoustic branch.

$$\eta = \sum \# \text{ states} * \frac{\# \text{ phonons}}{\text{State}} \rightarrow \int_{\omega_1}^{\omega_2} d\omega g(\omega) n(\omega)$$

where

$n(\omega)$ represents the average number of phonons per state at frequency ω
 $g(\omega)$ represents the density of states

We will see in a subsequent section that the average number of phonons in a given state can be calculated from the Bose–Einstein probability distribution $P(n)$ for equilibrium.

The allowed phonon frequencies used to calculate $g(\omega)$ can be traced back to the periodic boundary conditions and the resulting allowed k -states. An example appears in Figure 6.46 for the single acoustic branch for a monatomic crystal with atomic spacing a . The number of states in the FBZ must be given by

$$n = \frac{\text{FBZ width}}{\text{Minimum } k\text{-spacing}} = \frac{2\pi/a}{2\pi/L} = \frac{L}{a} = N$$

where N (on the order of 10^{24}) represents the number of atoms. The figure shows the same total number of k -states as for frequency states; however, the k -states have equal spacing but not the frequency ones. For Figure 6.46, there are 10 states when counted by k value or frequency value. Then eleventh one is part of another Brillouin zone.

The density of states $g(\omega)$ refers to the number of angular-frequency states rather than the number of allowed states per unit k -length (it is probably easier to think of $g(E)$ for physical applications). Figure 6.47 represents the allowed states by the dots on the dispersion curve. For $g(\omega)$, the number of states included in the range $\Delta\omega$ must include those states from both positive and negative values as shown, for example, in Figure 6.47 (monatomic, 1-D crystal). In this simple example, one counts approximately six states on either end for a range $\Delta\omega = 0.01$ (for example) which yields 1200 states/Hz (and per volume). Obviously, we need some relations between the k and ω values in order to count the allowed states. Those relations will certainly, ultimately involve the dispersion

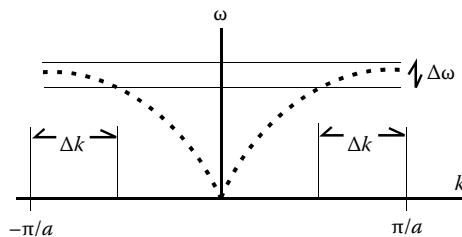


FIGURE 6.47 The states are equally spaced along the k -axis but the spacing in frequency depends on the group velocity.

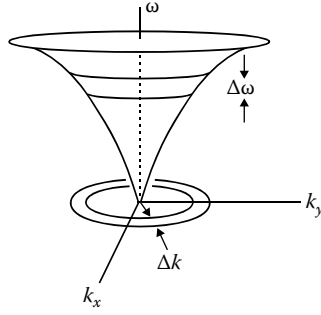


FIGURE 6.48 The density of states $g(\omega)$ for the 2-D crystal in the range $\Delta\omega$ is the same as the number of k -states in the annulus in the plane.

curves $\omega(k)$ since the shape of these curves determines the number of states in a given range of energy. Figure 6.47 also shows how those regions with the shallower slope (i.e., smaller group velocity) incorporate more states within the small range of frequencies.

The method of counting frequency states can be seen most clearly for a 2-D monatomic crystal (such as might be found on the surface of a table). In this case (Figure 6.48), the dispersion curve generates a surface for ω versus $\vec{k}$. The states within the range $\Delta\omega$ correspond to the equally spaced states within the annulus in the $k_x - k_y$ plane. The projection of the 3-D region denoted by $\Delta\omega$ produces the 2-D region in the plane with the difference in radii given by Δk . In short, calculating the number of points in the annulus in the $k_x - k_y$ plane therefore gives the number of states in the range $\Delta\omega$ required for calculating $g(\omega)$.

In the ensuing discussion, we define two types of state density for calculation (?) convenience; however, the density of states most often refers to $g(\omega)$. The density of states denoted by $g_{\vec{k}}$ refers to the states in the $\vec{k}$ plane. In what follows, we show first an example for the 2-D crystal owing to the ease of drawing figures. Afterward, we summarize this standard technique and then demonstrate the method for the 3-D crystal.

6.9.2 THE DENSITY OF STATES IN $\vec{k}$ -SPACE

Consider first (for ease of drawing the figures) the case of a 2-D arrangement of atoms forming a crystal in the x - y plane. Assume these atoms can vibrate along the z -direction with traveling waves instead of the standing ones. The traveling waves propagate along the x - and y -axes. Assume each side of the finite crystal has length L . The $\vec{k}$ -density of states determines the number of possible modes in a given region of $\vec{k}$ -space. Figure 6.49 shows a 2-D region of k -space for the vectors

$$\vec{k} = \frac{2\pi m}{L} \hat{x} + \frac{2\pi n}{L} \hat{y} \quad m, n = 0, \pm 1, \dots \quad (6.105)$$

where $\hat{x}, \hat{y}$ represent the unit vectors along the x - and y -directions, respectively. These allowed k -vectors come from the 2-D crystal with periodic boundary conditions on the length L as discussed in the previous section. The length L then relates to the number of atoms in the crystal $L = Na$ where a is the atomic spacing.

As previously discussed, the state density $g(\omega)$ can be found from the states in the $\vec{k}$ plane. First find the number of states per unit k -area. If we look at the horizontal direction for a moment then the k_x -distance between adjacent points must be given by

$$\frac{2\pi(m+1)}{L} - \frac{2\pi m}{L} = \frac{2\pi}{L} \quad (6.106a)$$

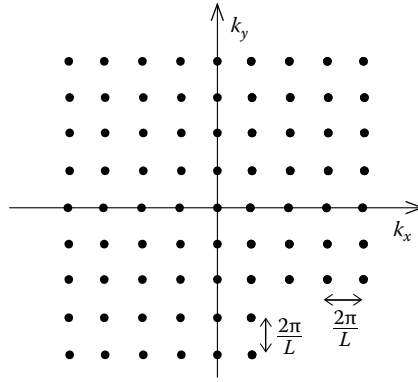


FIGURE 6.49 The states in the k -plane allowed by the periodic boundary conditions. Standing waves would produce k -values only in the positive quadrant but with four times the number of points shown here.

Therefore, each elemental area of k -space

$$\frac{2\pi}{L} \cdot \frac{2\pi}{L} = \left(\frac{2\pi}{L}\right)^2 \quad (6.106b)$$

has precisely one mode. The number of modes per unit area of $\vec{k}$ -space must then be given by

$$g_{\vec{k}}^{(2-D)} = \frac{1}{(2\pi/L)^2} = \frac{L^2}{4\pi^2} = \frac{A_{\text{xal}}}{4\pi^2} \quad (6.107)$$

where A_{xal} is the area of the crystal. The last equation can be normalized to the crystal area by dividing out the A_{xal} to find $g_{\vec{k}}^{(2-D)} = 1/(4\pi^2)$

A similar calculation provides the $\vec{k}$ -state density for a 3-D crystal. There is one mode in each elemental volume of k -space

$$g_{\vec{k}}^{(3-D)} = \frac{1}{(2\pi/L)^3} = \frac{L^3}{8\pi^3} = \frac{V_{\text{xal}}}{8\pi^3} \quad (6.108)$$

where V_{xal} is the total volume of the crystal (in direct space). A state density is often normalized to the crystal volume by dividing the last equation by V_{xal} . The $\vec{k}$ -state density then becomes $g_{\vec{k}}^{(3-D)} = 1/(8\pi^3)$.

Obviously, for one dimension, the k -density of states must be

$$g_{\vec{k}}^{(1-D)} = \frac{1}{(2\pi/L)} = \frac{L}{2\pi} \quad (6.109)$$

The previous equations show that the density of states for n -dimensions can be written as

$$g_{\vec{k}}^{(n-D)} = \frac{1}{(2\pi/L)^n} = \left(\frac{L}{2\pi}\right)^n \quad (6.110)$$

and can be normalized by dividing L^n if desired.

6.9.3 DENSITY OF STATES FOR 2-D CRYSTAL NEAR $k=0$ FOR THE ACOUSTIC BRANCH

The density of states $g(\omega)$ can be calculated for the 2-D monatomic crystal using the $\vec{k}$ -state density obtained in the previous section. We limit the range of $\vec{k}$ to small values so that the medium can be considered “nondispersive” in that the angular frequency can be related to the magnitude of the wave vector $k = |\vec{k}|$ through the speed v as

$$\omega = v|\vec{k}| = vk \quad (6.111)$$

The group velocity has the same value as the phase speed in this case. Restricting our attention to small k -values means the granularity of the k -values becomes more important. However, we will assume the minimum distance between k -states is small compared with our k -values of interest and thereby use an integral rather than a discrete summation for convenience. For large values of k , we would need to include the group velocity.

The total number of states within the area of a circle of radius k can be written as

$$\text{Total number} = \sum \frac{\text{Number}}{k\text{-area}} \Delta(k\text{-area}) = \sum g_{\vec{k}}^{(2-D)} dk |k| d\phi$$

We can rewrite this last expression as an integral using a dummy variable as

$$N_T = \int_0^k g_{\vec{k}}^{(2-D)} k' dk' d\phi = \int_0^k \frac{A_{\text{xal}}}{4\pi^2} k' dk' d\phi \quad (6.112a)$$

where $A_{\text{xal}} = L^2$. The last integral can also be written for the total number per unit crystal area as

$$N_A = \frac{N_T}{A_{\text{xal}}} = \int_0^k \frac{1}{4\pi^2} k' dk' d\phi$$

Integrating Equation 6.112a over the angle gives

$$N_T = \frac{A_{\text{xal}}}{2\pi} \int_0^k k' dk' \quad (6.112b)$$

The density of states per unit $|\vec{k}|$ (i.e., the magnitude) comes from the last equation by differentiating

$$g_k^{(2-D)} = \frac{\partial N_T}{\partial k} = \frac{A_{\text{xal}} k}{2\pi} \quad (6.113)$$

We can find the density of modes for ω -space by substituting $\omega = vk$ into Equation 6.112a

$$N_T = \frac{A_{\text{xal}}}{\pi v^2} \int_0^\omega \omega' d\omega' \quad (6.114)$$

As a side note, for N_T to be the same number using either $|\vec{k}|$ or ω , we require the limits on the two integrals (Equations 6.112b and 6.114) to be related through the applicable dispersion relation

which is $\omega = vk$ for this nondispersive case. Continuing with the integration in Equation 6.114, we find

$$N_T = \frac{A_{\text{sal}} \omega^2}{2\pi v^2}$$

Therefore, the number of states per unit angular frequency is given by

$$g_{\omega}^{(2-D)} = \frac{\partial N_T}{\partial \omega} = \frac{A_{\text{sal}} \omega}{\pi v^2} \quad (6.115)$$

where V_{sal} is the volume used to normalize a wave function for the periodic boundary conditions.

6.9.4 SUMMARY OF TECHNIQUE

In this section we briefly repeat the procedure used in the previous section to calculate the density of states $g(\omega)$. As before, we simplify the work by assuming that ω is isotropic in k -space so that

$$\omega = \omega(\vec{k}) = \omega(k_x, k_y, k_z) = \omega(k) \quad (6.116)$$

which indicates that ω and k are related in the same way regardless of the direction of propagation. This isotropy in k -space is important because, as shown in Figure 6.50, the condition of $\omega = \omega(k) = \text{constant}$ defines a circle which then requires us to integrate the area of a circle to find the total number of states. A subsequent section, for the case of electrons, will show the anisotropic case by using an ellipsoid rather than a spherical surface. In such a case, if $\omega = \omega(k) = \text{constant}$ defines another curve in the k -plane such as an ellipse, then we must be able to integrate the area to find the density of states. However, for the circle, we are able to easily integrate $g(\vec{k})$ over the angles to find $g(k)$. Carefully note the difference between $\vec{k}$ and $k = |\vec{k}|$ since k appears as a radius. We then have

$$N_T = \int_0^k dk' g(k') \quad (6.117)$$

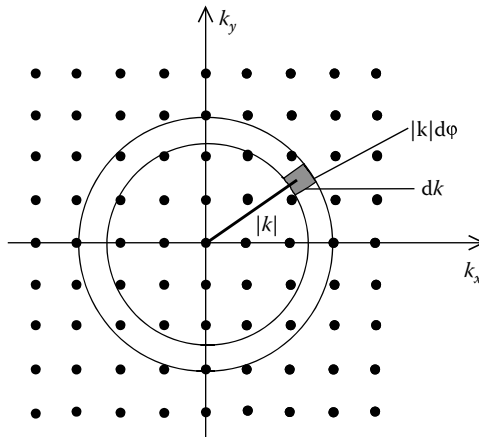


FIGURE 6.50 The frequency depends on the magnitude of the wave vector k , which also provides the approximate radius of the outer circle.

Now we can find $g(\omega)$ as follows

$$g(\omega) = \frac{dN_T}{d\omega} = \frac{dk}{d\omega} \frac{dN_T}{dk} = \frac{dk}{d\omega} \frac{d}{dk} \int_0^k dk' g(k') = g(k) \frac{dk}{d\omega} \quad (6.118)$$

We recognize the last derivate as being related to the group velocity.

$$\frac{dk}{d\omega} = \frac{1}{d\omega/dk} = \frac{1}{v_g} \quad (6.119)$$

Notice in this case that we have used v_g . In the previous section, we found

$$g(k) = \int d\varphi k g(\vec{k}) = \int d\varphi k \frac{A_{\text{xal}}}{(2\pi)^2} = \frac{A_{\text{xal}} k}{2\pi} \quad (6.120)$$

As a special important note, Equation 6.118 gives a special relation that can be used in a variety of circumstances

$$g(\omega)d\omega = g(k)dk \quad (6.121)$$

For example to find $g(\omega)$, solve for $g(\omega)$, substitute $d\omega/dk$, and remember to eliminate all k at the end in favor of ω .

The above analysis applies to the case of wave motion in a 2-D crystal when the coupling constants (i.e., spring constants) $\beta_x = \beta_y$ which results in the isotropic form of ω . Apparently, the isotropy of ω must be linked to the isotropy of the crystal. Also notice that we did not worry about the size of the crystal along x and y . It would only show up in the $g(k)$ as a slightly different crystal area $A_{\text{xal}} = L_x L_y$.

6.9.5 3-D CRYSTAL IN LONG-WAVELENGTH LIMIT

Consider a 3-D isotropic crystal with only one polarization for each propagation direction. We know that the density of states in k -space can be written as

$$g_k^{(3-D)} = \frac{V_{\text{xal}}}{8\pi^3} \quad (6.122)$$

The total number of states enclosed by a sphere of radius of $k = |\vec{k}|$ can be written as

$$N_T = \int_0^k dk' \int_0^{2\pi} k' d\phi \int d\theta k' \sin \theta g_{\vec{k}'}^{(3-D)}$$

where the integral has the usual spherical coordinates and a differential volume element of

$$(dk)(k d\phi)(d\theta k \sin \theta)$$

The two angular integrals can be evaluated since the density of states does not depend on the angles. We find

$$N_T = 4\pi \int_0^k dk k' 2g_k^{(3-D)} = \frac{V_{\text{xal}}}{2\pi^2} \int_0^k dk k'^2$$

The integral provides the total number of states enclosed by a sphere of radius $k = |\vec{k}|$

$$N_T = \frac{V_{\text{xal}} k^3}{6\pi^2} \quad (6.123)$$

Notice that this could have been immediately deduced without working through the integral for the isotropic crystal just by multiplying the k -density of states by the volume of sphere in k -space. Now we can find the frequency density by the following calculation. Remember to remove the crystal volume!

$$g(\omega) = \frac{1}{V_{\text{xal}}} \frac{dN_T}{d\omega} = \frac{1}{V_{\text{xal}}} \frac{dk}{d\omega} \frac{dN_T}{dk} = \frac{1}{v_g} \frac{k^2}{2\pi^2} \quad (6.124)$$

where $v_g = v$ (phase speed) for the present case.

The density of states for ω -space is found from Equation 6.124 by substituting $\omega = vk$ for the wave vector to get

$$g(\omega) = \frac{\omega^2}{2\pi^2 v^3} \quad (6.125)$$

The number of modes and the density of modes increases if more than one polarization is included. For phonons, there might be six modes and so the density of modes increases by a factor of 6. For light traveling in a medium, the constant c the speed of light in the medium replaces v . For photons in the coulomb gauge, there are two transverse modes and so the density of modes must double.

6.10 COMMENTS ON PHONON CRYSTAL MOMENTUM

The phonon momentum and energy have many important roles in semiconductor phenomena. An electron (atom) interacting (emitting or absorbing) with a phonon can make “nonvertical” transitions between the conduction and valence bands. The phonon affects the conductivity and mobility through the scattering processes. In this section, we discuss the phonon momentum and its relation to crystal momentum.

6.10.1 ANTICIPATIONS FOR MOMENTUM

Recall the acoustic dispersion curve for the phonon. Figure 6.51 shows an example of the extended band structure for the LA phonons in a 1-D monatomic crystal. Recall that the first Brillouin zone (FBZ) extends from $-\pi/a$ to π/a . For the 1-D simple cubic (SC) lattice, the distance between these two points must equal the smallest reciprocal lattice vector $G_1 = 2\pi/a$. Usually we restrict our attention to the FBZ and do not consider wave vectors outside of this region.

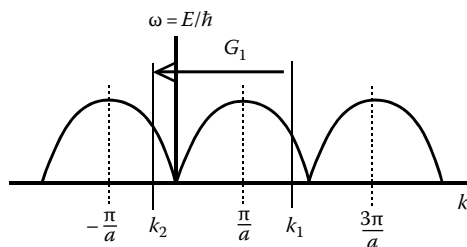


FIGURE 6.51 The extended band diagram for LA wave on a 1-D monatomic crystal.

Based on experience with momentum in free space, one might expect the phonon to have $\vec{p} = \hbar\vec{k}$. Further one might expect the conservation of momentum to hold. For example, if a neutron (or electron or ...) collides with a crystal atom and imparts momentum, the conservation of momentum should hold $\vec{p}_{\text{neutron}}^{\text{initial}} = \vec{p}_{\text{neutron}}^{\text{final}} + \hbar\vec{k}_1$. Likewise, if two phonons with wave vectors $\vec{k}$ and $\vec{K}$ combine (through crystal nonlinearities) then one expects to find a third phonon with wave vector $\vec{k}_1 = \vec{k} + \vec{K}$. In either case, energy must be conserved. The conservation of momentum agrees with intuition when the three phonons have wave vectors within the FBZ.

6.10.2 CONSERVATION OF MOMENTUM IN CRYSTALS

What if the conservation of momentum requires the final phonon momentum k_1 to be outside of the FBZ (Figure 6.51)? The question really addresses at least two issues. First the physical relevance of a wave vector outside the FBZ and second the role of the reciprocal lattice vectors for conservation of momentum. An unusual aspect of the problem concerns the fact that k_1 shown in Figure 6.51 has negative group velocity which indicates the phonon (perhaps produced by an impacting neutron from outside the crystal) moves in a direction opposite to that required by ordinary conservation of momentum. That is, based on wave vectors, momentum does not appear to be conserved; however, overall, the crystal as a block of total mass M along with an incident particle do conserve momentum. In fact, conservation of momentum holds for systems with infinitesimal translation symmetry whereas the crystal and lattice only have translational symmetry through a lattice vector.

Wave vectors beyond the FBZ have not any physical significance. As shown in previous sections, the sinusoidal wave corresponds to actual physical atoms only at specific locations. The wave does not have any physical significance for positions between the atoms. Further, the number of allowed wave vectors must be the same as the number of degrees of freedom. Wave vectors in the FBZ account for all of the degrees of freedom. The functions of physical significance for phonons (1-D for example) can be written as a Fourier summation over the traveling waves as

$$u(x_m) = \sum_n C_n e^{ik_n x_m} \quad (6.126)$$

where x_m refers to a lattice site and therefore must have the form of a direct lattice vector. Then if G is a reciprocal lattice vector then $\exp(Gx_m) = 1$ so the function u must be invariant with respect to changes in the wave vector by a reciprocal lattice vector.

$$\sum_n C_n e^{i(k_n + G)x_m} = \sum_n C_n e^{ik_n x_m} = u(x_m) \quad (6.127)$$

For the momentum, the usual procedure is to add or subtract a reciprocal lattice vector as shown in the figure such that $k_2 = k_1 + G_1$ can be found in the FBZ. Notice that the phonon still has negative group velocity. The final momentum of the phonon becomes

$$p_{\text{phonon}} = \hbar(k \pm G) \quad (6.128)$$

where the reciprocal lattice vector G is chosen so that the phonon wave vector lies in the FBZ. For the neutron collision, the momentum conservation would read

$$\hbar\vec{k}_{\text{neutron}}^{\text{initial}} = \hbar\vec{k}_{\text{neutron}}^{\text{final}} + \hbar(\vec{k}_{\text{phonon}} \pm G) \quad (6.129)$$

The difference in neutron momentum must be exhibited by the crystal as a whole in order to rigorously conserve momentum. The conservation of energy uses the wave vector in the FBZ.

For example, for a neutron colliding with an atom in a massive crystal (for which the change in kinetic energy can be considered negligible), one would write

$$E_{\text{neutron}}^{\text{initial}} = E_{\text{neutron}}^{\text{final}} \pm \hbar\omega_k$$

depending on whether the neutron produces or absorbs a phonon. Of course the value of ω does not depend on which Brillouin zone is being considered because of the periodic nature of the dispersion curves.

The Umklapp phonon process (or u -process) occurs when the resultant wave vector for a phonon occurs in the second (or larger) Brillouin zone as shown by Figure 6.51 for k_1 . As mentioned, the physically relevant wave vector has the value k_2 (negative) and represents a wave moving in a direction opposite to the initial wave. In particular, if two phonons (with positive values of k) interact through any nonlinearities of the crystal and produce a resultant phonon with wave vector k_1 (Figure 6.51), then the resultant phonon actually has wave vector k_2 and moves in the opposite direction. This Umklapp process occurs only for periodic structures and produces thermal resistance within the material. Those processes that do not exhibit wave vectors outside of the FBZ are the normal processes (or n -processes).

6.11 THE PHONON BOSE-EINSTEIN PROBABILITY DISTRIBUTION

The temperature of a material determines the number of phonons occupying each phonon mode. The occupancy has important implications for physical properties including specific heat, thermal conductivity, and electron mobility. The present section determines the Bose-Einstein probability distribution for phonons based on concepts of statistical mechanics for thermal equilibrium and the determination of temperature through the entropy. Once having found the probability distribution, the section discusses the statistical moments.

6.11.1 DISCUSSION OF RESERVOIRS AND EQUILIBRIUM

A material system can be maintained at a given temperature T by bringing it into thermal contact with a thermal reservoir similar to Figure 6.52. The reservoir and system interchange energy (heat) to bring the system to the same temperature as the reservoir. The temperature of the reservoir undergoes negligible change as a result of its very large number of degrees of freedom (compared with the system). The reservoir and system continuously interchange energy even after reaching thermal equilibrium in the form of fluctuations. However, this “to and fro” flow averages to an equilibrium value to maintain the system temperature at T .

The temperature of the system measures the energy in the system. Notice that this measure of energy must be related to the average energy per molecule or atom comprising the system. For if

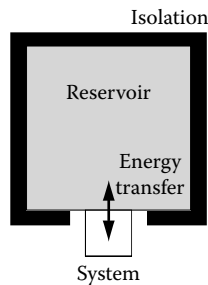


FIGURE 6.52 The thermal reservoir in thermal contact with a small piece of matter.

temperature was to refer to the total energy then doubling the size of the small system would double the temperature, which does not happen. The notion of temperature has general application to all substances, but perhaps examining a gas reservoir at temperature T with a small mercury thermometer as a monitor provides a good visual example. Hot gas molecules colliding with a cool thermometer, for example, transfer kinetic energy from the gas to the mercury in the thermometer. The transferred energy (1) increases the atomic motion for the mercury, (2) increases the separation of the mercury atoms through collisions and nonlinearities, and (3) thereby increases the height of the mercury column in the thermometer to indicate larger temperatures. The translational energy directly indicates the temperature according to $E_{\text{trans}} = kT/2$ per degree of freedom where k and T represent Boltzmann's constant and the temperature in Kelvin, respectively. The translational energy for an atom free to move in 3-D is $E_{\text{trans}} = 3kT/2$ and for N -atoms free to move in 3-D is $E_{\text{trans}} = \frac{3N}{2}kT$. The gas with N -atoms has a Hamiltonian with $3N$ terms for kinetic energy unlike the harmonic oscillator which has both the kinetic and potential energy terms. Each term in the Hamiltonian receives $kT/2$. If one could convert another degree of freedom to a visual indicator, then that other degree of freedom could be used to measure temperature. An indicator of temperature only needs to provide a measure of the energy per degree-of-freedom (DOF). For example, if the average classical rotation of the molecules (of a gas for example) could be measured then that average could be used as an indicator of temperature.

For normal systems, the quantity kT (and hence the temperature T) roughly represents the mean energy per degree-of-freedom (DOF) above the ground state. For example, consider atoms of type A that can vibrate along one single direction (1 degree of freedom for each atom) and atoms of type B that can vibrate along two directions (a total of two degrees of freedom for each atom). Following classical notions, these two types of atoms will be in thermal equilibrium with each other at temperature T provided each degree of freedom for each atom has roughly the energy kT . That is, type A atoms would have energy kT and type B atoms would have energy $2kT$. One might view the process of approaching equilibrium as the diffusion of energy from hot objects to cool ones until every degree of freedom has the same "temperature" (i.e., each "nook and cranny" is filled with energy to the same level). The measure of temperature provides a measure of the level of the energy as all "nooks and crannies" fill. A difference in temperature corresponds to two regions with differing amounts of energy per DOF which then sets the stage for energy to diffuse from the high-energy-density region to the low-density region.

The idea of temperature has a relation with specific heat in that specific heat describes the total internal energy of an object at a given temperature by adding together all of the energy for all of the degrees of freedom. For example, if gas molecules have three translational degrees of freedom and three rotational degrees of freedom then the specific heat includes the approximately $6NkT$ of energy (ignoring the factors of $1/2$). We should include the factor of $1/2$ for translations but harmonic oscillators do not have the factor of $1/2$.

Some systems do not have translational degrees of freedom but can still attain thermal equilibrium at temperature T . For example, consider a system composed of (1) electrons that do not have translational motion but can freely change spin between the up and down state (i.e., z -component of spin) and (2) a magnetic field so that the two states of electron spin correspond to different energy $E_1 < E_2$ (as discussed in Chapter 5). In such a case, one expects the ratio of the number of electrons with spin in the higher energy state and lower energy state to be proportional to the temperature. However, the temperature becomes negative when more spins occupy the upper energy level than the lower one (a population inversion) because the entropy decreases when all electrons have spin in the same energy state (see Chapter 8) (somewhat similar concepts apply to lasers that must achieve a population inversion in order to lase). If the ratio of the number of spins in two states has the form $N_2/N_1 \sim e^{\Delta E_{\text{spin}}/kT}$ (the Boltzmann factor, see below and Chapter 8) where $\Delta E_{\text{spin}} = E_2 - E_1$ then for very large temperature T , one at most achieves $N_2 = N_1$. Clearly the temperature must be negative to achieve the population inversion whereby $N_1 < N_2$. In semiconductor lasers, the electron population (without considering spin) can be inverted by charge injection methods such as attaching a battery.

For the present section, the number of phonons in a state occupies our attention. The phonons deal with the vibration of atoms and molecules in a material about an equilibrium position. In such a case, the temperature must be related to the vibrational energy of the molecules. In the case of phonons, the allowed energy states $E_k = \hbar\omega_k$ associated with the dispersion curves are related to the degrees of freedom—there is one such state for each degree of freedom. The frequencies ω_k correspond to the normal modes of vibration that appear similar to sinusoidal waves across the material produced by the collective motion of the atoms. In the case of traveling waves, ω_k and ω_{-k} represent distinct states but with identical energy. For fixed endpoint conditions, only positive k should be considered for the standing waves. The amplitude of the wave at frequency ω_k increases as $\sqrt{\text{#phonons in state } \omega_k}$. We might imagine the states on the dispersion curve as “buckets” that hold phonons. So now, thermal equilibrium means that each bucket should be filled with phonons until the contained energy reaches a “level” of approximately kT since each bucket represents a degree of freedom. One might expect the states with larger ω_k have fewer phonons because the phonons have larger energy and therefore, the state requires fewer of them to reach the energy kT .

6.11.2 EQUILIBRIUM REQUIRES EQUAL TEMPERATURES

One can show that two systems achieve thermal equilibrium with a common temperature T when the entropy (i.e., disorder) of the two systems attains a maximum. Consider a system S with N_S degrees of freedom. We previously discussed that the energy E_S of the system should roughly divide equally among all of the degrees of freedom at equilibrium (i.e., equipartition of energy, and ignoring factors of $1/2$, although the factors are necessary). The system moves away from equilibrium when the energy moves from equal division among DOF. Generally, not all degrees of freedom have equal energy due to normal thermal fluctuations at equilibrium or while the system approaches equilibrium at initial contact. The various arrangements of the microscopic particles or energy quanta among the various degrees of freedom (not necessarily equally) represents the microstates of the system.

Consider a system S in thermal contact with a reservoir R . At any instant of time, the system S will be in a particular microstate s . A particular microstate s refers to the particular arrangement of energy in the possible degrees of freedom. For example, if an atom has two translational degrees of freedom then the coordinates required to specify the state would include both position and momentum as (x, p_x) and (y, p_y) . For the translational microstate in this case, we are interested in the kinetic energy T content of the x and y degrees of freedom (T_x, T_y) . So the total energy E_S has been apportioned in a specific manner among the various DOF. However, there will generally be other arrangements that have the same energy E_S . For example (T_x, T_y) of $(1, 2)$ and $(2, 1)$ are different microstates but have the same energy E_S . Conceptually, the quantum systems make for easier computation since the microstates arise from discrete energy levels. For example, consider two quantum wells then (E_1, E_2) describe a particular microstate with an electron in state E_1 for well #1 and one in state E_2 for well #2.

Let $\Omega(E_S)$ be the number of microstates with the energy E_S . Similarly, because the combined system of $S + R$ satisfies the conservation of energy $E = E_S + E_R$ (assuming the combined system is isolated from external influences), the remaining amount of reservoir energy $E_R = E - E_S$ must be divided among the reservoir microstates that have total energy E_R . Let $\Omega_R(E_R)$ represent the total number of microstates available to the reservoir when it has energy E_R . Notice that the energy E_S determines the number of states accessible to both S and R because of energy conservation. The total number of states accessible to the combined system $\Omega(E_S)$ must be the product of the number of states accessible to the phonon system $\Omega_S(E_S)$ and the reservoir $\Omega_R(E_R)$

$$\Omega(E_S) = \Omega_S(E_S)\Omega_R(E_R) \quad (6.130)$$

Notice that the functional dependence $\Omega = \Omega(E_S)$ indicates that the entropy for the combined system depends on the energy of the small-phonon system. When the system S corresponds to a system of phonons, then the phonons arrange themselves in the various states $\omega_B(\vec{k})$ where “B” indexes the various branches.

We show that equilibrium occurs when the entropy $S = kLn(\Omega)$ of the combined system attains a maximum value (see Chapter 8 for a full discussion of entropy). Here k represents the Boltzmann constant. We wish to adjust the energy of the small system so as to achieve maximum disorder for the combined system as a requirement for thermal equilibrium. Taking the natural logarithm of Equation 6.130, and then differentiating and setting to zero provides

$$0 = \frac{d}{dE_S} [kLn(\Omega_S)] - \frac{d}{dE_R} [kLn(\Omega_R)] \quad (6.131a)$$

where, in view of energy conservation $E = E_S + E_R$, the energy of the combined system does not change ($dE = 0$) so that we were able to change the second differential dE_S to $-dE_R$. Using the definition of temperature as $T^{-1} = dS/dE$ (see Chapter 8), Equation 6.131a then shows that the maximum entropy for the combined system leads to identical temperatures for the small system and the thermal reservoir

$$T_S = T_R \quad (6.131b)$$

6.11.3 DISCUSSION OF BOLTZMANN FACTOR

The phonon system has energy levels E_s in thermal equilibrium at temperature T and an average of n_s phonons in each state. Here each “s” represents precisely one state even when there are two or more states with the same energy (such as for ω_k and ω_{-k}). The Boltzmann factor gives the probability of finding an oscillator in its n th state or equivalently the probability of a state $\omega(\vec{k})$ having n phonons

$$P(E_n) = Ce^{-E_n/kT} \quad P(n) = Ce^{-n\hbar\omega_k/kT} \quad (6.132)$$

where the $1/2$ in the harmonic oscillator energy $E_n = \hbar\omega_k(n + 1/2)$ has been dropped. A simple derivation of the Boltzmann factor will be considered next (and in Chapter 8). Also see for instance the books by Pathria or Reif.

If a small system S occupies a particular state s with energy E_S then the reservoir has energy $E - E_S$ when the combined system of $S + R$ has energy E . Suppose the system S has the number $\Omega_S(E_S)$ of such microstates with energy E_S . As a result of the system S having energy E_S , the reservoir will distribute the energy $E_R = E - E_S$ among its microstates. The number of microstates for the reservoir will be $\Omega_R(E_R)$. We assume that the small phonon system S has far fewer degrees of freedom than the reservoir R so that $E_S \ll E_R$ and $E_S \ll E$. The probability of finding the small system in a particular microstate s with energy E_S is calculated from the number of possible configurations of the combined system such that the small system has energy E_S . This calculation considers the number of possible arrangements of the internal constituents of the reservoir among all of its internal degrees of freedom so that its energy is $E - E_S$. Therefore the probability of finding the phonon system in microstate s with energy E_S must be proportional to the number of possible configurations $\Omega_R(E_R)$ of the reservoir

$$P(E_S) \sim \Omega_R(E_R) = \Omega_R(E - E_S) \quad (6.133)$$

Consider the natural logarithm and make a Taylor expansion in E_S , which is small compared with both E_R and E due to system S having a relatively small number of degrees of freedom.

$$\ln\{\Omega_R(E - E_S)\} \cong \ln\{\Omega_R(E)\} + (-E_S) \frac{\partial}{\partial E_R} \ln\{\Omega_R(E_R)\}_{E_R=E} \quad (6.134)$$

Use the definition of temperature T in terms of entropy $T^{-1} = dS/dE$ where $S = k \ln(\Omega)$ and k represents the Boltzmann constant. The temperature of the system matches that of the reservoir $T_R^{-1} = dS_R/dE_R = k \frac{\partial}{\partial E_R} \ln\{\Omega_R(E_R)\}$ and since $E_S \ll E_R$, we can evaluate the entropy with $E_R \sim E$. The result becomes

$$\ln\{\Omega_R(E - E_S)\} = \ln\{\Omega_R(E)\} - E_S/kT \quad \Rightarrow \quad \Omega_R(E - E_S) = \Omega_R(E)e^{-E_S/kT} \quad (6.135)$$

Equation 6.133 then shows that the probability of a harmonic oscillator occupying state E_S must be

$$P(E_S) = Ce^{-E_S/kT} \quad (6.136)$$

where C represents a constant to normalize the probability. If the system is a single-harmonic oscillator, then $E_S = E_n$. We must yet find the constant C which yields the Bose–Einstein statistics.

6.11.4 BOSE–EINSTEIN PROBABILITY DISTRIBUTION FOR PHONONS

The Bose–Einstein statistics can be found using a number of methods such as through the Lagrange multipliers (c.f., Chapter 8, Appendices H and J), but we follow a procedure often used for phonons. The states $\omega(\vec{k})$ refer to the collective motion of the atoms in the material. Previous sections such as Sections 4.5 and 6.6 show that N coupled oscillators produce N distinct states in a phonon band labeled by $\omega(\vec{k})$. These modes represent the collective motion of the atoms (i.e., normal modes). Each normal mode represents a single-harmonic oscillator as opposed to focusing on a single atom oscillating about an equilibrium point. Adding a phonon to the state $\omega(\vec{k})$ increases the amplitude of the collective motion.

The fraction of harmonic oscillators (i.e., normal modes) $P(n)$ with n quanta must be given by

$$P(n) = N_n / \sum_n N_n \quad (6.137)$$

where

N_n represents the total number of modes with n quanta

$N = \sum_n N_n$ represents the total number of modes (equal to the number of atoms for each branch)

The fraction $P(n)$ can be interpreted as the probability of finding an oscillator to have n quanta. However, the various modes have different energies which can affect the probability. A better way to view the situation consists of defining an ensemble of identically constructed systems and look at identical modes $\omega(\vec{k})$ in each system of the ensemble. Each system at temperature T will have different numbers of phonons in the state $\omega(\vec{k})$ due to typical thermal fluctuations. Then N_n will be the number of systems that have n phonons in the mode.

Assume each system in the ensemble is at equilibrium temperature T . Consider the harmonic oscillators having the specific frequency ω_k . The oscillators have energy levels

$$E_n = \hbar\omega_k \left(n + \frac{1}{2} \right) \quad (6.138)$$

Assume the expected number N_{n+1} of oscillators in state $n+1$ compared with those in state n are related by the Boltzmann factor

$$\frac{N_{n+1}}{N_n} = e^{-\frac{\hbar\omega_k}{k_b T}} \quad (6.139)$$

where

k_b is the Boltzmann constant

$\hbar\omega_k$ gives the energy separating adjacent levels

This last equation really describes the ratio of amplitudes of the waves with frequency ω_k . Using the ratio

$$\frac{N_n}{N_0} = \frac{N_n}{N_{n-1}} \frac{N_{n-1}}{N_{n-2}} \dots \frac{N_2}{N_1} \frac{N_1}{N_0} = \underbrace{e^{\frac{\hbar\omega_k}{k_b T}} e^{\frac{\hbar\omega_k}{k_b T}} \dots e^{\frac{\hbar\omega_k}{k_b T}} e^{\frac{\hbar\omega_k}{k_b T}}}_{n\text{-terms}} = e^{-\frac{n\hbar\omega_k}{k_b T}} \quad (6.140)$$

The probability of a mode with frequency ω_k having n phonons can be found from Equation 6.140 as

$$p(n) = \frac{N_n}{\sum_m N_m} = \frac{N_0 e^{-\frac{n\hbar\omega_k}{k_b T}}}{\sum_m N_0 e^{-\frac{m\hbar\omega_k}{k_b T}}} = \frac{e^{-\frac{n\hbar\omega_k}{k_b T}}}{\sum_m e^{-\frac{m\hbar\omega_k}{k_b T}}} \quad (6.141)$$

We can calculate the denominator in Equation 6.141. Assume very large N , the denominator has the form

$$y = \sum_{m=0}^N x^m = 1 + x(y - x^N) \quad (6.142a)$$

where $x = e^{-\frac{\hbar\omega_k}{k_b T}}$. We find

$$y = \frac{1 - x^{N+1}}{1 - x} \xrightarrow{N \rightarrow \infty} y = \frac{1}{1 - x} \quad (6.142b)$$

The probability of finding an ω_k oscillator having n photons in Equation 6.141 becomes

$$p(n) = \left(1 - e^{-\frac{\hbar\omega_k}{k_b T}}\right) e^{-\frac{n\hbar\omega_k}{k_b T}} \quad (6.143)$$

The thermal reservoir is responsible for this probability distribution.

6.11.5 STATISTICAL MOMENTS FOR PHONON BOSE-EINSTEIN DISTRIBUTION

The number of phonons in each mode fluctuates over a period of time even when the system has achieved equilibrium. However, the fluctuations always occur about an average value consistent with the conditions for thermal equilibrium. Holding the system parameters constant such as temperature (as a macroscopic average quantity) will produce an ergodic process.

The m th statistical moment of n has the definition

$$M_m(n) = \langle n^m \rangle = \sum_n n^m p(n) \quad (6.144)$$

Often one uses the alternative definition of the m th statistical moment (about the mean)

$$M'_m(n) = \langle (n - \bar{n})^m \rangle = \sum_n (n - \bar{n})^m p(n) \quad (6.145)$$

where the mean $\bar{n}$ has the definition

$$\bar{n} = \sum_n np(n) \quad (6.146)$$

The second statistical moment about the mean M'_2 is the variance and M'_3 and M'_4 are the skew and kurtosis, respectively.

The moments M_m (the moments in Equation 6.144) can most conveniently be determined by using a moment generating function G_m defined by

$$G_m(n) = \frac{\partial^m}{\partial t^m} \langle e^{nt} \rangle \Big|_{t=0} \quad (6.147)$$

Note that t represents a parameter and not related to time. For example, G_1 represents the average

$$G_1(n) = \frac{\partial}{\partial t} \langle e^{nt} \rangle \Big|_{t=0} = \left\langle \frac{\partial}{\partial t} e^{nt} \right\rangle_{t=0} = \langle n e^{nt} \rangle_{t=0} = \langle n \rangle$$

In order to use the moment generating functions G_m , one must calculate $\langle e^{nt} \rangle$ as follows

$$\langle e^{nt} \rangle = \sum_{n=0}^{\infty} e^{nt} p(n) = \left(1 - e^{-\frac{\hbar\omega_k}{kT}} \right) \sum_n e^{nt} e^{-\frac{n\hbar\omega_k}{kT}}$$

Defining $y = \sum_{n=0}^{\infty} x^n$ and $x = e^{t - \frac{\hbar\omega_k}{kT}}$, we can calculate the summation from

$$y = 1 + x + x^2 + \cdots = 1 + x(1 + x + x^2 + \cdots) = 1 + xy \quad \rightarrow \quad y = \frac{1}{1-x}$$

Convergence is not an issue since the t parameter can (and will be) made arbitrarily small. The average for the generating function becomes

$$\langle e^{nt} \rangle = \frac{1 - e^{-\frac{\hbar\omega_k}{kT}}}{1 - e^t e^{-\frac{\hbar\omega_k}{kT}}} \quad (6.148)$$

Now the moments can be calculated by differentiation.

The average number of quanta (phonons) per oscillator in the mode ω_k can be calculated using Equation 6.148.

$$\langle n \rangle = \frac{\partial}{\partial t} \langle e^{nt} \rangle \Big|_{t=0} = \frac{\partial}{\partial t} \frac{1 - e^{-\frac{\hbar\omega_k}{kT}}}{1 - e^t e^{-\frac{\hbar\omega_k}{kT}}} \Big|_{t=0} = \frac{1}{e^{\frac{\hbar\omega_k}{kT}} - 1} \quad (6.149)$$

Notice for phonon energy $E_k = \hbar\omega_k$ small compared with kT (by Taylor expanding the exponential and dropping second order and larger terms), the average number of phonons in the mode has the value kT/E_k . The mode contains the number of phonons necessary to give the oscillator (i.e., the degree of freedom) an energy of kT .

The variance σ_n^2 can be calculated in a similar manner using the moment generating function except it requires two moments to be included

$$\sigma_n^2 = \langle (n - \bar{n})^2 \rangle = \langle n^2 \rangle - \bar{n}^2 \quad (6.150)$$

where $\bar{n}$ represents the average value of n as $\bar{n} = \langle n \rangle = \sum_n np(n)$, which has already been determined in Equation 6.149. The average of the square can be determined as

$$\langle n^2 \rangle = \frac{\partial^2}{\partial t^2} \langle e^{nt} \rangle \Big|_{t=0} = \frac{e^{-\frac{\hbar\omega_k}{kT}}}{1 - e^{-\frac{\hbar\omega_k}{kT}}} + \frac{2e^{-\frac{2\hbar\omega_k}{kT}}}{\left(1 - e^{-\frac{\hbar\omega_k}{kT}}\right)^2} = \langle n \rangle + 2\langle n \rangle^2 \quad (6.151)$$

Therefore the variance becomes

$$\sigma_n^2 = \bar{n} + \bar{n}^2 \quad (6.152)$$

The variance describes the expected variation of the phonon number as the phonon system attempts to maintain equilibrium. Sometimes people refer to the probability as being “super-Poisson” since the variance is larger than the average.

In conclusion, Equation 6.149 shows that the number (average) of phonons in a given mode is controlled by the temperature of the system.

Example 6.9: Average Energy of an Oscillator

The average energy for the oscillators at ω_k can be calculated. Including the zero-point motion provides

$$\langle E_n \rangle = \left\langle \hbar\omega_k \left(n_k + \frac{1}{2} \right) \right\rangle = \hbar\omega_k \left(\langle n_k \rangle + \frac{1}{2} \right) = \frac{1}{2} \hbar\omega_k + \frac{\hbar\omega_k}{e^{\frac{\hbar\omega_k}{kT}} - 1} \quad (6.153)$$

6.12 INTRODUCTION TO SPECIFIC HEAT

An important application for the phonons is the determination of specific heat of a material. The section discusses the connection between temperature and the energy stored in degrees of freedom starting first with a gas and then for a solid. The phonon has important applications for conduction in electronic devices as well as for optical transitions.

6.12.1 DISCUSSION OF SPECIFIC HEAT

The specific heat c_v (energy per mass per degree Kelvin) or C_v (energy per degree Kelvin) of the body indicates the total energy stored (i.e., internal energy) within the material that can be exchanged with other objects during thermal interactions. The specific heat implicitly describes the various ways that an object can store energy. The temperature of the matter indirectly indicates the amount of stored energy. The specific heat can be defined (constant volume) as

$$dU = C_v dT \quad \text{or} \quad dU = mc_v dT \quad (6.154)$$

where

U represent the internal energy of the matter

T represents the temperature in Kelvin

For c_v , one divides the mass out of C_v so that the specific heat (i.e., stored energy) does not depend on the size of the object and it can be used as a material property. The specific heat is sometimes measured using constant pressure (C_p or c_p) that includes the energy required to expand the object (see review exercises for Chapter 6).

The *ideal* gas with N particles has only translational motion (in 3-D) so that

$$E_T = (3)(N)(kT/2) = 3Nk_bT/2 \quad (6.155)$$

represents the internal energy (translations) and therefore one expects the classical value of $C_v = \partial E / \partial T = 3Nk_b/2$ where N represents the total number of gas molecules. The Boltzmann constant k_b has the subscript “b” in this section to distinguish it from the wave vector k . In this case, one assumes that the equipartition theorem applies which requires each degree of freedom receive an equal share (i.e., $k_bT/2$) of the total energy. However, the law of equipartition of energy does not accurately describe quantum systems especially in those situations where the difference of energy between states exceeds k_bT .

Actual gasses have alternate energy storage mechanisms (degrees of freedom) than solely translations such as molecular rotation and the vibration between bonded atoms in molecules. Assuming that the total internal energy of a substance divides (equipartition theorem) among all the degrees of freedom (for all the gas molecules), the internal energy can be much larger than just the translational kinetic energy. Defining E_T , E_A to be the energy for the “translations” and the “alternate mechanisms,” respectively, the total energy would be $U = E_T(T) + E_A(T)$ or factoring out the translational part provides

$$U = [1 + E_A/E_T]E_T = [1 + E_A/E_T]3Nk_bT/2 \sim C_vT$$

which assumes E_T/E_A is approximately constant.

The translational degree of freedom was given special status since, for a gas, we are familiar with how the temperature directly controls the translational energy. The specific heat C_v can be seen to represent the ratio in this case, of the energy stored in “alternate mechanisms” to the energy in translations. However, the stored energy must be related to the number of different mechanisms for storing energy. Therefore, specific heat essentially measures the number of ways to store energy in the material. For a solid, the specific heat originates in several mechanisms. The molecules can vibrate with respect to their equilibrium position and they have some limited rotational energy (more like angular oscillations) since they bind to each other. Interestingly, still other storage mechanism might be available including, for example, “free” charge within the system.

Consider the monatomic phonon system with one polarization. Here the various phononic states represent the degrees of freedom. The number of atoms matches the number of states in the first Brillouin zone (FBZ). One might imagine that each state (degree of freedom) “fills up” with phonons, which have energy $E_k = \hbar\omega_k$, until the total phonon energy in the state matches the expected thermal energy k_bT . That is, if n denotes the average number of phonons in the state then $n = k_bT/E_k$ (see the discussion in Section 6.11). For phonons, the total energy would appear to be calculated as the product of the total number degrees of freedom (equal to N -atoms) times k_bT (which does not appear to involve the number of phonons n). In this case, one expects to find the specific heat of Nk_b (or $3Nk_b$ for three polarizations and N -atoms). However, this calculation for the number of phonons only has some validity for phonon energy $E_k = \hbar\omega_k$ small compared with k_bT as first shown by Equation 6.149 in the limit of small energy using

$$\langle n \rangle = \frac{1}{e^{\frac{\hbar\omega_k}{k_bT}} - 1} \quad (6.156a)$$

$$\langle n \rangle \sim \frac{k_bT}{\hbar\omega_k} \quad k_bT \gg \hbar\omega_k \quad (6.156b)$$

Instead we need to account for the distribution of phonons for any energy state and temperature T as will be accomplished in the next section. That is, one needs to incorporate the full expression for the number of phonons embodied by Equation 6.156a.

6.12.2 EINSTEIN MODEL FOR SPECIFIC HEAT

In the beginning, Dulong–Petit developed a model for the high temperature limit of the specific heat. In this case, they assumed atoms independently vibrate as classical harmonic oscillators about the equilibrium position. Averaging with a Boltzmann distribution, including N total atoms each having 3 degrees of freedom, produces an average total energy of

$$U = 3Nk_bT \quad (6.157a)$$

and therefore a specific heat of

$$C_v = \partial U / \partial T = 3Nk_b \quad (6.157b)$$

where as usual, k_b represents the Boltzmann constant. While the Dulong–Petit model predicts the high temperature limit for the specific heat C_v , it does not explain the decrease of C_v to zero as the temperature decreases for solids (see the references such as the books by Kittel or Blakemore). The problem concerns the equipartition theorem and the use of Boltzmann statistics rather than the Bose–Einstein distribution.

Just after M. Planck developed the quantization associated with light emission, Einstein explored models for the quantization associated with phonons and specific heat. He assumed the atoms oscillated with the same frequency around equilibrium but that they do not interact. Einstein used the correct distribution and applied it to N -atoms with three degrees of freedom each oscillating with angular frequency ω_k about equilibrium.

The total internal energy for the material must then be given by

$$U = 3N \langle n \rangle \hbar \omega_k = \frac{3N \hbar \omega_k}{e^{\frac{\hbar \omega_k}{k_b T}} - 1} \quad (6.158a)$$

The specific heat becomes

$$C_v = \frac{\partial U}{\partial T} = 3Nk_b \left(\frac{\hbar \omega_k}{k_b T} \right)^2 e^{\frac{\hbar \omega_k}{k_b T}} \left(e^{\frac{\hbar \omega_k}{k_b T}} - 1 \right)^{-2} \quad (6.158b)$$

One can see that C_v approaches the Dulong–Petit law at high temperatures. Einstein compared the result with diamond and used $T_E = \hbar \omega_k / k_b$ as an adjustable parameter. He found excellent agreement and the parameter T_E led to an average value for ω_k . Despite the success of the model, one needs to circumvent the restrictive assumptions such as independent atoms and a single-oscillation frequency. We have seen that atoms oscillate in a collective mode and with possibly different frequencies.

6.12.3 DEBYE MODEL FOR SPECIFIC HEAT

The Debye model presented here associates the phonon quantum of energy with the collective motion of the atoms. The quantum field theory for the phonon shows that the creation operator adds a single phonon to the collective mode rather than to the single harmonically oscillating atom.

The total internal energy U (per unit volume) for a volume of N -atoms executing 1-D motion must be

$$U = \sum d\omega \frac{\# \text{ states}}{\text{freq.} * \text{Vol.}} \frac{\# \text{ phonons}}{\text{State}} \frac{\text{Energy}}{\text{Phonon}} = \int d\omega g(\omega) n(\omega) \hbar \omega \quad (6.159)$$

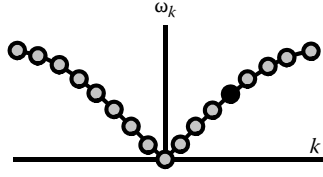


FIGURE 6.53 Phonon states in the FBZ for the monatomic crystal with one polarization.

where $g(\omega)$ represents the number of states per unit energy per unit crystal volume and $E = \hbar\omega$, and as previously discussed, $n(\omega) = \langle n \rangle = \left(e^{\frac{\hbar\omega}{k_b T}} - 1 \right)^{-1}$ represents the number of phonons per state. The macroscopic sized boundary conditions produce states k very close together in Figure 6.53 and so the sum over all the states can be taken as an integral in Equation 6.159.

Assuming a 3-D monatomic crystal, the density of states in Equation 6.159 must account for three polarizations (two transverse and one longitudinal mode). Previous sections in this chapter provide

$$g_{\omega}^{(3-D)} = \frac{3\omega^2}{2\pi^2 v^3} \quad (6.160)$$

for phonons in the acoustic branch in the long-wavelength limit with equal propagation speeds v in the three propagation directions. The factor of three in this last equation accounts for the three possible polarizations.

It might appear reasonable to integrate from 0 to ∞ in Equation 6.159 since the number of phonons $\langle n \rangle$ exponentially decreases with energy as shown in Equation 6.156a. However, rather than taking the integral to infinity, Debye suggested a maximum cutoff energy $E_{\max} = \hbar\omega_{\max}$ found from the density of states $g(\omega)$. Assuming N -atoms in the volume V , the total number of degrees of freedom must be

$$3N/V = \int_0^{\omega_{\max}} g(\omega) d\omega \quad (6.161a)$$

The integration provides

$$\omega_{\max} = \left(\frac{6\pi^2 v^3 N}{V} \right)^{1/3} \quad (6.162b)$$

where

N/V gives the number of atoms per crystal volume

v denotes the phonon speed

Debye defined the Debye temperature T_D as $E_{\max} = k_b T_D$.

The internal energy (per unit volume) in Equation 6.159 becomes

$$U = \frac{3\hbar}{2\pi^2 v^3} \int_0^{\omega_{\max}} d\omega \frac{\omega^3}{e^{\frac{\hbar\omega}{k_b T}} - 1} = \frac{3k_b^4 T^4}{2\pi^2 v^3 \hbar^3} \int_0^{x_{\max}} dx \frac{x^3}{e^x - 1} \quad (6.163)$$

where $x = E/(k_b T) = \hbar\omega_k/(k_b T)$ and $x_{\max} = E_{\max}/(k_b T) = \hbar\omega_{\max}/(k_b T) = T_D/T$. Differentiating the energy integral with respect to temperature T in Equation 6.163 produces the specific heat C (per volume).

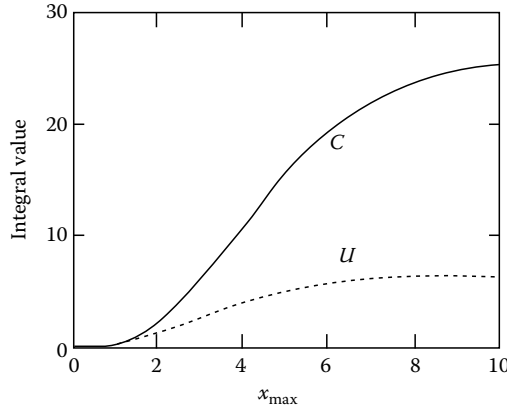


FIGURE 6.54 The integrals in Equations 6.163 and 6.164 (represent by U and C) reach asymptotic values of 6.5 and 26.0, respectively.

$$C = \frac{\partial U}{\partial T} = 9k_b \frac{N}{V} \left(\frac{T}{T_D} \right)^3 \int_0^{x_{\max}} dx \frac{x^4 e^x}{(e^x - 1)^2} \quad (6.164)$$

The integrals in Equations 6.163 and 6.164 appear in Figure 6.54.

Example 6.10

Show that the specific heat decreases with temperature at low T .

SOLUTION

The integral in Equation 6.164 is a constant. Therefore, the expression for C decreases as T^3 as the temperature decreases. This occurs because at low temperature, fewer modes come into play and therefore the number of “places” to store energy must decrease.

6.13 QUANTUM MECHANICAL DEVELOPMENT OF PHONON FIELDS

This section states and then quantizes the Hamiltonian for the phonon. The generalized coordinates consist of the position and momentum of each atom. The amplitudes in the Fourier transform of these phase space coordinates can be taken as operators satisfying commutation relations. These operators operate on an amplitude Hilbert space to define the amplitude of the phonon wave. The resulting Hamiltonian has the simple form of a harmonic oscillator form with the energy of the waves can be deduced in terms of energy quanta. The procedure is very similar to that for electromagnetic waves covered in detail in the *Physics of Optoelectronics*. In this section, we discuss the simplest case of a linear array of atom undergoing 1-D wave motion.

The phase space coordinates are defined on the lattice and are required to satisfy periodic boundary conditions. The boundary conditions produce the discrete states. Reminiscent of the basis vectors in the linear algebra chapters, we first show the basis vectors for the discrete Fourier series. Next, we develop the Lagrangian in order to identify the correct conjugate variables. Then, the Hamiltonian is stated and reduced to the form for simple harmonic oscillators.

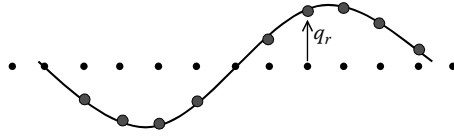


FIGURE 6.55 The mode and the displacement q_r for atom $\#r$.

6.13.1 BASIS STATES FOR FOURIER SERIES WITH PERIODIC BOUNDARY CONDITIONS

In this section we use the conventional notation of q_r, p_r to represent the displacement and momentum of an atom at lattice position r (an index as shown in Figure 6.55). The collective motion of the atoms produces traveling plane waves.

We assume the waves obey periodic boundary conditions over the length L , which encloses N -atoms. This means that the q_r, p_r must repeat every length L according to $q_{r+N} = q_r$ and $p_{r+N} = p_r$. The Fourier series must contain components that have k -vectors of the form

$$k = \frac{2\pi n}{aN} \quad n = \pm 1, \pm 2, \dots, \pm \frac{N}{2} \quad (6.165)$$

where a represents the lattice constant. We know that the Fourier summation can only use these wave vectors. We therefore use

$$q_r = \sum_k Q_k \frac{e^{ikra}}{\sqrt{N}} \quad (6.166)$$

where

k is one of the vectors in Equation 6.165

r is an integer “specifying” a lattice point so that ra is the lattice point

We do not specify the p_r at this time since we need to correctly identify the conjugate momentum.

The basis vectors for the expansion in Equation 6.166 consist of the complex exponential $e^{ikra}/\sqrt{N}$. First, let us demonstrate that the basis vectors have the correct normalization.

$$\left\langle \frac{e^{iKra}}{\sqrt{N}} \left| \frac{e^{ikra}}{\sqrt{N}} \right. \right\rangle \equiv \sum_{r=1}^N \frac{e^{i(k-K)ra}}{N} = \delta_{kK} \quad (6.167)$$

The result is obvious for the case of $k = K$. On the other hand, suppose $k \neq K$. Then

$$(K - k)ra = \frac{2\pi(n - m)}{aN} ra = \frac{2\pi n'}{N}$$

where $n' = r(n - m)$ must be an integer. Because $e^{\frac{2\pi n'}{N}}$ is periodic over N , we can assume that n' is between 0 and N . For example, $e^{\frac{2\pi(n' - N)}{N}} = e^{\frac{2\pi n'}{N}}$. We can either use a graphical approach as illustrated in Figure 6.56 for the example case $N = 8$ or use an algebraic approach. The graphical case produces zero since the vectors are equally spaced around the circle in the complex plane and must therefore add to zero. For the algebraic approach, we use the result

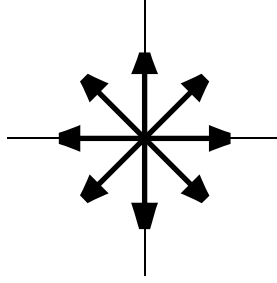


FIGURE 6.56 The vectors $e^{i\frac{2\pi n}{N}}$ for $n = 0$ to $N - 1$.

$$y = \sum_{n=0}^{N-1} x^n = 1 + x(1 + x + \cdots + x^{N-2}) = 1 + x(y - x^{N-1}) \rightarrow y = \frac{1 - x^N}{1 - x}$$

with $x = e^{i\frac{2\pi}{N}}$. We therefore find $\sum_{n=0}^{N-1} e^{i\frac{2\pi n}{N}} = \frac{1 - e^{i\frac{2\pi N}{N}}}{1 - e^{i\frac{2\pi}{N}}} = 0$ as required.

Using the same ideas we can also show the closure relation trivially follows. Both types of phase space coordinates must be real. Working with q_r we then find conditions on the Q_k as follows.

$$q_r = q_r^* \rightarrow \sum_k Q_k \frac{e^{ikra}}{\sqrt{N}} = \sum_k Q_k^* \frac{e^{-ikra}}{\sqrt{N}}$$

The last summation is over all k both positive and negative so that a sum over $-k$ must be the same as the sum over $+k$. The last equation yields

$$\sum_k Q_k \frac{e^{ikra}}{\sqrt{N}} = \sum_k Q_{-k}^* \frac{e^{ikra}}{\sqrt{N}}$$

Multiplying both sides by the projector

$$\left\langle \frac{e^{iKra}}{\sqrt{N}} \right| \sim \sum_{r=1}^N \frac{e^{-iKra}}{\sqrt{N}}$$

which is considered to be an operator, produces the results

$$\sum_{r=1}^N \frac{e^{-iKra}}{\sqrt{N}} \sum_k Q_k \frac{e^{ikra}}{\sqrt{N}} = \sum_{r=1}^N \frac{e^{-iKra}}{\sqrt{N}} \sum_k Q_{-k}^* \frac{e^{ikra}}{\sqrt{N}} \rightarrow \sum_k Q_k \sum_{r=1}^N \frac{e^{i(k-K)ra}}{N} = \sum_k Q_{-k}^* \sum_{r=1}^N \frac{e^{i(k-K)ra}}{N}$$

Using the Kronecker delta relation in Equation 6.167 produces

$$Q_k = Q_{-k}^* \quad (6.168)$$

6.13.2 LAGRANGIAN FOR LINE OF ATOMS

We develop the Lagrangian for several reasons. First, it provides the Hamiltonian. Second, and most important, it provides the momentum conjugate to the generalized coordinates q_r and Q_k . Third, it leads the correct Fourier series expansion of the momentum. We assume N -atoms in a length L with

spacing a . For an infinitely long array of atoms, we still apply the periodic boundary conditions. However, the interpretation of the energy remains in question since an infinitely long line would have infinite energy. We will see similar behavior in the light field.

The Lagrangian incorporates the kinetic and potential energy for each atom in the line.

$$L = T - V = \sum_{r=1}^N \frac{m\dot{q}_r^2}{2} - \sum_{r=1}^N \frac{\gamma}{2} (q_{r+1} - q_r)^2 \quad (6.169)$$

where m and γ represent the mass of each atom and the effective spring constant, respectively. We can see that the potential energy agrees with the force equation given at the start of Section 6.6 by finding the force on atom #s as follows.

$$F = -\frac{\partial V}{\partial q_\alpha} = -\frac{\gamma}{2} \frac{\partial}{\partial q_\alpha} [\cdots + (q_{\alpha+1}^2 + q_\alpha^2 - 2q_{\alpha+1}q_\alpha) + (q_\alpha^2 + q_{\alpha-1}^2 - 2q_\alpha q_{\alpha-1}) + \cdots] \quad (6.170a)$$

Differentiating this last equation produces

$$F = -\frac{\gamma}{2} [2q_\alpha - 2q_{\alpha+1} + 2q_\alpha - 2q_{\alpha-1}] = \gamma[(q_{\alpha+1} - q_\alpha) - (q_\alpha - q_{\alpha-1})] \quad (6.170b)$$

which agrees with the results given in Section 6.6

First we find the momentum conjugate to q_r in the usual manner

$$p_r = \frac{\partial L}{\partial \dot{q}_r} = m\dot{q}_r \quad (6.171)$$

The other canonical equation reproduces Newton's second law appearing in Section 6.6

$$F_r = \dot{p}_r = \frac{\partial L}{\partial q_r} = -\frac{\partial V}{\partial q_r} \quad (6.172)$$

as found in Equation 6.179.

Next we find the momentum conjugate to the coordinate Q_k which then leads to the appropriate Fourier series expansion of the momentum p_r . We start with the Lagrangian in Equation 6.169

$$L = T - V = \sum_r \frac{m\dot{q}_r^2}{2} - \sum_r \frac{\gamma}{2} (q_{r+1} - q_r)^2 \quad (6.173)$$

and substitute the quantities

$$q_r = \sum_k Q_k \frac{e^{ikra}}{\sqrt{N}} \quad \dot{q}_r = \sum_k \dot{Q}_k \frac{e^{ikra}}{\sqrt{N}} \quad (6.174)$$

The kinetic energy becomes

$$T = \frac{m}{2} \sum_{r=1}^N \sum_{k,K} \dot{Q}_k \dot{Q}_K e^{i(k+K)ra} = \frac{m}{2} \sum_{k,K} \dot{Q}_k \dot{Q}_K \sum_{r=1}^N e^{i(k+K)ra} = \frac{m}{2} \sum_{k,K} \dot{Q}_k \dot{Q}_K \delta_{k,-K} = \frac{m}{2} \sum_k \dot{Q}_k \dot{Q}_{-k} \quad (6.175)$$

The potential energy

$$V = \frac{\gamma}{2} \sum_{r=1}^N (q_{r+1} - q_r)^2 = \frac{\gamma}{2} \sum_{r=1}^N (q_{r+1}^2 + q_r^2 - 2q_{r+1}q_r) \quad (6.176)$$

can be written in terms of the Fourier amplitudes by substituting Equation 6.174

$$V = \frac{\gamma}{2} \sum_{k,K} \sum_{r=1}^N \left(\underbrace{Q_k Q_K \frac{e^{i(k+K)(r+1)a}}{N}}_{\text{Term 1}} + \underbrace{Q_k Q_K \frac{e^{i(k+K)ra}}{N}}_{\text{Term 2}} - \underbrace{2Q_k Q_K \frac{e^{ik(r+1)a}}{\sqrt{N}} \frac{e^{iKra}}{\sqrt{N}}}_{\text{Term 3}} \right) \quad (6.177)$$

The first term provides

$$\text{Term 1} = \sum_{k,K} e^{i(k+K)a} Q_k Q_K \sum_{r=1}^N \left(\frac{e^{i(k+K)ra}}{N} \right) = \sum_{k,K} e^{i(k+K)a} Q_k Q_K \delta_{k,-K} = \sum_k Q_k Q_{-k} \quad (6.178a)$$

Similarly, we can rewrite the second term

$$\text{Term 2} = \sum_{k,K} Q_k Q_K \sum_{r=1}^N \frac{e^{i(k+K)ra}}{N} = \sum_{k,K} Q_k Q_K \delta_{k,-K} = \sum_k Q_k Q_{-k} \quad (6.178b)$$

Examine the third term. Since Term 3 has a factor of 2, the summand can be split into two terms as follows:

$$\text{Term 3} = - \sum_{k,K} \sum_{r=1}^N 2Q_k Q_K \frac{e^{ik(r+1)a}}{\sqrt{N}} \frac{e^{iKra}}{\sqrt{N}} = - \sum_{k,K} \sum_{r=1}^N Q_k Q_K \left(\frac{e^{ik(r+1)a}}{\sqrt{N}} \frac{e^{iKra}}{\sqrt{N}} + \frac{e^{ikra}}{\sqrt{N}} \frac{e^{iK(r+1)a}}{\sqrt{N}} \right)$$

Notice the $(r+1)$ has been shifted from the k -part to the K -part in the second set of fractions. We can do this because the k and K in the summations are dummy indices and can be interchanged. Now we find using the orthonormality relation:

$$\text{Term 3} = - \sum_{k,K} \sum_{r=1}^N Q_k Q_K \left(e^{ikra} \frac{e^{i(k+K)ra}}{N} + e^{iKra} \frac{e^{i(k+K)ra}}{N} \right) = - \sum_k Q_k Q_{-k} (e^{ika} + e^{-ika}) \quad (6.178c)$$

Combine all three terms in Equations 6.178a through c back into Equation 6.176 to get

$$V = \gamma \sum_k Q_k Q_{-k} [1 - \cos(ka)] \quad (6.179)$$

Finally, combining Equations 6.175 and 6.179 into Equation 6.173 produces the full Lagrangian

$$L = T - V = \frac{m}{2} \sum_k \dot{Q}_k \dot{Q}_{-k} - \gamma \sum_k Q_k Q_{-k} [1 - \cos(ka)] \quad (6.180)$$

Now we can find the momentum conjugate to the Q_k . The momentum P_k is defined as

$$P_k = \frac{\partial L}{\partial \dot{Q}_k} \quad (6.181)$$

To correctly evaluate this for momentum P_s for example, we need to include both $+s$ and $-s$ from the first term in the Lagrangian in Equation 6.180.

$$L = T - V = \frac{m}{2} (\cdots + \dot{Q}_s \dot{Q}_{-s} + \dot{Q}_{-s} \dot{Q}_s + \cdots) - \gamma \sum_k Q_k Q_{-k} [1 - \cos(ka)]$$

Differentiation provides

$$P_s = \frac{\partial L}{\partial \dot{Q}_s} = m \dot{Q}_{-s} \quad (6.182)$$

Finally, we can write the appropriate Fourier series expansion for the conjugate momentum p_r to accompany the expansion of the generalized coordinates q_r in Equation 6.166.

$$q_r = \sum_k Q_k \frac{e^{ikra}}{\sqrt{N}}$$

Differentiating with respect to time, multiplying by mass m , and using Equations 6.171 and 6.182 gives

$$p_r = m \dot{q}_r = \sum_k m \dot{Q}_k \frac{e^{ikra}}{\sqrt{N}} = \sum_k P_{-k} \frac{e^{ikra}}{\sqrt{N}} = \sum_k P_k \frac{e^{-ikra}}{\sqrt{N}} \quad (6.183)$$

where the last summation follows by redefining the index. Therefore, the Fourier series for q_r and p_r differ in the sign of the argument of the exponential.

6.13.3 CLASSICAL HAMILTONIAN

The Hamiltonian can be written by either adding together the kinetic T and potential V energy terms in Equation 6.173 or by using the definition of the Hamiltonian through the Legendre transformation given in the chapter on *Dynamics* (Chapter 4).

$$H = \sum_r p_r \dot{q}_r - L$$

In either case, the Hamiltonian becomes

$$H = T + V = \sum_r \frac{p_r^2}{2m} + \sum_r \frac{\gamma}{2} (q_{r+1} - q_r)^2 \quad (6.184)$$

where the Hamiltonian uses the momentum rather than the generalized velocities. Similar to the development for the Lagrangian, the potential energy can be rewritten as in Equation 6.179

$$V = \gamma \sum_k Q_k Q_{-k} [1 - \cos(ka)] \quad (6.185)$$

The kinetic term can also be rewritten using the Fourier series in Equation 6.183 and the orthonormality relation in Equation 6.167

$$\begin{aligned} T &= \sum_r \frac{p_r^2}{2m} = \frac{1}{2m} \sum_r \sum_{k,K} P_k \frac{e^{-ikra}}{\sqrt{N}} P_K \frac{e^{-iKra}}{\sqrt{N}} = \frac{1}{2m} \sum_{k,K} P_k P_K \sum_r \frac{e^{-i(k+K)ra}}{N} = \frac{1}{2m} \sum_{k,K} P_k P_K \delta_{k,-K} \\ &= \frac{1}{2m} \sum_k P_k P_{-k} \end{aligned}$$

Therefore, the Hamiltonian becomes

$$H = T + V = \frac{1}{2m} \sum_k P_k P_{-k} + \gamma \sum_k Q_k Q_{-k} [1 - \cos(ka)] \quad (6.186)$$

Note that the Q_k , Q_{-k} , P_k , and P_{-k} are considered to be independent.

6.13.4 INTRODUCTION TO QUANTIZING PHONON FIELD AND HAMILTONIAN

The phonon field is quantized by requiring the Fourier series amplitudes to be operators satisfying commutation relations. We require the conjugate variables

$$\hat{q}_r = \sum_k \hat{Q}_k \frac{e^{ikra}}{\sqrt{N}} \quad \text{and} \quad \hat{p}_r = \sum_k \hat{P}_k \frac{e^{-ikra}}{\sqrt{N}} \quad (6.187)$$

to satisfy

$$[\hat{q}_r, \hat{p}_s] = i\hbar \delta_{r,s} \quad [\hat{q}_r, \hat{q}_s] = 0 \quad [\hat{p}_r, \hat{p}_s] = i\hbar \delta_{r,s} \quad (6.188a)$$

These lead to the commutation relations for Q_k and P_k .

$$[Q_k, P_K] = \left[\sum_{r=1}^N q_r \frac{e^{-ikra}}{\sqrt{N}}, \sum_{s=1}^N p_s \frac{e^{+iKsa}}{\sqrt{N}} \right] = \sum_{r,s=1}^N \frac{e^{-ikra} e^{+iKsa}}{N} [q_r, p_s] = \sum_{r,s=1}^N \frac{e^{-ikra} e^{+iKsa}}{N} \delta_{rs} = \sum_{r,s=1}^N \frac{e^{i(K-k)ra}}{N}$$

With similar results for the other commutators in Equation 6.187

$$[Q_k, P_K] = i\hbar \delta_{k,K} \quad [Q_k, Q_K] = 0 \quad [P_k, P_K] = 0 \quad (6.188b)$$

We will need to specify a Hilbert space in order to give these operators meaning. But first, we develop the quantum phonon Hamiltonian.

We now show the quantum phonon Hamiltonian. We could start with Equation 6.184, make the dynamical variables into operators, and substitute the expansions while using the commutation relations. However, the development used for the Fourier series version of the classical Hamiltonian did not commute dynamical variables that cannot be quantum mechanically commuted. Therefore, we start with the result given in Equation 6.186. We use the Fourier series since each k refers to an allowed state on the dispersion curve.

$$\hat{H} = \frac{1}{2m} \sum_k \hat{P}_k \hat{P}_{-k} + \gamma \sum_k \hat{Q}_k \hat{Q}_{-k} [1 - \cos(ka)] \quad (6.189a)$$

We would like to define creation $\hat{a}^+$ and annihilation $\hat{a}$ operators in order to give the Hamiltonian the same form as for the electron harmonic oscillator discussed in the quantum theory

$$\hat{H} = \sum_{k=1}^N \hbar \omega_k (\hat{n}_k + 1/2) \quad \text{with} \quad \omega_k = \sqrt{\frac{\gamma}{m}} \sqrt{2[1 - \cos(ka)]} \quad (6.189b)$$

and $\hat{n}_k = \hat{a}_k^+ \hat{a}_k$. Notice that $\omega_{\vec{k}} = \omega_{-\vec{k}}$. The reader might find it useful to read part of the next section to better understand the notation and physical significance. Basically, the number operator $\hat{n}_{\vec{k}}$ gives the number of phonons in mode $\vec{k}$. Therefore Equation 6.189b adds up the energy of all phonons in all modes. The Hamiltonian has the form of a harmonic oscillator since the individual atoms execute harmonic motion.

To show the number representation of the phonon Hamiltonian given in Equation 6.189b, we define the creation and annihilation operators in a manner similar to that for the electron harmonic oscillator.

$$\hat{a}_k = \frac{m\omega_k}{\sqrt{2\hbar m\omega_k}} \hat{Q}_{-k} + \frac{i}{\sqrt{2\hbar m\omega_k}} \hat{P}_k \quad \hat{a}_k^+ = \frac{m\omega_k}{\sqrt{2\hbar m\omega_k}} \hat{Q}_k - \frac{i}{\sqrt{2\hbar m\omega_k}} \hat{P}_{-k} \quad (6.190)$$

where $\hat{Q}_k = \hat{Q}_{-k}^+$ and $\hat{P}_k = \hat{P}_{-k}^+$. Using the commutation relations in Equation 6.188b, we can find the commutation relations for the creation and annihilation operators. For example,

$$[\hat{a}_k, \hat{a}_k^+] = \frac{-im\omega_k}{2\hbar m\omega_k} [\hat{Q}_{-k}, \hat{P}_{-k}] + \frac{im\omega_k}{2\hbar m\omega_k} [\hat{P}_k, \hat{Q}_k] = \frac{-i}{\hbar} i\hbar = 1$$

We find the following relations

$$[\hat{a}_{\vec{k}}, \hat{a}_{\vec{k}}^+] = \delta_{\vec{k}, \vec{k}} \quad [\hat{a}_{\vec{k}}, \hat{a}_{\vec{k}}] = 0 \quad [\hat{a}_{\vec{k}}^+, \hat{a}_{\vec{k}}^+] = 0 \quad (6.191)$$

Expressions for Q and P can be found by simultaneously solving Equation 6.190.

$$\hat{Q}_k = \sqrt{\frac{\hbar}{2m\omega_k}} [\hat{a}_{-k} + \hat{a}_k^+] \quad \hat{P}_k = -i\sqrt{\frac{\hbar m\omega_k}{2}} [\hat{a}_{\vec{k}} - \hat{a}_{-\vec{k}}^+] \quad (6.192)$$

The Hamiltonian in Equation 6.189a can now be written in terms of the creation and annihilation operators. Starting with

$$\hat{H} = \frac{1}{2m} \sum_k \hat{P}_k \hat{P}_{-k} + \gamma \sum_k \hat{Q}_k \hat{Q}_{-k} [1 - \cos(ka)]$$

we find

$$\hat{H} = \sum_k \frac{-\hbar\omega_k}{4} [\hat{a}_{\vec{k}} - \hat{a}_{-\vec{k}}^+] [\hat{a}_{-\vec{k}} - \hat{a}_{+\vec{k}}^+] + \gamma \sum_k \frac{\hbar}{2m\omega_k} [\hat{a}_{-\vec{k}} + \hat{a}_{\vec{k}}^+] [\hat{a}_{\vec{k}} + \hat{a}_{-\vec{k}}^+] [1 - \cos(ka)]$$

Notice that the last term can be rewritten using the second of Equation 6.189b.

$$\hat{H} = \sum_k \frac{-\hbar\omega_k}{4} [\hat{a}_{\vec{k}} - \hat{a}_{-\vec{k}}^+] [\hat{a}_{-\vec{k}} - \hat{a}_{+\vec{k}}^+] + \sum_k \frac{\hbar\omega_k}{4} [\hat{a}_{-\vec{k}} + \hat{a}_{\vec{k}}^+] [\hat{a}_{\vec{k}} + \hat{a}_{-\vec{k}}^+]$$

Using commutators, we can cancel some terms to find

$$\hat{H} = \sum_k \frac{\hbar\omega_k}{4} [\hat{a}_{\vec{k}}\hat{a}_{\vec{k}}^+ + \hat{a}_{-\vec{k}}^+\hat{a}_{-\vec{k}} + \hat{a}_{-\vec{k}}\hat{a}_{-\vec{k}}^+ + \hat{a}_{\vec{k}}^+\hat{a}_{\vec{k}}] = \sum_k \frac{\hbar\omega_k}{2} [\hat{a}_{\vec{k}}\hat{a}_{\vec{k}}^+ + \hat{a}_{\vec{k}}^+\hat{a}_{\vec{k}}]$$

The second equality obtains since the sums in the Hamiltonian contain both positive and negative k and the Hamiltonian can therefore be rewritten in terms of $+k$. Next, using the commutation relations in Equation 6.191 yields

$$\hat{H} = \sum_k \hbar\omega_k [\hat{a}_{\vec{k}}^+\hat{a}_{\vec{k}} + 1/2] = \sum_k \hbar\omega_k [\hat{n}_{\vec{k}} + 1/2] \quad (6.193)$$

6.13.5 INTRODUCTION TO PHONON FOCK STATES

For the operators to have meaning, we must specify a Hilbert space.

The Hilbert space (amplitude space) can be defined by the Fock basis vectors that have the form

$$|n_1, n_2, \dots, n_k, \dots, n_N\rangle$$

where each position represents one of the states in the dispersion curve. The integer n_k may take on any non-negative value. Figure 6.57 shows that we think of each *position* in the ket as if it were a bucket and n_k represents the number of phonons in that bucket. The creation operators add one phonon to a given state while the annihilation operator removes one phonon. The following relations hold.

$$\begin{aligned} \hat{a}_k^+ |n_1, n_2, \dots, n_k, \dots, n_N\rangle &= \sqrt{n_k + 1} |n_1, n_2, \dots, n_k + 1, \dots, n_N\rangle \\ \hat{a}_k |n_1, n_2, \dots, n_k, \dots, n_N\rangle &= \sqrt{n_k} |n_1, n_2, \dots, n_k - 1, \dots, n_N\rangle \\ \hat{a}_k |0, 0, \dots, 0, \dots, 0\rangle &= 0 \end{aligned} \quad (6.194)$$

The modes can be written as the tensor product

$$|n_1, n_2, \dots, n_k, \dots, n_N\rangle = |n_1\rangle |n_2\rangle \cdots |n_k\rangle \cdots |n_N\rangle$$

The orthonormality relation has the form

$$\langle m_1, m_2, \dots, m_k, \dots, m_N | n_1, n_2, \dots, n_k, \dots, n_N \rangle = \delta_{m_1, n_1} \delta_{m_2, n_2} \cdots \quad (6.195a)$$

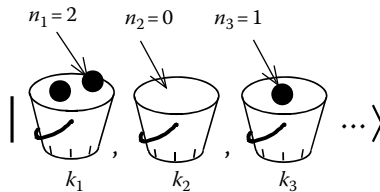


FIGURE 6.57 The Fock state describes the number of particles in the modes or states of the system. The diagram represents the ket $|2, 0, 1, \dots\rangle$.

or in shortened notation

$$\langle \{m_k\} | \{n_k\} \rangle = \delta_{\{m_k\}, \{n_k\}} \quad (6.195b)$$

The closure relation has the form

$$\hat{1} = \sum_{n_1, n_2, \dots, n_N=1}^{\infty} |n_1, n_2, \dots, n_k, \dots, n_N\rangle \langle n_1, n_2, \dots, n_k, \dots, n_N| \quad (6.196a)$$

and in shortened notation

$$\hat{1} = \sum_{\{n_k\}} |\{n_k\}\rangle \langle \{n_k\}| \quad (6.196b)$$

The relations can be written in a coordinate representation but we will reserve the development until the chapter on light.

The number operator $\hat{n}_k = \hat{a}_k^\dagger \hat{a}_k$ provides the number of phonons in mode k

$$\hat{n}_k |n_1, n_2, \dots, n_k, \dots, n_N\rangle = \hat{a}_k^\dagger \hat{a}_k |n_1, n_2, \dots, n_k, \dots, n_N\rangle = n_k |n_1, n_2, \dots, n_k, \dots, n_N\rangle$$

Therefore, the number operators have the Fock states as eigenstates, which provides a concise definition for the Fock states. The operator for the total number of phonons in the system must be

$$\hat{N} = \sum_k \hat{n}_k \quad (6.197a)$$

so that

$$\hat{N} |n_1, n_2, \dots, n_k, \dots, n_N\rangle = \sum_a \hat{n}_a |n_1, n_2, \dots, n_k, \dots, n_N\rangle = \sum_a n_a |n_1, n_2, \dots, n_k, \dots, n_N\rangle \quad (6.197b)$$

and the total number must be $\sum_a n_a$.

According to the previous section, the total energy in Equation 6.193 becomes

$$\hat{H} |n_1, n_2, \dots, n_k, \dots, n_N\rangle = \sum_k \hbar \omega_k \left[n_k + 1/2 \right] |n_1, n_2, \dots, n_k, \dots, n_N\rangle \quad (6.198)$$

The vacuum state $|0\rangle = |0, 0, \dots, 0\rangle$ represents the case when the system does not have any available quanta. However, the vacuum state has a residual energy

$$\sum_k \frac{1}{2} \hbar \omega_k$$

This corresponds to the zero-point motion of the field.

6.14 PHONONS AND CONTINUOUS MEDIA

Previous sections develop expressions for phonon wave motion and the phonon velocity. The previous chapter developed the quantum mechanical expressions for the Hamiltonian and momentum associated with displacement. The present section produces the Hamiltonian and equations of motion for continuous media.

6.14.1 WAVE EQUATION AND SPEED

Consider the case a transverse wave moving through a continuous medium as a continuation of the development in Section 4.6. The Lagrange density must have the form of the kinetic energy minus the potential energy. Given that the Lagrange density $\mathcal{L}$ represents energy per volume, the kinetic and potential energy terms must have similar units. For the case of 1-D wave motion, the kinetic energy per length has the form

$$T = \frac{KE}{\Delta z} = \frac{1}{\Delta z} \frac{1}{2} m \dot{\eta}^2 = \frac{1}{2} \frac{m}{\Delta z} \dot{\eta}^2 = \frac{1}{2} \rho \dot{\eta}^2 \quad (6.199)$$

where ρ gives the mass density.

Now we must find the potential energy per unit volume. Figure 6.58 schematically represents the forces acting between adjacent elements of mass by interconnected springs. We assume the springs exert negligible force along the z -direction. The force on the i th mass must be related to its relative displacement from mass $(i - 1)$ and mass $(i + 1)$. The force on the i th mass must be

$$F_i = -k(n_i - n_{i-1}) - k(n_i - n_{i+1}) \quad (6.200)$$

We can either incorporate the force into the Lagrangian formalism or else note that the springs provide a conservative force and can be represented by a potential function. The correct potential function can be seen to be

$$PE = \frac{1}{2} \sum_i k(\eta_{i+1} - \eta_i)^2 = \frac{k}{2} [\cdots + (\eta_{i+2} - \eta_{i+1})^2 + (\eta_{i+1} - \eta_i)^2 + (\eta_i - \eta_{i-1})^2 + \cdots]$$

by calculating

$$\begin{aligned} F_i &= \frac{\partial V}{\partial \eta_i} = \frac{k}{2} \frac{\partial}{\partial \eta_i} [\cdots + (\eta_{i+1} - \eta_i)^2 + (\eta_i - \eta_{i-1})^2 + \cdots] \\ &= \frac{k}{2} [0 + 2(\eta_{i+1} - \eta_i)(-1) + 2(\eta_i - \eta_{i-1})(+1) + \cdots] \end{aligned}$$

Notice that the potential (PE) gives the potential energy for the entire medium since the sum includes all the pieces denoted by i . By going to the limit of small volume ($\Delta z \rightarrow 0$ for the 1-D case shown in the figure), we can find the potential energy per unit volume.

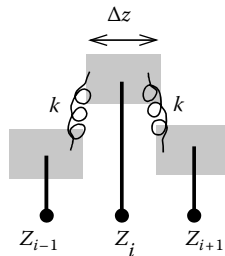


FIGURE 6.58 Force between neighboring mass elements.

$$\text{PE} = \sum_i \Delta z \frac{(k\Delta z)}{2} \left(\frac{\eta_{i+1} - \eta_i}{\Delta z} \right)^2 \rightarrow \int dz \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \quad (6.201)$$

where Young's modulus $Y = \lim_{\Delta z \rightarrow 0} k\Delta z$ reaches a nonzero limit because as spring length decreases, the spring constant increases. Equation 6.201 yields the potential energy per unit distance (1-D)

$$U = \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \quad (6.202)$$

Combining Equations 6.199 and 6.202 with the Lagrange density provides

$$\mathcal{L} = T - U = \frac{1}{2} \rho \dot{\eta}^2 - \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \quad (6.203)$$

We can find the equations of motion by applying Lagrange's equation found in Section 4.6 (here, use the repeated index convention: the last term has a summation over i and where $\partial_i = \partial/\partial x_i$ and $x_1 = x$, $x_2 = y$, $x_3 = z$)

$$\frac{\partial \mathcal{L}}{\partial \eta} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} - \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} = 0 \quad (6.204)$$

Calculating the following derivatives

$$\frac{\partial \mathcal{L}}{\partial \eta} = \frac{\partial}{\partial \eta} \left[\frac{1}{2} \rho \dot{\eta}^2 - \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \right] = 0 \quad (6.205a)$$

$$\frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} = \frac{\partial}{\partial t} \frac{\partial}{\partial \dot{\eta}} \left[\frac{1}{2} \rho \dot{\eta}^2 - \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \right] = \frac{\partial}{\partial t} (\rho \dot{\eta}) = \rho \ddot{\eta} \quad (6.205b)$$

$$\frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} = \frac{\partial}{\partial (\partial_i \eta)} \left[\frac{1}{2} \rho \dot{\eta}^2 - \frac{Y}{2} (\partial_z \eta)^2 \right] = -\delta_{iz} Y \partial_z \eta \rightarrow \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} = -\delta_{iz} Y \partial_i \partial_z \eta = -Y \frac{\partial^2 \eta}{\partial z^2} \quad (6.205c)$$

(where we sum over repeated indices in the last equation) and combining the last three results into Equation 6.204 provides

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \eta} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} - \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} &= 0 \\ 0 - \rho \ddot{\eta} + Y \frac{\partial^2 \eta}{\partial z^2} &= 0 \\ \frac{\partial^2 \eta}{\partial z^2} - \frac{1}{(Y/\rho)} \frac{\partial^2 \eta}{\partial t^2} &= 0 \end{aligned} \quad (6.206)$$

This is a wave equation with a propagation speed

$$v = \sqrt{Y/\rho} \quad (6.207)$$

Next, we derive the Hamiltonian for the continuous system.

6.14.2 HAMILTONIAN FOR ONE-DIMENSIONAL WAVE MOTION

We now know that the Lagrangian for a continuous system depends on a set of generalized coordinates $\{\eta_{\vec{r}}(t) = \eta(\vec{r}, t)\}$, where each point in space has a coordinate associated with it (for example, displacement). There can be more than one coordinate associated with each point. For example, at each point we can have the three components of polarization. In general, the Lagrangian depends on the generalized velocity and the spatial derivatives of the fields as well.

$$L = L(\eta, \dot{\eta}, \partial_i \eta) = L(\eta, \dot{\eta}, \partial_1 \eta, \partial_2 \eta, \partial_3 \eta) \quad (6.208)$$

We define the Lagrange density (energy per volume)

$$\mathcal{L}(\eta, \dot{\eta}, \partial_i \eta) \quad (6.209)$$

where, for notational convenience, we use i to mean the 1, 2, and 3 appearing in the previous equation. By minimizing the action

$$I = \int_{t_1}^{t_2} dt L = \int_{t_1}^{t_2} \int_{\vec{r}_1}^{\vec{r}_2} dt d^3x \mathcal{L}(\eta, \dot{\eta}, \partial_i \eta) \quad (6.210)$$

The minimization procedure produces Lagrange's equations

$$\frac{\partial \mathcal{L}}{\partial \eta} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{\eta}} - \partial_i \frac{\partial \mathcal{L}}{\partial (\partial_i \eta)} = 0 \quad (6.211)$$

where we must sum over the repeated index.

Now we can demonstrate the Hamiltonian and Hamiltonian density for the continuous system. The previous section shows that the Hamiltonian density comes from the Hamiltonian by allowing the cell size to approach zero $\Delta V_i \rightarrow 0$. We find

$$H = H(q_i, p_i) = \sum_i p_i \dot{q}_i - L(q_i, \dot{q}_i, \dots) \xrightarrow{\Delta V_i \rightarrow 0} \int dV \{ \pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \mathcal{L}(\eta, \dot{\eta}, \dots) \} = \int dV \mathcal{H}$$

where the Hamiltonian density has the form

$$\mathcal{H} = \pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \mathcal{L}(\eta, \dot{\eta}, \dots) \quad (6.212)$$

The Hamiltonian density gives the energy per unit volume. Integrating over all space gives the total energy. The momentum density is defined by

$$\pi = \frac{\partial \mathcal{L}}{\partial \dot{\eta}} \quad (6.213)$$

similar to the definition of ordinary momentum.

Continuing with the demonstration of the Hamiltonian density, we calculate the momentum density in Equation 6.213 using the Lagrange density in Equation 6.203

$$\mathcal{L} = \frac{1}{2} \rho \dot{\eta}^2 - \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2$$

$$\pi = \frac{\partial \mathcal{L}}{\partial \dot{\eta}} = \rho \dot{\eta}$$

The Hamiltonian density becomes

$$\begin{aligned} \mathcal{H} &= \pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \mathcal{L}(\eta, \dot{\eta}, \dots) = \pi(\vec{r}, t) \dot{\eta}(\vec{r}, t) - \frac{1}{2} \rho \dot{\eta}^2 + \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \\ &= \pi \frac{\pi}{\rho} - \frac{1}{2} \rho \left(\frac{\pi}{\rho} \right)^2 + \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 = \frac{\pi^2}{2\rho} + \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 \end{aligned}$$

REVIEW EXERCISES

Note to the reader: The first seven exercises review elementary concepts in solid state, refer to a freshman physics or chemistry book.

- 6.1** Derive the pressure $P = n_o m v^2 / 3 = \rho v^2 / 3$, which is force per area, exerted by an ideal gas on the sidewalls of a container having sides of length d . Follow the substeps below. Here n_o , m , v , and ρ are the number of molecules per unit volume, the mass of a molecule, the average speed of the molecule, and the mass density, respectively.
- Explain why $n_o A d / 6$ represents the average number of molecules traveling in the $+x$ direction where $A = d^2$. All of these will hit the wall during a time d/v . Show the total change in momentum must be $m v n_o A d / 3$.
 - For a molecule traveling along $+x$, describe under what conditions the change in momentum will be $\Delta p = 2mv$. Explain why the impulse can be written as $Ft = Fd/v$ where F represents the average force exerted on the wall by the molecules but t is not the time of impact for a single molecule.
 - Combine parts a and b to find the desired results.
- 6.2** Recall that Avogadro's number $N_A = 6.022 \cdot 10^{23}$ particles/(g mol) represents the quantity in grams of N_A particles. The (g mol) can be found by using the atomic weight of a substance from the periodic table as if it were grams.
- Find the mass of an atom of Silicon.
 - If 1 g of silicon has a volume of 0.429 cm^3 then find the mass density.
 - Using N_A for silicon, find the average spacing between silicon atoms.
 - Find N_A in particles/(kg mol).
- 6.3** Derive the relation between temperature scales
- Using the boiling point of water as 100°C (212°F) and the freezing point as 0°C (32°F), derive a formula relating Fahrenheit (F) to Centigrade (C).
 - Find a relation between Kelvin (K) and Fahrenheit (F).
- 6.4** Recall that the ideal gas law has the form $PV = n_m RT$ where P , V , n_m , R , and T represent the pressure, volume, number of moles, universal gas constant, and the temperature in Kelvin, respectively. Recall also that Boltzmann's constant can be written as $k = R/N_A$ where N_A symbolizes Avogadro's number. Using the results of Problem 6.1 above, show the average kinetic energy of a molecule can be written as

$$\frac{1}{2} m v^2 = \frac{3}{2} k T$$

Hint: Multiply both sides of the results for Problem 6.1 by V . Use the total mass as $M = \rho V$. Combine this result with the ideal gas law. Finally note that the total number of particles in volume V must be $n_m N_A$.

- 6.5** The first law of thermodynamics (conservation of energy) states that energy ΔE added to a system can increase the internal energy ΔU and can be used by the system to do work W so that $\Delta E = \Delta U + W$. The specific heat c_v (constant volume) of a mass M at constant volume relates the internal energy to the temperature T of the mass by $c_v = \frac{1}{M} \frac{\partial U}{\partial T}$ where the volume is held constant. For a solid body with negligible change of volume with temperature, this can be written as $\Delta U = Mc_v \Delta T$. Figure P6.5 shows a simple method for maintaining a constant pressure on a volume V of ideal gas while adding energy from below. The piston has a fixed weight and can move along the y -direction allowing the volume of the gas to adjust as necessary. Define the constant pressure specific heat by $c_p = \frac{1}{M} \frac{\partial E}{\partial T}$ (keeping P constant).
- Show the work done by the gas in moving the piston through a distance of d can be written as $W = PAd = nk\Delta T$ where the symbols (in order of appearance) represent the work (W) done by the gas, the pressure, the cross sectional area (A) of the cylinder, the distance (d) the piston moves, the total number (n) of moles of gas, Boltzmann's constant (k) (see Problems 6.1 and 6.4 above), and the change in temperature.
 - Show $c_v = c_p - R$

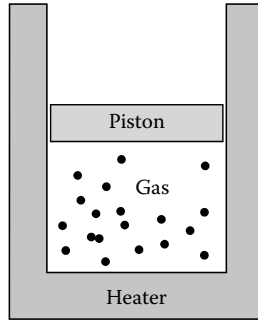


FIGURE P6.5 Gas in the cylinder can move the piston. The piston keeps the gas pressure constant.

- 6.6** The classical equipartition theorem states that internal energy will be divided equally among all degrees of freedom. A degree of freedom defines a possible motion of the system. Consider a molecule consisting of two point masses m interconnected by a massless spring. Assume the center of mass can move in three directions (x, y, z) but that only rotations of the entire system (two masses plus the spring) about the y and z axis lead to any energy; this molecule appears to have 5 degrees of freedom. Explain why the internal energy must have the form $\Delta U = \frac{5}{2}nk\Delta T$ where n represents the total number of molecules, k represents Boltzmann's constant, and ΔT represents the change in temperature. Show $c_v = 2.5R$ based on the definition given in Problem 6.5.
- 6.7** Silicon has a diamond structure with a conventional lattice constant $a = 0.543$ nm. Silicon has an atomic weight of 28. Calculate the mass density. Protons and neutrons have atomic weights of approximately 1.67×10^{-27} kg.
- 6.8** Show the functions $|S\rangle, |X\rangle, |Y\rangle, |Z\rangle$ are orthonormal.
- 6.9** Consider the function

$$f(x) = \begin{cases} +1 & 0 < x < L \\ -1 & -L < x < 0 \end{cases}$$

Find the Fourier expansion coefficients using the basis

$$\left\{ \frac{e^{in\pi x/L}}{\sqrt{2L}} : n = 0, \pm 1, \pm 2, \dots \right\}$$

- 6.10** Find the primitive reciprocal lattice vectors corresponding to the direct primitive lattice vectors given by

$$\vec{a}_1 = \hat{x} \quad \vec{a}_2 = 2\hat{y} \quad \vec{a}_3 = 3\hat{z}$$

- 6.11** Find the primitive reciprocal lattice vectors corresponding to the FCC direct lattice. Show the reciprocal lattice vectors span a BCC lattice.
- 6.12** Show the reciprocal lattice vectors corresponding to the BCC direct lattice produce an FCC reciprocal lattice.
- 6.13** Show that $|(\vec{A} \times \vec{B}) \cdot \vec{C}|$ gives the volume enclosed by the three vectors.
- 6.14** Write $(\vec{A} \times \vec{B}) \cdot \vec{C}$ in terms of the totally antisymmetric tensor $\epsilon_{ijk} \dots$ defined in Chapter 3. Use the antisymmetric tensor to show $\vec{A} \times \vec{B} = -\vec{B} \times \vec{A}$. What happens when the three vectors in $(\vec{A} \times \vec{B}) \cdot \vec{C}$ are permuted?
- 6.15** Assume $\vec{A} = A_x\hat{x} + A_y\hat{y} + A_z\hat{z}$, $\vec{B} = B_x\hat{x} + B_y\hat{y} + B_z\hat{z}$, $\vec{C} = C_x\hat{x} + C_y\hat{y} + C_z\hat{z}$ and show that

$$\vec{A} \cdot (\vec{B} \times \vec{C}) = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}$$

- 6.16** Show that the reciprocal lattice vector $\vec{G} = h\vec{b}_1 + k\vec{b}_2 + l\vec{b}_3$ can be written as

$$\vec{G} = \frac{2\pi}{a^2} (h\vec{a}_1 + \kappa\vec{a}_2 + l\vec{a}_3)$$

for a cubic lattice with spacing a .

- 6.17** Using Problem 6.16, show the reciprocal lattice vector $\vec{G}$ is perpendicular to the plane (h, κ, l) in the cubic lattice. Assume the intersection points

$$\vec{r}_1 = n_1\vec{a}_1 \quad \vec{r}_2 = n_2\vec{a}_2 \quad \vec{r}_3 = n_3\vec{a}_3$$

where n_1, n_2, n_3 are integers and the ratios $h : \kappa : l = \frac{1}{n_1} : \frac{1}{n_2} : \frac{1}{n_3}$ are maintained by the constant c in $h = \frac{c}{n_1}$ $\kappa = \frac{c}{n_2}$ $l = \frac{c}{n_3}$. *Hint:* Calculate $(\vec{r}_1 - \vec{r}_2) \cdot \vec{G}$ etc. (refer to Figure P6.17).

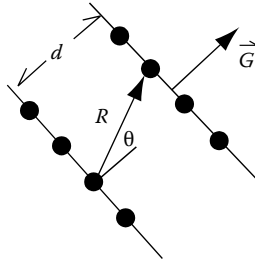


FIGURE P6.17 Adjacent planes.

- 6.18** Show $d = 2\pi/|\vec{G}|$ is the separation between adjacent planes with the vector $\vec{G}$ a reciprocal lattice vector perpendicular to the plane. *Hint:* Consider Figure P6.17 and the definition of the reciprocal lattice vectors in terms of complex exponentials.

6.19 For Section 6.4, show

$$\begin{aligned}\hat{T}_{ma}f(x) &= A_0 + \sum_{n=1}^{\infty} \left[A_n \hat{T}_{ma} \cos\left(\frac{2n\pi x}{a}\right) + B_n \hat{T}_{ma} \sin\left(\frac{2n\pi x}{a}\right) \right] \\ &= A_0 + \sum_{n=1}^{\infty} \left[A_n \cos\left(\frac{2n\pi x}{a}\right) + B_n \sin\left(\frac{2n\pi x}{a}\right) \right] \\ &= f(x)\end{aligned}$$

- 6.20** Consider a 2-D lattice that exhibits 180° rotation symmetry.
- Draw the 2-D lattice. Start with one point at the origin of an x - y -coordinate system.
 - Draw the two primitive vectors and write them in terms of $\tilde{x}$, $\tilde{y}$.
 - Write the rotation matrix for a 180° rotation in terms of the coefficients of the primitive vectors.
 - What is the coordinate transformation matrix $\underline{S}$?
 - What is the rotation matrix in the $\tilde{x}$, $\tilde{y}$ basis for the rotation?
- 6.21** Repeat Problem 6.20 for the case of 120° rotation symmetry.
- 6.22** Repeat Problem 6.20 for the case of 60° rotation symmetry.
- 6.23** Show that when $m = M$ for the diatomic array that the dispersion ω versus k reduces to that for the monatomic array.
- 6.24** Discuss the derivation of the dispersion curves for a triatomic linear crystal. Compare and contrast with that for the monatomic- and diatomic crystal.
- 6.25** Derive all steps in Section 6.6.1 regarding the coupled oscillators and normal modes.
- 6.26** Repeat the analysis in Section 6.6.1 for springs with equal spring constants $\beta_{12} = \beta$. State the oscillation frequencies.
- 6.27** For the coupled oscillators in Section 6.6.1, assume $\beta_{12} \ll 2\beta$ and make a Taylor series expansion of the frequencies to show $\omega_1 \cong \omega_o (1 + \varepsilon)$ and $\omega_2 \cong \omega_o (1 - \varepsilon)$ where $\varepsilon = \beta_{12}/2\beta$.
- 6.28** Find the density of energy states for a 3-D monatomic crystal allowing only one transverse mode for each direction (for phonons). Assume periodic boundary conditions and linear dispersion.
- 6.29** Consider phonons in a 3-D crystal and suppose the periodic boundary conditions use unequal repetition lengths for the three directions L_x , L_y , and L_z . Find the density of states for k_x large compared with $1/L_x$, etc.
- 6.30** The chapter discusses periodic boundary conditions. Assume we have an infinite 3-D crystal having orthogonal sides of length L_x , L_y , and L_z . These lengths define a rectangular lattice.
- Find the primitive reciprocal lattice vectors.
 - Show that the allowed wave vectors for the phonon modes must be one of the reciprocal lattice vectors.
- 6.31** For the 1-D monatomic crystal with dispersion curves

$$\omega_k = 2\sqrt{\frac{\beta}{m}} \left| \sin\left(\frac{ka}{2}\right) \right| = \sqrt{\frac{\beta}{m}} \sqrt{2[1 - \cos(ka)]}$$

Find an expression for the group velocity over the FBZ. Make rough plots of both the dispersion curve and the group velocity on the same plot.

- 6.32** Find the density of states for 1-D longitudinal propagation along the z -axis in the long-wavelength limit.
- 6.33** For the linear monatomic crystal show that neighboring atoms oscillate 180° out-of-phase when k coincides with the wave vector $G_1/2$ demarking the edge of the FBZ.

- 6.34** For the diatomic crystal, determine and show mathematically whether the following statement is true or not: neighboring atoms move 180° out-of-phase for the optical branch while neighboring primitive cells (with atoms m and M as the basis) move 180° out-of-phase for the acoustic branch when k coincides with the wave vector $G_1/2$ demarking the edge of the FBZ.
- 6.35** Find the phonon density of states for 1-D longitudinal propagation along the z -axis for wavelengths up to the edges of the FBZ.
- 6.36** Repeat Problem 6.35 for propagation vectors having components k_x and k_y .
- 6.37** Consider $P(n) = \left(1 - e^{-\frac{\hbar\omega_k}{k_b T}}\right) e^{-\frac{n\hbar\omega_k}{k_b T}}$
- Plot $P(n)$ versus n for $\hbar\omega = kT$, and for a couple of cases for different c in $c\hbar\omega = kT$.
 - Plot $\bar{n}$ and σ_n^2 versus temperature.
 - Find an equation for the skew S defined by $S^3 = \langle (n - \bar{n})^3 \rangle$. Reduce your answer as much as possible to terms involving $\bar{n}$, σ_n^2 .
- 6.38** For the phonon Bose–Einstein distribution, the average number of quanta (phonon) per oscillator in the mode ω_k can be written as

$$\langle n \rangle = \sum_n n P(n) = \left(1 - e^{-\frac{\hbar\omega_k}{k_b T}}\right) \sum_n n e^{-\frac{n\hbar\omega_k}{k_b T}}$$

Plot the average versus kT for several values of oscillator energy.

- 6.39** Perform the following steps for the equation of the average given in problem 6.38.
- Show for n ranging from 0 to N and defining $y = 1 + 2x + \dots + Nx^{N-1}$ that the summation can be written as

$$z = x + 2x^2 + \dots + Nx^N = x \frac{dy}{dx} \xrightarrow{N \rightarrow \infty} z = \frac{x}{(1-x)^2}$$

- Show the relation $x \frac{dy}{dx} \xrightarrow{N \rightarrow \infty} z = \frac{x}{(1-x)^2}$ by showing and then using

$$y = \sum_{m=0}^N x^m = 1 + x(y - x^N)$$

- Finally show that

$$\langle n \rangle = \frac{1}{e^{\frac{\hbar\omega_k}{k_b T}} - 1}$$

- 6.40** For the phonon Bose–Einstein probability distribution, find $\langle n^2 \rangle$ without using generating functions. Define $x = e^{-\frac{\hbar\omega_k}{kT}}$ and $z = \sum_{n=0}^{\infty} n^2 x^n$. Then show

$$z = x \frac{d}{dx} x \frac{d}{dx} \sum_n x^n = x \frac{d}{dx} x \frac{d}{dx} \frac{1}{1-x}$$

- 6.41** For phonon Bose–Einstein statistics, show all steps for the following two results

$$\begin{aligned} \langle n \rangle &= \frac{\partial}{\partial t} \langle e^{nt} \rangle \Big|_{t=0} = \frac{\partial}{\partial t} \frac{1 - e^{-\frac{\hbar\omega_k}{kT}}}{1 - e^t e^{-\frac{\hbar\omega_k}{kT}}} \Big|_{t=0} = \frac{1}{e^{\frac{\hbar\omega_k}{kT}} - 1} \\ \langle n^2 \rangle &= \frac{\partial^2}{\partial t^2} \langle e^{nt} \rangle \Big|_{t=0} = \frac{e^{-\frac{\hbar\omega_k}{kT}}}{1 - e^{-\frac{\hbar\omega_k}{kT}}} + \frac{2e^{-\frac{2\hbar\omega_k}{kT}}}{\left(1 - e^{-\frac{\hbar\omega_k}{kT}}\right)^2} = \langle n \rangle + 2\langle n \rangle^2 \end{aligned}$$

6.42 Show the variance can be written as the sum of two terms

$$\sigma_n^2 = \langle (n - \bar{n})^2 \rangle = \langle n^2 \rangle - \bar{n}^2$$

6.43 Use the moment generating function to find a formula for the skew for the phonon Bose–Einstein distribution.

6.44 Show the high temperature limit of the specific heat from the Einstein model

$$C_v = \frac{\partial U}{\partial T} = 3Nk_b \left(\frac{\hbar\omega_k}{k_b T} \right)^2 e^{\frac{\hbar\omega_k}{k_b T}} \left(e^{\frac{\hbar\omega_k}{k_b T}} - 1 \right)^{-2}$$

agrees with that of the Dulong–Petit model.

6.45 Find the average amplitude of a mode (frequency ω) and temperature T .

6.46 For the Lagrange density given by

$$\mathcal{L} = \frac{1}{2} \rho \dot{\eta}^2 - \frac{Y}{2} \left(\frac{\partial \eta}{\partial z} \right)^2 - \frac{c}{2} \eta^2$$

- Find the equation of motion (i.e., wave equation).
- Find the canonical momentum density.
- Determine the Hamiltonian density.

6.47 For the Lagrange density given by

$$\mathcal{L} = \frac{1}{2} \rho \dot{\eta}^2 - \frac{Y_i}{2} (\partial_i \eta)^2 - \frac{c}{2} \eta^2$$

- Show $\sum_i Y_i \partial_i^2 \eta = \rho \ddot{\eta}$
- Find the momentum density
- Write the result for part a in terms of the Dyad $\vec{Y}$
- Find the Hamiltonian density

REFERENCES AND FURTHER READINGS

General References for Structure, Phonons, Waves, Reciprocal Lattice

1. Tiwari S., *Compound Semiconductor Device Physics*, Academic Press, New York (1992). This book has in-depth discussion of phonon effects.
2. Blakemore J.S., *Solid State Physics*, 2nd ed., W.B. Saunders Company, Philadelphia, PA (1974).
3. Ashcroft N.W. and Mermin N.D., *Solid State Physics*, Holt, Rinehart & Winston, New York (1976).
4. Kittel C., *Introduction to Solid State Physics*, 5th ed., John Wiley & Sons, New York (1976).
5. Kittel C., *Quantum Theory of Solids*, John Wiley & Sons, New York (1987).
6. Bhattacharya P., *Semiconductor Optoelectronic Devices*, 2nd ed., Prentice Hall, Upper Saddle River, NJ (1997). Good general reference on most aspects of solid state including fabrication, electronic processes, bands, junctions, and optoelectronic devices.
7. Rosenberg H.M., *The Solid State*, 2nd ed., Oxford University Press, New York (1984).
8. Davies J.H., *The Physics of Low Dimensional Semiconductors*, Cambridge University Press, Cambridge, U.K. (1998).
9. Yu P.Y. and Cardona M., *Fundamentals of Semiconductors: Physics and Materials Properties*, 2nd ed., Springer, Berlin (1999).
10. Datta S., *Quantum Phenomena, Volume VIII in the Modular Series on Solid State Devices*, R.F. Pierret and G.W. Neudeck, eds., Addison-Wesley Publishing Company, Reading, MA (1989).

11. Lundstrom M., *Fundamentals of Carrier Transport, Volume X in the Modular Series on Solid State Devices*, R.F. Pierret and G.W. Neudeck, eds., Addison-Wesley Publishing Company, Reading, MA (1990).
12. Johnson S.C. and Gutierrez T.D. Visualizing the phonon wave function, *Am. J. Phys.* 70(3) 227–237 (2002).
13. Klingshirn C.F., *Semiconductor Optics*, Springer, New York (1997).

Chemical Bonding

14. Coulson C.A., *Valence*, 2nd ed., Oxford University Press, London (1961). This is a classic with clear discussion.
15. Tang C.L., *Fundamentals of Quantum Mechanics for Solid State Electronics and Optics*, Cambridge University Press, Cambridge, U.K. (2005).

Classical Mechanics and Normal Coordinates

16. Marion J.B., *Classical Dynamics*, 2nd ed., Academic Press, New York (1970).
17. Goldstein R., *Classical Mechanics*, Addison-Wesley, Reading, MA (1950).

Statistical Mechanics

18. Reif F., *Statistical Physics, Berkeley Physics Course*, Vol. 5, McGraw-Hill Book Company, New York (1965).
19. Pathria R.K., *Statistical Mechanics, International Series in Natural Philosophy*, Vol. 45, Butterworth-Heinemann Ltd., Oxford. First printing 1972 and reprinted through 1995. This is one of the most readable treatments.
20. Datta S., *Quantum Transport: Atom to Transistor*, Cambridge University Press, Cambridge, U.K. (2005).

7 Solid-State: Conduction, States, and Bands

The operation of the vast majority of modern electronic components can only be described through the band theory. The crystalline and near-crystalline forms of matter produce bands as described by the models presented in this chapter including the Kronig–Penney, tight binding, and k–p models. The bands produce an effective mass for the electron and hole, which can be many orders of magnitude smaller than the mass of the free electron. The effective mass has very important consequences for electrical conduction and the high-frequency performance of many devices. Purely crystalline materials do not have states in the energy bandgap. However, defects and doping do produce localized states within the gap that tend to trap the electrons and holes in a specific region of space and energy.

One of the most exciting areas of research focuses on the theory, fabrication, and experiments on reduced dimensional structures. These structures can have fewer than a hundred atoms. Such small sizes induce quantum confinement effects in the systems that radically affect the band structure and most of the optoelectronic properties. Some of the earliest devices incorporated nanostructures by epitaxial growth whereby the composition of the material changes along the growth axis to form a quantum well. As a result, this chapter calculates not only the density of states for bulk semiconductors but also for reduced dimensional structures.

7.1 THE EQUATION OF CONTINUITY

The equation of continuity describes the rate of change of charge in a small volume in terms of the current and the charge generation and recombination. The details of the classical and quantum equations of continuity differ in the descriptions of the current and charge distribution. This section reviews the drift mobility and its relation to Ohm’s law. We review the classical equation of continuity in preparation for the quantum mechanical version that comes from the time-dependent Schrödinger wave equation. Elementary quantum theory defines the charges and currents in terms of the wave functions. Later sections apply the quantum currents to tunneling and electron-resonant devices.

The classical free electron model, related to the Drude model describing conduction in metals, treats charge transport using many of the tools and concepts from the kinetic theory of gases. In the Drude model, electrons collide with immobile ions, and other electrons and phonons; these collisions limit the electron mobility. Several approximations can be made. The independent electron approximation neglects the electron–electron scattering events. The free electron approximation neglects the electron–ion interactions. The Drude model assumes that the collisions occur instantaneously and that the velocity changes randomly. The electrons achieve thermal equilibrium with the lattice through the collisions.

7.1.1 CLASSICAL DC CONDUCTION

Elementary studies of electricity and magnetism define the current density $\vec{J}$ (Amps/area) in terms of the velocity of a charge carrier as

$$\vec{J} = \rho \vec{v} \quad (7.1a)$$

where ρ represents the charge per volume that has velocity $\vec{v}$. For homogeneous materials, one can link the charge due to an electric field $\vec{E}$ through Ohm's law as

$$\vec{J} = \sigma \vec{E} \quad (7.1b)$$

The conductivity σ is positive number so that the current density must be parallel to the field. That this holds for negative charge can be seen from Equation 7.1a since $\vec{v}$ will be antiparallel to $\vec{E}$ but the sign of the charge density then reverses the direction of $\rho\vec{v}$ to be parallel to $\vec{E}$. Recall that current is taken as the flow of positive charge. The conductivity σ and the resistivity $r = 1/\sigma$ represent material properties for the conduction of current. Recall, they describe “the ease” with which current flows in response to a given applied field. The quantities appearing in Equation 7.1 represent averages over the material. As another simple reminder, the usual form of Ohm's law relates the voltage $V = EL$ across the length L of the material to the current $I = JA$ where A represents the cross sectional area. Ohm's law can therefore be rewritten in the familiar form

$$\frac{I}{A} = \sigma \frac{V}{L} \rightarrow V = \frac{1}{\sigma} \frac{L}{A} I = RI \quad (7.2)$$

In principle, it is possible for the conductivity and mobility (see below) to be tensors whereby the motional response and the electric field are no longer parallel. In such a case, Ohm's law uses conductance as a second rank tensor. For example if the fields E_x, E_y along the x - and y - directions, respectively, produce current density J_x and J_y , which are not equal as shown in Figure 7.1, then the response $\vec{J}$ will not be parallel to the field $\vec{E}$. The figure then indicates $J_x = \sigma_x E_x$ and $J_y = \sigma_y E_y$ so that one can write

$$\begin{pmatrix} J_x \\ J_y \end{pmatrix} = \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \quad \text{or} \quad \vec{J} = \vec{\sigma} \cdot \vec{E} \quad (7.3)$$

where the last form makes use of the dyadic notation from Chapter 3.

The resistivity r , conductivity σ , and the drift mobility μ represent more complicated physical phenomenon at the microscopic level. Ohm's law uses these constants to describe current flow without reference to the underlying physical processes. The drift mobility describes the average speed of the carrier in response to an impressed electric field. In a material, the carrier accelerates to an average speed controlled by the collisions with other particles within the material. This behavior is contrary to that in vacuum where the speed can continue to increase. The drift mobility $\mu \geq 0$ relates the average carrier speed (in a material for a particular type of carrier) to the driving field through the relation in Equation 7.4 below. However, for vector equations, one must include a minus sign so that the velocity of electrons or other negatively charged particles will have a direction opposite to that for the electric field. One can include the minus sign by defining the sign function as s_q which has the value -1 for negative charge and $+1$ for positive charge.

$$v = \mu E \quad \text{or} \quad \vec{v} = s_q \mu \vec{E} \quad (7.4)$$

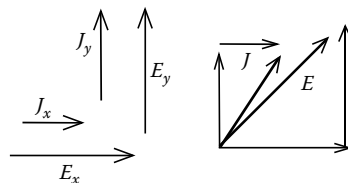


FIGURE 7.1 Example for nonparallel current density and electric field.

This last equation indicates that the electric field directly controls the average speed of the moving charge within the material. Collisions between the charge carriers and particles intrinsic to the material (such as atoms, phonons, and crystal defects) damp their motion. The drift mobility describes the average effect of many collisions experienced by the moving charges. The drift mobility is a scalar as defined in the first of Equations 7.4. Later sections will recast the mobility into a tensor form for the case when the motional response of the carriers to the electric is not parallel to the field. The value of the drift mobility differs for electrons and holes.

How can one relate the conductivity to the mobility? For multiple species of charge carriers denoted by i with density ρ_i (charge per volume), and velocity $\vec{v}_i$, the current density can be written as

$$\vec{J} = \sum_i \rho_i \vec{v}_i = \sum_i q_i \eta_i \vec{v}_i \quad (7.5a)$$

where η_i represents the number of carriers with charge q_i per volume. Assuming the charge carriers can only be electrons and holes with number density n_i and p_i (number per volume), respectively, then this last relation becomes

$$\vec{J} = q_n n \vec{v}_n + q_p p \vec{v}_p \quad (7.5b)$$

where

$q = -e$ for electrons ($e > 0$)

$\vec{v}_n, \vec{v}_p$ represent the speed of the corresponding carrier

Therefore the current density can be written in terms of the drift mobility according to

$$\vec{J} = q_n n \vec{v}_n + q_p p \vec{v}_p = q_n n (s_n \mu_n \vec{E}) + q_p p (s_p \mu_p \vec{E}) = (en\mu_n + ep\mu_p) \vec{E} \quad (7.6)$$

Comparing Equations 7.1b and 7.6, one finds an expression for conductivity

$$\sigma = e\mu_n n + e\mu_p p \quad (7.7)$$

where both terms have positive values. The discussion centers on the motion of electrons until having developed the concept of holes in semiconductors. The ability of a wire to conduct current depends on the number of free electrons n and the mobility μ_n .

Often people imagine the hole to be a particle with positive charge which proves useful for practical applications. However, the hole represents a missing electron in an interatomic bond. As such, the hole does not have a charge. The unbalanced positive charge appears on nearby nuclear cores; consequently, the charge of the hole is not localized to the empty state although the hole behaves as a positive charge. One might imagine that conduction of holes can occur when an electron tunnels from a nearby filled state (bond) into the empty state. The transition rate exponentially decreases with the separation distance. One does not expect the mobility of holes to be the same as the mobility of electrons in the conduction band.

7.1.2 COLLISIONS AND DRIFT MOBILITY

This section briefly demonstrates how collisions determine the drift mobility of a particle (Figure 7.2). A moving electron can collide with ions, phonons, and other electrons. All of the various types of collisions affect the drift mobility. Adjusting the crystal temperature can control the number of phonon collisions. As the temperature of a material increases and the lattice atoms

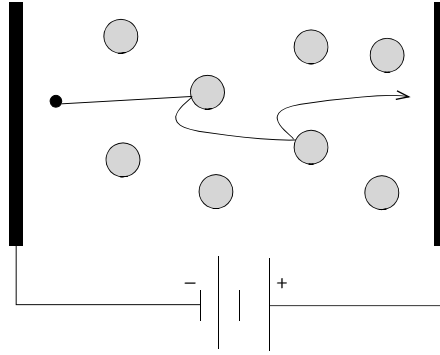


FIGURE 7.2 The scattering of an electron.

experience increased motion, the number of phonons must increase, the number of electron–phonon collisions increases and therefore the drift mobility must decrease. Decreasing the temperature reduces the phonon collisions but does not affect collisions with the lattice defects. Therefore, we expect the mobility to increase to a limit as the temperature decreases. These effects become more evident once we adopt a model (Figure 7.2).

Assume an electron moves through a region of space and collides with a number of obstacles. We apply Newton's second law $\vec{F} = m\vec{a} = \dot{\vec{p}}$ where p represents the momentum of the electron. Assume $t=0$ refers to the electron immediately after a first collision while $t=\tau$ refers to the electron just prior to the second collision. We take the time τ to be the average time between collisions. Newton's second law can be integrated to provide

$$\vec{p}(t) = \vec{p}(0) + \int_0^t d\tau \vec{F} = \vec{p}(0) + \int_0^t d\tau q\vec{E} = \vec{p}(0) + q\vec{E}t \quad (7.8)$$

where E represents the electric field (assumed independent of time). We assume that the electron moves in a random direction just after the last collision (on average). The momentum $\vec{p}(0)$ therefore averages to zero. The momentum at the time of the next collision obtains from Equation 7.8 by taking the average yields $\langle \vec{p}(t) \rangle = qE\langle t \rangle$. Using $\vec{p} = m\vec{v}$ provides the average velocity at the average time t of impact (note the symbols p , v , t now represent an average)

$$\vec{v} = \frac{q\tau \vec{E}}{m} \quad (7.9)$$

Sometimes one can find the same results but implicitly treating the medium as a viscous fluid with the averages already incorporated into the equations of motion. For such a case, one can write

$$\dot{\vec{p}} = q\vec{E} - \vec{p}/\tau \quad (7.10)$$

where the last term p/τ represents the damping term. At steady state, $\dot{p} = 0$ and then one obtains Equation 7.9 again.

Comparing Equations 7.4 and 7.9, we identify the drift mobility $\mu = v/E$ as

$$\mu = \frac{q\tau}{m} \quad (7.11)$$

and the electron conductivity as

$$\sigma = \frac{q^2 n \tau}{m} \quad (7.12)$$

with m being the mass of the electron. We see that the mobility depends on the time between collisions. One can calculate the relation between the density of defects and phonons to find the mobility.

Having discussed the origin of Ohm's law and drift mobility, consider charge flowing within an electronic circuit. Under transient conditions, current can flow into a region of space (or an electrical junction or capacitor . . .) and increase the charge there. Under steady-state conditions, the charge in the region must remain constant; consequently, any current flowing into the region must be balanced by an equal amount leaving at the same time. Similarly, for any point in space (rather than a region), the current flowing into the point must be balanced by the current flowing out of it under steady-state conditions (in the absence of charge generation and recombination). You can probably recognize this as the conservation of current usually employed to analyze elementary electrical circuits. The equation of continuity reviewed in the next section provides a general description of the process.

7.1.3 CLASSICAL EQUATION OF CONTINUITY

Recall from elementary electromagnetic studies, several mechanisms can change the amount of charge Q in a volume V of space. First consider the process of current flow. Assume the charge Q consists of electrons. Denote the number of electrons per unit volume by n and the charge density by $\rho = -en$ (Coulombs per volume). Figure 7.3 suggests the charge Q must decrease as the current flows through the sides of the volume V . The total current flow out of the volume reduces the enclosed charge and provides

$$\frac{dQ}{dt} + I = 0 \quad (7.13)$$

This last equation provides one form of the equation of continuity.

Other processes can change the amount of charge within a region. One might imagine, for example, a generation process that creates a single type of charge q at some rate G (charge/second). Likewise we might imagine a "sink" of charge that destroys a single type of charge at a rate R (charge/second). The rate equations in Equation 7.13 must become

$$\frac{dQ}{dt} + I = G - R \quad (7.14)$$

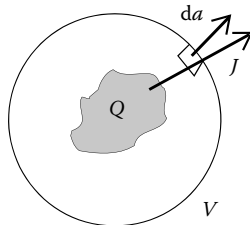


FIGURE 7.3 The current density J through the surface reduces the charge Q stored inside.

One can easily see this last equation must be true by assuming $I = 0$ and then realizing the difference $G - R$ must increase the amount of charge inside the volume. In semiconductors, we usually create charge pairs and not charge of a single sign. For example, charge can be generated in a semiconductor when it absorbs light. However, the process creates equal numbers of positively charged holes and negatively charged electrons. For this reason, the term “generation” more accurately applies to electron-hole pairs since the total charge adds to zero. Recombination refers to the process whereby an electron loses sufficient energy for it to return to the valence band. Recombination reduces the number of electron-hole pairs. However, even for pair creation, one can still focus on a single type of charge so that Equation 7.14 holds. In this section, we neglect generation–recombination events and instead focus on changing the amount of charge in a region of space by conduction processes.

The equation of continuity in Equation 7.14 with $G - R = 0$ can also be written as an integral relation. The total charge within the volume can be rewritten in terms of the electron density ρ

$$Q = \int_V dV \rho \quad (7.15)$$

Likewise, the total current flow through the surface of the volume V can also be written in terms of the current density $\vec{J}$. We think of current density as the number of amperes flowing through each unit area of a surface (i.e., units of Amps/area). The total current is then found by summing all of the small currents $\vec{J} \cdot d\vec{a}$ at the surface of V —notice the dot product. The total current through the surface V is then

$$I = \int_A d\vec{a} \cdot \vec{J} \quad (7.16)$$

where A is the bounding surface. The integral form of the equation of continuity can be found by combining the last three equations into

$$\int_V dV \frac{d\rho}{dt} = - \int_A d\vec{a} \cdot \vec{J} \quad (7.17)$$

One can also write the equation of continuity in the most familiar differential form. Recall the divergence theorem

$$\int_V dV \nabla \cdot \vec{J} = \int_A d\vec{a} \cdot \vec{J} \quad (7.18)$$

This equation basically says that current diverges away from a volume when it flows through the surface. Combining the last two equations provides

$$\int_V dV \left[\nabla \cdot \vec{J} + \frac{d\rho}{dt} \right] = 0 \quad (7.19)$$

Assuming that Equation 7.19 holds regardless of the integration region, we arrive at the differential form of the equation of continuity.

$$\nabla \cdot \vec{J} + \frac{d\rho}{dt} = 0 \quad (7.20a)$$

Letting $\mathcal{G}$ and $\mathcal{R}$ be the generation and recombination terms with units of charge per second per volume

$$\nabla \cdot \vec{J} + \frac{d\rho}{dt} = \mathcal{G} - \mathcal{R} \quad (7.20b)$$

The equation of continuity provides the basis for statements such as “charge” is conserved. It is not destroyed but moved from one place to another (in the absence of generation and recombination). Similar equations of continuity can be found for energy except energy cannot be created or destroy but can be transformed from one form to another (such as from energy to mass). The “heat” can take the place of current and then “internal energy” takes the place of the charge density. The equation of continuity is also the reason that Kirkoff’s current laws for electrical circuits works.

Example 7.1

Suppose $\vec{J} = 2x\hat{x}$ which means that the current grows with x . The rate of change of the charge density at $x = 1$ is then

$$\frac{d\rho}{dt} = -\frac{dJ}{dx} = -2 \rightarrow \rho(t) = \rho(0) - 2t$$

The only reason that the current increases with x is that the stored charge at x is being reduced and flowing toward positive x .

7.1.4 EQUATION OF CONTINUITY FOR QUANTUM PARTICLES

We now develop an equation of continuity for the charged quantum particle. The process identifies an expression for the charge and current density (in first quantization). In quantum theory, one thinks of particles as waves. If these waves move along the x -axis (for example), the charged particles move as well; consequently, current must flow. We also discuss how the normalization of a wave might be modified to include a number of electrons in nearly identical states in order to discuss current consisting of more than one charged particle. We must be careful not to violate the Pauli exclusion principle that allows only one electron per quantum state.

Suppose $\psi(\vec{r}, t)$ satisfies Schrödinger’s wave equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\vec{r}, t) + V\psi(\vec{r}, t) = i\hbar\frac{\partial}{\partial t}\psi(\vec{r}, t) \quad (7.21)$$

Notice that if we take the complex conjugate of Equation 7.21 then ψ^* must satisfy

$$-\frac{\hbar^2}{2m}\nabla^2\psi^* + V\psi^* = -i\hbar\frac{\partial}{\partial t}\psi^* \quad (7.22)$$

Next multiply Equation 7.21 by ψ^* and Equation 7.22 by ψ , and then subtract the two resulting equations to find

$$-\frac{\hbar^2}{2m}[\psi^*\nabla^2\psi - \psi\nabla^2\psi^*] = i\hbar\left[\psi^*\frac{\partial}{\partial t}\psi + \psi\frac{\partial}{\partial t}\psi^*\right] \quad (7.23)$$

This last result comes from observing that $\psi V \psi^* - \psi^* V \psi = 0$ because V is not an operator in the coordinate representation. Now using properties of differential operators, namely

$$\begin{aligned}\frac{\partial}{\partial t}(\psi^* \psi) &= \frac{\partial \psi^*}{\partial t} \psi + \frac{\partial \psi}{\partial t} \psi^* \\ \nabla \cdot (\psi^* \nabla \psi) &= \nabla \psi^* \cdot \nabla \psi + \psi^* \nabla^2 \psi\end{aligned}$$

Equation 7.23 can be written as

$$-\frac{i\hbar}{2m} \nabla \cdot [\psi^* \nabla \psi - \psi \nabla \psi^*] = -\frac{\partial}{\partial t} [\psi^* \psi] \quad (7.24)$$

The quantity $\psi^* \psi$ represents a probability density. The electron can be imagined as smeared out like a cloud (at least in a statistical sense) even though we always detect it in one spot. If we think of $e\psi^* \psi$ as being similar to a charge density (i.e., charge per unit volume), then Equation 7.24 starts to look more like an equation of continuity. Multiplying Equation 7.24 by q provides

$$-\frac{qi\hbar}{2m} \nabla \cdot [\psi^* \nabla \psi - \psi \nabla \psi^*] = -\frac{\partial}{\partial t} [q\psi^* \psi] \quad (7.25)$$

where $q = +e$ for a hole and $q = -e$ for an electron. The current density can be identified as

$$\vec{J} = -\frac{qi\hbar}{2m} [\psi^* \nabla \psi - \psi \nabla \psi^*] \quad (7.26)$$

(sometimes called a “probability current density”) and the charge density as

$$\rho_q = q\psi^* \psi \quad (7.27)$$

Equation 7.25 then becomes the equation of continuity for the quantum particle.

$$\nabla \cdot \vec{J} + \frac{\partial}{\partial t} \rho_q = 0 \quad (7.28)$$

Example 7.2

Represent an electron by a plane wave

$$\psi = \sqrt{\frac{1}{V}} e^{ikx - i\omega t} \quad (7.29a)$$

The charge density can be calculated as

$$\rho_q = q\psi^* \psi = \frac{q}{V} \quad (7.29b)$$

The normalization of the plane wave in Equation 7.29a, namely $\sqrt{1/V}$, provides one charged particle in volume V according to the result q/V in Equation 7.29b.

Sometimes people represent a collection of interacting particles by the same wave function. Strictly speaking, because Pauli’s exclusion principle does not allow two or more electrons in

exactly the same state, each electron must be represented by slightly different wave functions. For example, electrons in neighboring quantum wells with overlapping wave functions (interacting) give rise to energy levels just slightly separated and eigenfunctions just slightly different. To some order of approximation we might consider N electrons to be in approximately the same translational state. We might consider normalizing the wave function according to

$$\psi = \sqrt{\frac{N}{V}} e^{ikx - i\omega t}$$

Now notice that the wave functions can no longer normalize to 1. The charge density becomes $\rho_q = \psi^* \psi = N/V$. As an alternative procedure, if we have N particles with charge q , then we can just make the replacement $q \rightarrow Nq$ in the formulas.

Example 7.3

For the same plane wave in Example 7.2 with real k , find the current density.

SOLUTION

The current density must be given by

$$\vec{J} = -\frac{qi\hbar}{2m} [\psi^* \nabla \psi - \psi \nabla \psi^*] = \frac{q\hbar k}{mV} \tilde{x}$$

If we make use of the classical expression for momentum $p = mv$ and combine it with $p = \hbar k$, we find the familiar form of the current density

$$J = \frac{q}{V} v = \rho_q v$$

where ρ_q is the charge density.

Example 7.4

Suppose a wave function has the form (ignoring the correct normalization) of $\psi(x, t) = ae^{ikx - i\omega t}$ where $k = k_r + ik_i$, k_r and k_i are real, and $i = \sqrt{-1}$. Find the current density.

SOLUTION

In this case, one must be careful to correctly calculate the complex conjugate of the exponential. A Taylor expansion shows

$$(e^{ikx})^* = \left(\sum_n \frac{(ik)^n}{n!} x^n \right)^* = \sum_n \frac{(-ik^*)^n}{n!} x^n = e^{-ik^* x}$$

One must remember to complex conjugate the k . Now the current density can be calculated according to

$$\vec{J} = -\frac{qi\hbar}{2m} [\psi^* \nabla \psi - \psi \nabla \psi^*] = -\frac{qi\hbar}{mV} |a|^2 ik_r e^{-2k_i x} \tilde{x} = \frac{q\hbar k_r}{mV} |a|^2 e^{-2k_i x} \tilde{x}$$

Notice the complex k produces an exponentially decaying current density. The imaginary part of the wave vector controls the rate of decay along the x -direction. Further, if $k_i = 0$ then the result

reduces to that in the previous example. For the cases of step potentials in the next couple of sections, the wave vector will be either entirely real or entirely imaginary. Based on this last equation, the current density must be zero for entirely imaginary wave vectors. Such a case corresponds to an electron for example, penetrating a classically forbidden region (quantum tunneling) where the energy of the particle is smaller than the potential energy in that region.

7.2 SCATTERING MATRICES

Chapter 6 solve Schrödinger's wave equation for bound particles. Separating variables leads to an eigenvector equation for the Hamiltonian that constitutes the Sturm–Liouville problem found in studies of boundary value problems and partial differential equations. The boundary conditions for the Sturm–Liouville problem provide the quantization and the basis functions (the eigenfunctions). A bound particle or one with periodic boundary conditions has discrete energy levels since the wavelength can only take on certain values. We picture the eigenstates as standing waves (i.e., wave functions without time dependence). A superposition of the basis functions provides a solution to the Schrödinger wave equation. We know the probability of finding a quantum particle in a given eigenstate comes from the expansion coefficients for the superposition.

Some systems consist of free particles (i.e., unbounded). We again separate variables in the Schrödinger equation and focus on the Sturm–Liouville equation. Again, the boundary conditions lead to the allowed energies and eigenvectors. However, the boundary conditions lead to a continuous range of allowed energy and to a continuous basis set. One basis set consists of plane waves. An electron in one of the basis states can be pictured as moving and the particle is free from restraints. The superposition wave function must be an integral over these plane waves (i.e., a Fourier integral).

In preparation for band theory, we examine linear systems theory for electronic devices with quantum sizes (typically about 150 Å and smaller). We have primary interest in the theory of reflection and transmission through multiple electronic devices. This involves the nonclassical behavior of the electron described by its wave function. The reflected or transmitted amplitudes and their phase can easily be calculated using the so-called scattering and transfer matrices. The formalism resembles that used for optical and microwave systems.

Section 7.2.1 discusses the scattering theory in general terms as motivation for the remainder of the section. Section 7.2.2 shows how the simplest scattering matrix can be found based on an example with boundary conditions. Although the scattering-matrix equation relates the amplitude of an output beam to the amplitude of an input beam, it is not the most convenient representation of the system with multiple elements. The transfer matrix is better suited for stacked elements and can be found from the scattering matrix. The transfer matrices just need to be multiplied together for the more complex cases.

7.2.1 INTRODUCTION TO SCATTERING THEORY

For device work, one typically wants to know the current transmitted through a device as a function of various applied voltages. The circuits and systems can use these signals in a variety of ways such as for feedback or to drive other devices. As for most propagating waves, we need to find the reflected and transmitted waves. For quantum particles, the “waves” refer to the wave functions. Figure 7.4 shows a semiconductor material (GaAs) with a built-in barrier layer (AlGaAs). You can imagine a stack of multiple layers with electrodes on the left- and right-hand sides that can be used to modulate the barrier and well potentials. It would be helpful to have a simple matrix that predicts the amount of current (i.e., number of carriers) transmitted through the barrier. We need the amplitude and phase of the reflected and transmitted wave function. For now, we do not explicitly distinguish between propagation through a vacuum and a semiconductor.

Consider the example shown in Figure 7.4. The incident beam strikes the barrier at point *a*. Some of the electron wave reflects into beam 2 while some of the wave propagates to point *b*. The wave

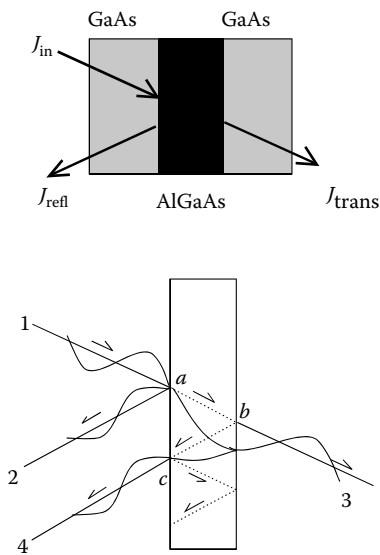


FIGURE 7.4 Wave picture of reflected and transmitted beams.

again divides into a reflected and transmitted piece. The wave reflected from point b travels to point c where another reflection can occur. If the total distance the wave travels from a to b to c can be made to $\lambda/2$ (i.e., barrier thickness of $\lambda/4$) and assuming a phase reversal at one of the interfaces, then beams 2 and 4 will be in phase and constructively interfere to produce a “bright spot.” In this case, the quarter-wave barrier might function as a fairly good electron mirror. Changing the thickness of the barrier should also make it possible to reduce the reflected component to below that for a single interface (similar to an antireflective coating on optical components). The output beams can constructively and destructively interfere. By controlling either the speed of the incident particle or by adjusting the barrier properties, we can use the barrier as either an energy filter or as an electronic switch.

The multiple reflected and transmitted beams in Figure 7.4 can be schematically represented as in Figure 7.5. To find the total reflected amplitude b_1 and transmitted amplitude b_2 we must add

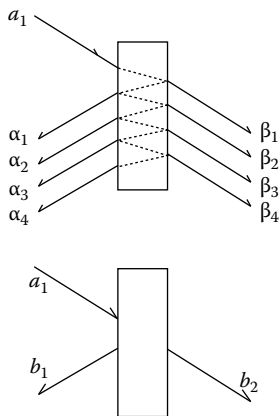


FIGURE 7.5 The reflected α_i and transmitted β_i amplitudes add together to produce b_1 and b_2 , respectively.

up all of the individual amplitudes α_i of the reflected beams and the amplitudes β_i of the transmitted beams.

$$b_1 = \sum_i \alpha_i \quad b_2 = \sum_i \beta_i$$

In general, the amplitudes must include phase information. The phase can be affected by the thickness of the element, as well as the reflection and transmission coefficients at an interface.

It should be clear that the two output amplitudes b_1, b_2 must be linearly related to the input amplitude a_1 assuming the region does not incorporate a nonlinear mechanism. We can write

$$\begin{aligned} b_1 &= S_{11}a_1 \\ b_2 &= S_{21}a_1 \end{aligned}$$

The scattering matrix consists of the set of numbers S_{ij} . The scattering matrix describes the particular electronic element.

The situation can be generalized by including two input beams as illustrated in Figure 7.6. The two outputs must be linearly related to the two inputs according to

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \quad (7.30a)$$

or, in matrix notation

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \underline{S} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad (7.30b)$$

Having developed the general notion of the scattering matrix, we now must specify the amplitudes using the quantum mechanical current density.

7.2.2 AMPLITUDES

The simplest prescription defines the amplitudes (including phase) in terms of traveling waves. The overall normalization of the wave functions will not be an issue as the primary interest will be in the reflectivity and transmissivity of the interface which comes from the ratio of amplitudes; consequently an overall normalization factor cancels out. The amplitude for an electron in the plane wave state

$$\psi(x, t) = a_0 e^{ikx - i\omega t} \quad (7.31a)$$

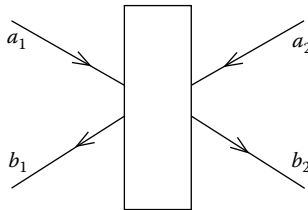


FIGURE 7.6 An electron beam strikes an electronic device from either side. The figure can be further generalized by having any number of beams from left or right.

can be written as

$$a = a_0 e^{ikx} \quad (7.31b)$$

This represents a plane wave moving along the positive x -axis. For an electron in a plane wave moving along the negative x -axis

$$\psi(x, t) = a'_0 e^{-ikx - i\omega t} \quad (7.32a)$$

then the amplitude can be written as

$$a' = a'_0 e^{-ikx} \quad (7.32b)$$

At other times, a person might be more interested in the current density $\vec{J}$ rather than the wave functions. We know the current density once we know all of the amplitudes according to

$$\vec{J} = -\frac{q\hbar}{2m} [\psi^* \nabla \psi - \psi \nabla \psi^*] \quad (7.33)$$

However, we should not apply this formula directly to the transmission or reflection of an electron at a barrier because we lose the phase information embodied in the wave function ψ . Example 7.2 in the previous section shows that the current density for one-dimensional (1-D) motion reduces to $J = \frac{q}{v} v = \rho_q v$. We do not see any phase (i.e., a complex exponential that depends on x) necessary to produce interference.

All coherent phenomena have similar behavior. For example, optical or microwave studies first must calculate the fields (i.e., wave functions). For any coherent phenomenon (such as a two slit experiment), these wave functions must be first added together and then one can calculate the power (refer to books on optics or lasers regarding the Poynting vector). In this case, we cannot first calculate the currents and then add them together since we lose the phase information and will not see any interference effect.

7.2.3 REFLECTIVITY AND TRANSMISSIVITY

In this section, consider the simplest case of an electron incident on a step potential as shown in Figure 7.7. For example, the electron might initially propagate through GaAs and then strike a GaAs–AlGaAs interface; however, we must include effective mass effects. We will find the reflected and transmitted components at the interface by solving Schrödinger's equation and then identifying and relating the amplitudes of the plane waves. From there we will set up the scattering matrix. Once we determine the reflection coefficients, there would not be any need to solve the problem again for arbitrary combinations of layers. This analysis holds for both the free electron

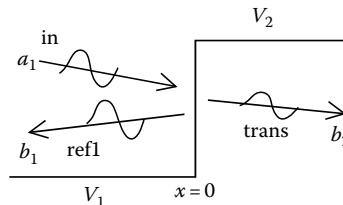


FIGURE 7.7 Incident wave is transmitted and reflected at interface.

model of conduction and the nearly free model (by identifying the effective mass and envelope wave functions).

The time-independent Schrödinger's equation for the step potential has the form

$$-\frac{\hbar^2}{2m_x} \frac{d^2}{dx^2} \psi(x) + V(x)\psi(x) = E\psi(x) \quad (7.34)$$

For now, consider the free-space case. However, later we will investigate heterostructure materials where m must be the effective mass of the particle (electron or hole) sometimes denoted by m^* (not to be confused with the complex conjugate). Depending on the type of material, the effective mass of an electron in the conduction band can differ by a factor of thousand or more from the electron rest mass. The subscript x on the mass denoted by m_x in Equation 7.34 indicates the possibility that the mass might depend on position x . For two layers that meet at $x=0$, one might have

$$m_x = \begin{cases} m_1 & x < 0 \\ m_2 & x > 0 \end{cases} \quad (7.35)$$

For GaAs–AlGaAs, we assume the difference is small enough to be neglected. However, Section 7.2.4 shows the necessary modifications should the difference in mass be considered significant.

Returning to the free-space problem, the step potential in Figure 7.7 can be written as

$$V(x) = \begin{cases} V_1 & x < 0 \\ V_2 & x > 0 \end{cases} \quad (7.36)$$

with the wave function defined according to

$$\psi(x) = \begin{cases} \phi(x) & x < 0 \\ \gamma(x) & x > 0 \end{cases} \quad (7.37)$$

Often for simplicity, one assumes $V_1=0$ for example, and $V_2=V_0$ for the barrier or step height.

Using complex notation, the solutions can be written as

$$\phi(x) = a_{01}e^{ik_1x} + b_{01}e^{-ik_1x} \quad x < 0 \quad (7.38a)$$

$$\gamma(x) = b_{02}e^{ik_2x} + a_{02}e^{-ik_2x} \quad x > 0 \quad (7.38b)$$

The solutions have complex exponentials that incorporate the $i = \sqrt{-1}$ but the wave vectors can still be complex. The imaginary part of the wave vector k will produce exponentially increasing or decreasing wave functions. Do not worry about the notation for the coefficients as it will become clear later. For now, just assume that a indicates a wave moving into the element and b indicates a wave moving away from it (see Figure 7.7). Notice the subscripts on the wave vectors. The wave vectors k_1 and k_2 can be found by substituting e^{ik_ix} into Schrödinger's wave equation to find

$$k_1 = \sqrt{\frac{2m}{\hbar^2}(E - V_1)} \quad \text{and} \quad k_2 = \sqrt{\frac{2m}{\hbar^2}(E - V_2)} \quad (7.39)$$

Notice k_1 and k_2 can be complex depending on the relation among E , V_1 , and V_2 .

We have boundary conditions that must be satisfied. We assume that the particle (and hence the initial wave) comes from the left. The barrier reflects and transmits some of the incident wave.

This means that a_{01} , b_{01} , and b_{02} cannot be zero. However, the term with a_{02} represents a wave that starts at $+\infty$ and travels toward the barrier; therefore, we take a_{02} to be zero. The other boundary conditions (BCs) concern the interface. We require the wave functions and the first derivatives to be continuous across the interface. (*Note:* for actual heterostructure, the BCs must include the effective mass for the materials) The boundary conditions are

$$\phi(0) = \gamma(0) \quad (7.40a)$$

$$\frac{d\phi(0)}{dx} = \frac{d\gamma(0)}{dx} \quad (7.40b)$$

The boundary conditions produce two equations with three unknowns. As always, one coefficient comes from the normalization conditions. We leave the coefficients in terms of a_{01} since it represents the incoming wave, which should be known.

$$\begin{aligned} a_{01} + b_{01} &= b_{02} \\ k_1 a_{01} - k_1 b_{01} &= k_2 b_{02} \end{aligned} \quad (7.41a)$$

Solving the system of equations provides

$$b_{01} = \frac{k_1 - k_2}{k_1 + k_2} a_{01} \quad b_{02} = \frac{2k_1}{k_1 + k_2} a_{01} \quad (7.41b)$$

Now we can identify the plane waves and the amplitudes. The original incoming plane wave moves from $-\infty$ to $x=0$ according to

$$a_1(x) = a_{01} e^{ik_1 x} \quad (7.42)$$

where we choose the normalization a_{01} to suit our needs. The reflected wave moves away from the electronic barrier toward $-\infty$

$$b_1 = b_{01} e^{-ik_1 x} = \frac{k_1 - k_2}{k_1 + k_2} a_{01} e^{-ik_1 x} \quad (7.43)$$

and the transmitted wave moves away from the barrier toward $+\infty$ as

$$b_2 = b_{02} e^{ik_2 x} = \frac{2k_1}{k_1 + k_2} a_{01} e^{-ik_2 x} \quad (7.44)$$

Keep in mind that a_1 and a_2 indicate waves moving toward the barrier while b_1 and b_2 indicate waves moving away from the barrier.

At this point we can identify reflectivity and transmissivity from Figure 7.8. The reflectivity r is the ratio of the amplitude of the incident to reflected wave. It can therefore be determined by comparing its definition in

$$b_1(0) = r a_1(0) \quad \text{or} \quad b_{01} = r a_{01} \quad (7.45)$$

with the results given in Equations 7.43 and 7.44 to find

$$r = \frac{k_1 - k_2}{k_1 + k_2} \quad (7.46)$$

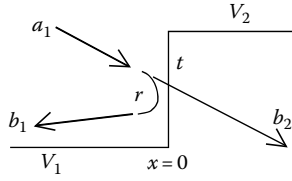


FIGURE 7.8 The reflectivity and transmissivity.

Notice that a wave traveling from $+\infty$ toward the barrier would experience a reflection of

$$r' = \frac{k_2 - k_1}{k_1 + k_2} = -r \quad (7.47)$$

Comparing Equations 7.46 and 7.47 shows the subscripts on k have been interchanged. Similarly, the transmissivity indicates how much of the wave passes the barrier and can be found by comparing its definition in

$$b_2(0) = t a_1(0) \quad \text{or} \quad b_{02} = t a_{01} \quad (7.48)$$

with Equation 7.44 to find

$$t_{1 \rightarrow 2} = \frac{2k_1}{k_1 + k_2} \quad (7.49)$$

The symbol $t_{1 \rightarrow 2}$ will be given the shortcut notation of t_{12} . Obviously, the transmissivity for a wave traveling from medium 2 to medium 1 (i.e., traveling from right to left across the barrier) would be

$$t_{2 \rightarrow 1} = \frac{2k_2}{k_1 + k_2} \quad (7.50)$$

Similarly at times, the symbol $t_{2 \rightarrow 1}$ will be represented by t_{21} .

Combining all of the previous work, we can now predict the outgoing waves in terms of the incident one. At this point, we need the coefficients but do not need the exponential factors (phase factors). In fact, in all of our future work, we want all of the effects to be included in the coefficients such as b_{01} which do not depend on the x -coordinate. Substituting $x=0$ for the position of the step provides

$$b_{01} = r a_{01} \quad b_{02} = t_{1 \rightarrow 2} a_{01} \quad (7.51)$$

or in matrix form

$$\begin{pmatrix} b_{01} \\ b_{02} \end{pmatrix} = \begin{pmatrix} r & 0 \\ t_{12} & 0 \end{pmatrix} \begin{pmatrix} a_{01} \\ 0 \end{pmatrix}$$

The 2×2 matrix is the scattering matrix.

As indicated in Figure 7.9, we can generalize the situation to include two waves incident on the barrier. One wave travels from $+\infty$ to the barrier while the other one travels from $-\infty$ to the barrier. These are incoming waves denoted by a . The principle of superposition can be used to relate the outgoing b waves to the incoming a waves according to

$$\begin{aligned} b_{01} &= r a_{01} + t_{2 \rightarrow 1} a_{02} = r a_{01} + t_{21} a_{02} \\ b_{02} &= t_{1 \rightarrow 2} a_{01} + r' a_{02} = t_{12} a_{01} - r a_{02} \end{aligned} \quad (7.52)$$

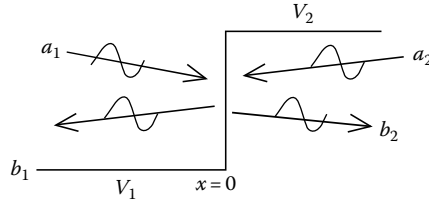


FIGURE 7.9 The output waves are due to two incoming waves.

In matrix notation, these equations can be written as

$$\begin{pmatrix} b_{01} \\ b_{02} \end{pmatrix} = \begin{pmatrix} r & t_{21} \\ t_{12} & -r \end{pmatrix} \begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix} \quad (7.53)$$

The scattering matrix for the simple interface is

$$\underline{S} = \begin{pmatrix} r & t_{21} \\ t_{12} & -r \end{pmatrix} \quad (7.54a)$$

where

$$r = \frac{k_1 - k_2}{k_1 + k_2} \quad t_{12} = \frac{2k_1}{k_1 + k_2} \quad t_{21} = \frac{2k_2}{k_1 + k_2} \quad (7.54b)$$

Finally, one can combine Equation 7.54b to find a relation useful for the transfer matrix

$$r^2 + t_{1 \rightarrow 2} t_{2 \rightarrow 1} = 1 \quad (7.54c)$$

7.2.4 MODIFICATIONS FOR HETEROSTRUCTURE

The previous analysis carried through for free space. Heterostructure material composed of multiple layers of different materials makes the effective mass of the charge carrier depend on position. The changes in potential V are produced by the differences in the bandgaps of the materials where they form an interface.

The boundary conditions account for the difference in mass through the first derivatives

$$\phi(0) = \gamma(0) \quad (7.55a)$$

$$\frac{1}{m_1} \frac{d\phi(0)}{dx} = \frac{1}{m_2} \frac{d\gamma(0)}{dx} \quad (7.55b)$$

where as before ϕ represents the wave for $x < 0$ and γ for $x > 0$ (see Figure 7.8 for example). As before, we leave the coefficients in terms of a_{01} since it represents the incoming wave, which should be known.

$$\begin{aligned} a_{01} + b_{01} &= b_{02} \\ \frac{k_1}{m_1} a_{01} - \frac{k_1}{m_1} b_{01} &= \frac{k_2}{m_2} b_{02} \end{aligned} \quad (7.56a)$$

Notice the second equation is similar to Equation 7.41a except k_1 and k_2 are replaced with k_1/m_1 and k_2/m_2 . Solving the simultaneous equations then produces similar results to 7.41b

$$b_{01} = \frac{k_1/m_1 - k_2/m_2}{k_1/m_1 + k_2/m_2} a_{01} \quad b_{02} = \frac{2k_1/m_1}{k_1/m_1 + k_2/m_2} a_{01} \quad (7.56b)$$

or equivalently,

$$r = \frac{k_1/m_1 - k_2/m_2}{k_1/m_1 + k_2/m_2} \quad t_{1 \rightarrow 2} = \frac{2k_1/m_1}{k_1/m_1 + k_2/m_2} \quad (7.56c)$$

7.2.5 REFLECTANCE AND TRANSMITTANCE

The reflectance and transmittance (as distinguished from reflectivity and transmissivity), constitute the primary quantities of interest for currents. The current density is calculated from the relation (see Section 7.1)

$$\vec{J} = -\frac{q\hbar}{2m} [\psi^* \nabla \psi - \psi \nabla \psi^*] \quad (7.57)$$

Recall from the examples at the end of the previous section, that a region where the energy of the particle is smaller than the potential energy $E < V$, then the complex wave vector produces an exponentially decreasing wave function. The portion of $\gamma(x)$ traveling from the step interface to $+\infty$, the transmitted part, produces current density at $x=0$ of

$$J_{\text{trans}} = \left[\frac{q\hbar \text{Re}(k_2)}{m} |b_{02}|^2 e^{-2k_{2i}x} \right]_{x=0} = \frac{q\hbar \text{Re}(k_2)}{m} |b_{02}|^2 \quad (7.58)$$

While the wave function can remain nonzero for $x > 0$, the current density will be nonzero there so long as $E > V_2$ in $k_2 = \sqrt{2m/\hbar^2(E - V_2)}$ because of the $\text{Re}(k_2)$ factor in Equation 7.58. For those cases when $E > V_2$, one can obviously replace $\text{Re}(k_2)$ by k_2 . Similarly, one can calculate the incident and reflected current density as

$$J_{\text{inc}} = \frac{q\hbar \text{Re}(k_1)}{m} |a_{01}|^2 \quad \text{and} \quad J_{\text{refl}} = \frac{q\hbar \text{Re}(k_1)}{m} |b_{01}|^2 \quad (7.59)$$

Equations 7.58 and 7.59 set $x=0$ for the position of the interface step. Values for $x > 0$ can be found using a simple “waveguide” approach as detailed in the next section. As previously mentioned for this simple step potential, the transmitted current density will be zero $J_{\text{trans}} = 0$ when $E < V_0$ and therefore, the incident charge will be totally reflected by the step. One might expect that a narrow barrier rather than the step barrier will allow charge to tunnel from one side to the other. This is a purely quantum mechanical effect since classically, the charge does not have sufficient energy to surmount even a narrow barrier. In most cases, one “shoots” a beam of electrons into a region where $E > V$ and the wave function has real k . These electrons can encounter a finitely wide barrier where some will be reflected and some will be transmitted through the barrier (quantum tunneling).

The reflectance defined as $R = J_{\text{refl}}/J_{\text{inc}}$ can be written as

$$R = \frac{J_{\text{refl}}}{J_{\text{inc}}} = \frac{b_{01}^*}{a_{01}^*} \frac{b_{01}}{a_{01}} = r^* r \quad (7.60a)$$

By substituting for r , this last relation can also be written as

$$R = \left| \frac{k_1 - k_2}{k_1 + k_2} \right|^2 \quad (7.60b)$$

Similarly, the transmittance $T = J_{\text{trans}}/J_{\text{inc}}$ can be determined from Equations 7.58 and 7.59

$$T = \frac{\text{Re}(k_2)}{k_1} \frac{4|k_1|^2}{|k_1 + k_2|^2} \quad (7.60c)$$

Assuming nonabsorbing interfaces, one can find a type of conservation equation as

$$J_{\text{inc}} = J_{\text{ref}} + J_{\text{trans}} \quad (7.61a)$$

Dividing through by J_{inc} provides

$$R + T = 1 \quad (7.61b)$$

7.2.6 CURRENT-DENSITY AMPLITUDES

The development of the reflection and transmission of a particle at an interface starts with the wave function and applies boundary conditions. However as a side issue, in some cases it might be convenient to normalize the wave functions in terms of current. That is, one can define amplitudes that have units of (current density)^{1/2} while retaining the phase information. Out of interest, it is worth seeing the procedure although it will provide only limited usage in this book. In particular, these amplitudes should not be used for heterostructure where the effective mass varies with position. The amplitudes j_0 explicitly display the phase information through the factor of $e^{\pm ikx}$

$$j_0 = c_0 e^{\pm ikx} \quad (7.62)$$

but still produce the current density according to $J = j_0^* j_0$. The amplitudes behave more like the wave functions and yet make it easy to calculate current. The amplitudes can be found starting with the plane wave (assuming k real)

$$\psi = c_0 e^{\pm ikx - i\omega t} \quad (7.63)$$

and substituting into the expression for current density found in the previous section.

$$J = -\frac{q\hbar}{2m} [\psi^* \nabla \psi - \psi \nabla \psi^*] = \frac{q\hbar k}{m} |c_0|^2$$

where k must be replaced by $\text{Re}(k)$ for complex k . Therefore, defining the amplitude j_0 by

$$j_0 = \sqrt{\frac{q\hbar k}{m}} |c_0|^2 e^{\pm ikx} \quad (7.64)$$

produces the current density

$$J = j_0^* j_0 = \frac{q\hbar k}{m} |c_0|^2 \quad (7.65)$$

If we normalize the wave function by setting $c_0 = \sqrt{N/V}$ in Equation 7.63 (rather than $\sqrt{1/V}$) similar to the end of Section 7.1 (N = number of electrons), the amplitude of interest becomes

$$j_0 = \sqrt{\frac{q\hbar k N}{mV}} e^{\pm ikx} = \sqrt{\rho_q v} e^{\pm ikx}$$

which essentially renormalizes the wave function to have a form similar to

$$\psi = \sqrt{\rho_q v} e^{\pm ikx - i\omega t}$$

Notice $\rho_q v$ agrees with the definition of current density found at the start of Section 7.1.

7.3 THE TRANSFER MATRIX

As demonstrated in the previous section, the scattering matrix for a simple step at $x = 0$ has the form

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} r & t_{21} \\ t_{12} & -r \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad \text{or} \quad \underline{b} = \underline{S} \underline{a} \quad (7.66)$$

where b_1 is used in place of b_{01} etc. since the previous section placed the interface at $x = 0$ so that $a_1 = a_{01} e^{ikx} |_{x=0} = a_{01}$. As will be shown, the “simple waveguide” element can be used to account for interfaces away from $x = 0$.

Of importance at the moment, Equation 7.66 multiplies the “physical inputs” (a_1, a_2) by the scattering matrix to determine the physical outputs (b_1, b_2). This section refers to “physical inputs” and “physical outputs” since respectively, these beams actually enter and leave the device. A “mathematical input” or “mathematical output” can be any combination of “physical inputs” and “physical outputs” that help solve the problem independent of their physical origins.

The arrangement of the physical input–output variables makes the scattering matrix inconvenient for solving more complicated problems. For example, consider Figure 7.10 showing an electronic device consisting of multiple elements labeled 1, 2, and 3. It would be possible to find the output parameters $\underline{b}$ if the input parameters $\underline{a}$ were known. However, the figure shows that the effect of the right-most (#3) and left-most (#1) elements must be known prior to being able to find the parameters $\underline{a}$. It is possible to write a matrix equation to include the other two elements (#1 and #3), but then there are three sets of simultaneous equations. A simpler method consists of looking at a figure (such as Figure 7.10) and writing a matrix for each element in the order that it occurs. These matrices would be multiplied in the same order as each electronic element occurs in the sequence. This is where the transfer matrix comes into play.

Figure 7.11 shows an expanded view of stacked electronic elements. The middle part of the figure separates the elements and labels the amplitudes of the input and output beams. The bottom portion shows how the transfer-matrix equation compares with each element. We want to use beams A_2 and B_2 as the input to optical element No. 1. We want to interpret the amplitudes A_1 and B_1 as the output from the optical element. In this way, the amplitudes A_2 and B_2 can be interpreted as the input

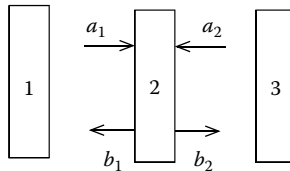


FIGURE 7.10 A multielement electronic device.

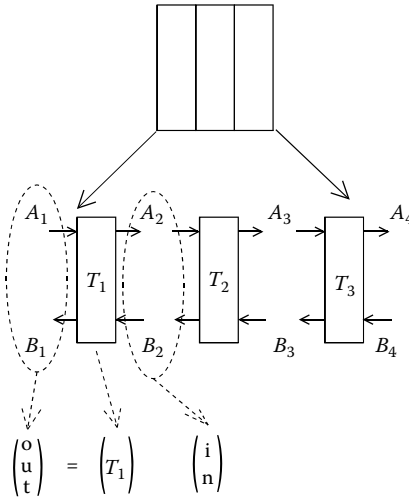


FIGURE 7.11 Stacked electronic elements. The transfer-matrix equation is shown for the first element.

to the first optical element and as the output from the second optical element. We picture the inputs to an element as residing on the right-hand side while the outputs from an element are on the left-hand side. In matrix notation, the first two elements produce an equation of the form

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = T_1 \begin{pmatrix} A_2 \\ B_2 \end{pmatrix} = T_1 T_2 \begin{pmatrix} A_3 \\ B_3 \end{pmatrix} \quad (7.67)$$

Clearly, a multielement device as in the figure can easily be represented by a series of multiplied transfer matrices. As with the scattering matrix S , the transfer matrix T represents the specifics of the multielement device and accounts for all coherent effects between components of the wave.

Now consider the important distinction between the transfer and scattering matrices. The right-hand side of a transfer-matrix equation (such as Equation 7.67) has a column vector containing “inputs” and the left-hand side has “outputs.” Consider the first element in Figure 7.11. Physically speaking, the A_1 amplitude represents an incident beam (i.e., an input beam) but it appears as an output variable in Equation 7.67. A similar comment applies to the amplitude A_2 appearing as an input variable even though that amplitude represents an output beam. Therefore, the mathematical inputs to the transfer-matrix equation must be different than the physical inputs to the scattering-matrix equation. For the transfer matrix, amplitudes on a given side of an electronic element (as indicated in Figure 7.11) have the same location in the equation. For the transfer-matrix equation, the mathematical input consists of a mixture of the physical inputs and outputs. For the scattering matrix, physical input amplitudes are placed together in a single-column vector.

For the transfer-matrix equation, how can physical output variables appear on the input side of the equation? The procedure works because the systems are linear in amplitudes. The variables in the scattering-matrix equation can be rearranged to give the variables for the transfer-matrix equation.

Scattering	Transfer	
$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$	$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} A_2 \\ B_2 \end{pmatrix}$	(7.68)
$b_1 = S_{11}a_1 + S_{12}a_2$	$A_1 = T_{11}A_2 + T_{12}B_2$	
$b_2 = S_{21}a_1 + S_{22}a_2$	$B_1 = T_{21}A_2 + T_{22}B_2$	

Consider the first electronic element in Figure 7.11. For the scattering matrix, we denote the input amplitudes by a_i and the output amplitudes by b_i . A comparison of Figure 7.11 and one similar to Figure 7.10 indicates the scattering and transfer variables must be related by $A_1 = a_1$, $A_2 = b_2$, $B_1 = b_1$, and $B_2 = a_2$ —one only needs to compare beam directions and their location with respect to the electronic element in question. A relation can be found between the scattering and transfer matrices. Start with the scattering-matrix equation

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \quad (7.69)$$

Next eliminate the scattering variables in favor of the transfer variables.

$$\begin{aligned} B_1 &= S_{11}A_1 + S_{12}B_2 \\ A_2 &= S_{21}A_1 + S_{22}B_2 \end{aligned} \quad (7.70)$$

Equation 7.70 must be compared with the defining relation for the transfer matrix in Equation 7.68. Equation 7.70 needs to be rearranged. Move A_2 and B_2 to the right-hand side and A_1 and B_1 to the left-hand side of the equation. The coefficients of A_2 and B_2 will be the elements of the transfer matrix. We find from the second of Equation 7.70

$$A_1 = \frac{1}{S_{21}}A_2 - \frac{S_{22}}{S_{21}}B_2$$

Substituting into the first of Equation 7.70 provides

$$B_1 = S_{11}A_1 + S_{12}B_2 = \frac{S_{11}}{S_{21}}A_2 - \frac{S_{11}S_{22} - S_{12}S_{21}}{S_{21}}B_2$$

The right-hand sides of the previous two lines provide the elements of the transfer matrix.

$$\underline{T} = \frac{1}{S_{21}} \begin{pmatrix} 1 & -S_{22} \\ S_{11} & -\det(\underline{S}) \end{pmatrix} \quad (7.71)$$

where “det” stands for the determinant. We could just as easily demonstrate the scattering matrix in terms of the transfer matrix

$$\underline{S} = \frac{1}{T_{11}} \begin{pmatrix} T_{21} & \det \underline{T} \\ 1 & -T_{12} \end{pmatrix} \quad (7.72)$$

Note that $\underline{T}$ refers to the transfer matrix and not the transmittance—two very different objects!

7.3.1 SIMPLE INTERFACE

Consider the interface between two media as shown in Figure 7.12. Assume that $E > V_0$. The previous section gives the scattering matrix as

$$\underline{S} = \begin{pmatrix} r & t_{2 \rightarrow 1} \\ t_{1 \rightarrow 2} & -r \end{pmatrix} \quad (7.73)$$

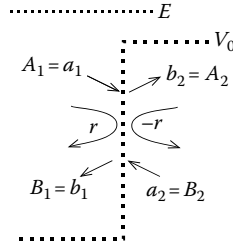


FIGURE 7.12 The simple interface. Side #1 has $x < 0$ and side #2 has $x > 0$.

Therefore the transfer matrix in

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = T \begin{pmatrix} A_2 \\ B_2 \end{pmatrix} \quad T = \frac{1}{S_{21}} \begin{pmatrix} 1 & -S_{22} \\ S_{11} & -\det(S) \end{pmatrix}$$

must be given by

$$T = \frac{1}{t_{12}} \begin{pmatrix} 1 & r \\ r & r^2 + t_{12}t_{21} \end{pmatrix} = \frac{1}{t_{12}} \begin{pmatrix} 1 & r \\ r & r^2 + t^2 \end{pmatrix} \quad (7.74)$$

where we define $t^2 = t_{12}t_{21}$. Recall that $r^2 + t^2 = 1$ from the previous section. Therefore the transfer matrix becomes

$$T = \frac{1}{t_{12}} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \quad (7.75)$$

Two important notes: (1) if $-r$ and r are interchanged in the figure, they would also be interchanged in the scattering and transfer matrices (i.e., r becomes $-r$); (2) if the minus signs appear as shown in Figure 7.12 but A_2 and B_2 become the output variables for the transfer matrix, then $-r$ and r must be interchanged in the transfer matrix.

The single interface provides a first example for the transfer matrix. The next example considers a particle propagating along a waveguide (along the z -direction) without any real interfaces. We will find the input and output differ by only a phase factor e^{ikz} .

7.3.2 SIMPLE ELECTRONIC WAVEGUIDE

Consider two waves (propagating along the horizontal z -direction) with amplitudes a_1 and a_2 incident on the left-hand and right-hand boundaries inside a chunk of material (Figure 7.13).

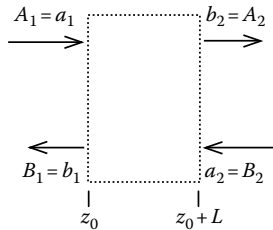


FIGURE 7.13 Block diagram for the simple waveguide.

We assume waves do not reflect from any of these internal virtual interfaces since they do not demark any separation between dissimilar materials and they do not represent any type of boundary between potential energies. We further assume that the electron beams propagate straight through the material. The forward propagating wave (from left to right) has the form $a_1 \sim a_{01} \exp(ikz)$ while the backward propagating wave has the form $a_2 \sim a_{02} \exp(-ikz)$. Notice the same wave vector k appears in both of these formulas. The amplitude b_2 at $z_0 + L$ must be related to the amplitude a_1 at z_0 by a phase factor (note $a_1 = a_1(z)$ is a function of z and similar for b_1, a_2, b_2)

$$b_2 = a_1(z_0 + L) = a_{01} \exp[ik(z_0 + L)] = a_1 \exp(ikL)$$

The backward propagating wave with amplitude b_1 at z_0 must be related to the wave with amplitude a_2 at $z_0 + L$.

$$a_2 = b_1(z_0 + L) = b_{01} \exp[-ik(z_0 + L)] = b_1 \exp(-ikL)$$

or, in other words,

$$b_1 = a_2 \exp(ikL)$$

The scattering-matrix equation is therefore

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 & \exp(ikL) \\ \exp(ikL) & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

Therefore, the transfer matrix T in the equation

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = T \begin{pmatrix} A_2 \\ B_2 \end{pmatrix}$$

must be given by

$$T = \frac{1}{S_{21}} \begin{pmatrix} 1 & -S_{22} \\ S_{11} & -\text{Det}(S) \end{pmatrix} = \frac{1}{\exp(ikL)} \begin{pmatrix} 1 & 0 \\ 0 & \exp(2ikL) \end{pmatrix} = \begin{pmatrix} \exp(-ikL) & 0 \\ 0 & \exp(ikL) \end{pmatrix} \quad (7.76)$$

Now we can discuss a more realistic device consisting of two interfaces that resembles an optical Fabry-Perot cavity. We must consider two boundaries and the interior. We will see that electrons with only certain initial speeds can be transmitted through the device. The dependence on the speed comes through the wave vector k , which depends on the De Broglie wavelength though $\lambda = h/(mv)$.

7.3.3 TRANSFER MATRIX FOR ELECTRON-RESONANT DEVICE

Consider a slab of material embedded within another material as shown in Figure 7.14. Assume that reflections occur at each of the two boundaries and that these two boundaries are parallel to each other. We assume the only input beam comes from the left and so $B_1 = 0$. Notice the reflectivity is assumed positive for waves reflecting off the inner surfaces. Starting with the right-hand interface we find using Equation 7.75,

$$\begin{pmatrix} A_2 \\ B_2 \end{pmatrix} = \frac{1}{t_{21}} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} A_1 \\ 0 \end{pmatrix}$$

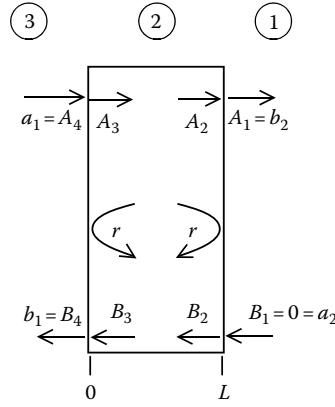


FIGURE 7.14 The resonant device. The circled numbers correspond to a given material. Materials 1 and 3 are assumed to be of the same type.

The subscript “21” on $t_{21} = t_{2 \rightarrow 1}$ indicates the signals moving from right to left across the right-hand interface for the formulas stated in previous sections. In principle, the reflectivity of the two facets can differ depending on the potential energy within each of the three regions in Figure 7.14. The waveguide (excluding the interfaces) has a transfer matrix as given above

$$\begin{pmatrix} A_3 \\ B_3 \end{pmatrix} = \begin{pmatrix} e^{-i\phi} & 0 \\ 0 & e^{i\phi} \end{pmatrix} \begin{pmatrix} A_2 \\ B_2 \end{pmatrix}$$

where $\phi = k_2 L$. The transfer matrix for the left-hand side is different from Section 7.3.1. Note that output side has $-r$ rather than $+r$ as in that section. We find

$$\begin{pmatrix} A_4 \\ B_4 \end{pmatrix} = \frac{1}{t_{32}} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \begin{pmatrix} A_3 \\ B_3 \end{pmatrix}$$

Assume the same type of materials for regions #3 and #1 so that $t_{32} = t_{12}$. Multiplying the three individual matrices provides the total transfer matrix

$$\begin{pmatrix} A_4 \\ B_4 \end{pmatrix} = \frac{1}{t_{12}} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \begin{pmatrix} e^{-i\phi} & 0 \\ 0 & e^{i\phi} \end{pmatrix} \frac{1}{t_{21}} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}.$$

Calculating the product, we find the total transfer matrix to be

$$\begin{pmatrix} A_4 \\ B_4 \end{pmatrix} = \frac{1}{t^2} \begin{pmatrix} e^{-i\phi} - r^2 e^{+i\phi} & r e^{-i\phi} - r e^{+i\phi} \\ -r e^{-i\phi} + r e^{+i\phi} & -r^2 e^{-i\phi} + e^{+i\phi} \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 = 0 \end{pmatrix} \quad (7.77)$$

The phase $\phi = k_2 L$ can have a complex wave vector. The complex part of k describes an exponentially decreasing or increasing wave function.

7.3.4 RESONANCE CONDITIONS FOR ELECTRON RESONANCE DEVICE

This section discusses the results of the application of a transfer matrix for an electron resonance device with a single input beam incident on the right-hand side. Equation 7.77 provides

$$\begin{pmatrix} A_4 \\ B_4 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 = 0 \end{pmatrix} \quad (7.78a)$$

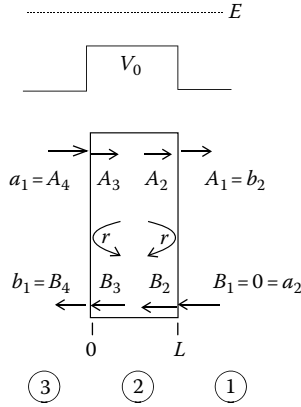


FIGURE 7.15 The amplitudes for the scattering matrix (lower case letters) and for the transfer matrix (upper case letters). Regions 1, 2, 3 describe $x < 0$, $0 < x < L$, $x > L$, respectively.

where

$$\underline{T} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} = \frac{1}{t^2} \begin{pmatrix} e^{-i\phi} - r^2 e^{i\phi} & re^{-i\phi} - re^{i\phi} \\ -re^{-i\phi} + re^{i\phi} & -r^2 e^{-i\phi} + e^{i\phi} \end{pmatrix} \quad (7.78b)$$

with the phase $\phi = k_2 L$ (Region 2 in Figure 7.15). For the sake of argument, assume that the potential in the region $(0, L)$ must be positive $V_0 > 0$ and zero everywhere else. Assume the electron energy $E > V_0$ so that the k -vectors are all real and assume “normal incidence” (i.e., the beams propagate perpendicular to the interfaces). Although the transfer matrix is a very useful mathematical abstraction, we eventually require the output amplitudes. The scattering matrix is better suited for this purpose. Recall the basic definition of the scattering matrix

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \underline{S} \begin{pmatrix} a_1 \\ a_2 = 0 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 = 0 \end{pmatrix} \quad (7.79)$$

The various amplitudes appear in Figure 7.15 for the scattering and transfer matrices. Equation 7.72 gives the relation between the two types of matrices.

$$S = \frac{1}{T_{11}} \begin{pmatrix} T_{21} & \det T \\ 1 & -T_{12} \end{pmatrix} = \frac{t^2}{e^{-i\phi} - r^2 e^{i\phi}} \begin{pmatrix} T_{21} & \det T \\ 1 & -T_{12} \end{pmatrix} \quad (7.80)$$

For the resonant device, we are interested in the output signal as a function of the input signal. We can solve for either the transmitted or reflected signal. Suppose we want to find the amplitude of the reflected signal. The scattering matrix provides the reflected signal as

$$b_1 = S_{11} a_1 \quad (7.81)$$

Equations 7.78 and 7.80 therefore provide the relevant transfer function

$$\frac{\text{Output}}{\text{Input}} = \frac{b_1}{a_1} = S_{11} = \frac{T_{21}}{T_{11}} = \frac{-re^{-i\phi} + re^{i\phi}}{e^{i\phi} - r^2 e^{i\phi}} = -r \frac{1 - e^{2i\phi}}{1 - r^2 e^{2i\phi}} \quad (7.82)$$

The interfaces between the two media have the same reflectivity. Notice the denominator might approach zero for certain values of phase ϕ which might be expected to indicate a resonance.

The current flowing in and out of the resonant device must be proportional to the square of the amplitudes.

$$J_{a1} = \frac{q\hbar k_1}{m} |a_1|^2 = \frac{q\hbar k_1}{m} |a_{01}|^2 \quad J_{b1} = \frac{q\hbar k_1}{m} |b_1|^2 = \frac{q\hbar k_1}{m} |b_{01}|^2 \quad (7.83)$$

as discussed in the previous two sections. Note that the reality of the k -vectors allows the current density to be nonzero and $k = \text{Re}(k)$. Changing notation from the subscripts a and b to the reflected current density $J_{\text{ref}} = J_{b1}$ and the incident current density $J_{\text{in}} = J_{a1}$ and using Equation 7.82 provides

$$J_{\text{ref}} = |S_{11}|^2 J_{\text{in}} \quad (7.84)$$

The reflected current J_{ref} actually originates from two sources. The first source consists of waves reflecting from the left-hand interface. The second source consists of waves that enter the middle layer, bounce around and then pass back out through the left-hand interface. The reflected current J_{ref} actually represents the superposition of many reflected beams from within the middle layer. Calculating the square of the complex transfer function S_{11} (Equation 7.82) we find

$$|S_{11}|^2 = S_{11} S_{11}^* = |r|^2 \frac{(1 - e^{2i\phi})(1 - e^{2i\phi})^*}{(1 - r^2 e^{2i\phi})(1 - r^2 e^{2i\phi})^*} \quad (7.85)$$

As a note on terminology, S_{11} refers to a transfer function (even though it appears as an element of the scattering matrix) because it relates an output parameter to an input parameter.

A few definitions will be helpful at this point. The phase factor $\phi = k_2 L$ will be real when $E > V$ but it will be complex in the most general case and have the form

$$k_2 = k_{2r} - ik_{2i}$$

(note the minus sign for convenience). The phase factor becomes

$$\phi = \phi_r + i\phi_i = k_{2r}L - ik_{2i}L$$

We can later set the imaginary part to zero; however some devices might have a purely imaginary phase depending on the relation between the energy of the particle and the magnitude of the potential. We can write the reflectance coefficient R in terms of the reflectivity $r = (k_2 - k_1)/(k_1 + k_2)$ as

$$R = |r|^2$$

Define an effective reflectance

$$\mathcal{R} = R \exp(-2\phi_i) \quad (7.86)$$

and write the (potentially) imaginary reflectivity as $r = |r| e^{i\alpha}$.

The current transfer function can now be written as

$$|S_{11}|^2 = R \frac{1 + \exp(-4\phi_i) - 2 \exp(-2\phi_i) \cos(2\phi_r)}{1 + R^2 \exp(-4\phi_i) - 2R \exp(-2\phi_i) \cos(2\phi_r + 2\alpha)} = R \frac{1 + \frac{\mathcal{R}^2}{R^2} - 2 \frac{\mathcal{R}}{R} \cos(2\phi_r)}{1 + \mathcal{R}^2 - 2\mathcal{R} \cos(2\phi_r + 2\alpha)}$$

Using the cosine expansion $\cos(2\phi_r) = \cos^2(\phi_r) - \sin^2(\phi_r) = 2 - 2\sin^2(\phi_r)$, we find

$$|S_{11}|^2 = R \frac{\left[1 - \frac{\mathcal{R}}{R}\right]^2 + 4 \frac{\mathcal{R}}{R} \sin^2 \phi_r}{[1 - \mathcal{R}]^2 + 4\mathcal{R} \sin^2(\phi_r + \alpha)} \quad (7.87)$$

The relation between the reflected current and the input current must be given by

$$J_{\text{ref}} = |S_{11}|^2 J_{\text{in}} = R \frac{\left[1 - \frac{\mathcal{R}}{R}\right]^2 + 4 \frac{\mathcal{R}}{R} \sin^2 \phi_r}{[1 - \mathcal{R}]^2 + 4\mathcal{R} \sin^2(\phi_r + \alpha)} J_{\text{in}} \quad (7.88)$$

Assume the phase ϕ is real (i.e., $\phi_i = 0$ and $\phi = \phi_r$), and the reflectivity is real ($\alpha = 0$) as required for $E > V$ everywhere so that $\mathcal{R} = R = r^2$. Equation 7.88 becomes

$$J_{\text{ref}} = |S_{11}|^2 J_{\text{in}} = r^2 \frac{2R - 2R \cos(2k_2L)}{1 + R^2 - 2R \cos(2k_2L)} J_{\text{in}} \quad (7.89a)$$

A similar procedure applies to the transmitted current

$$J_{\text{trans}} = |S_{21}|^2 J_{\text{in}}$$

since regions 1 and 3 produce the same k -vectors. Again assuming $E > V$ everywhere, the result has the form

$$|S_{21}|^2 = \frac{(1 - R)^2}{1 + R^2 - 2R \cos(2k_2L)} \quad (7.89b)$$

Figure 7.16 shows the (normalized) transmitted current for real $r = 0.56$ as a function of the phase $\phi_r = k_2L$. Here L represents the width of the barrier. The phase can be controlled by adjusting the height of the barrier V_0 or the energy E of the incident electron and $k_2 = \sqrt{2m(E - V_0)/\hbar^2}$. Notice that electrons with specific energy will be transmitted through the device to the other side. Larger values of R increase the selectivity for the specific energies.

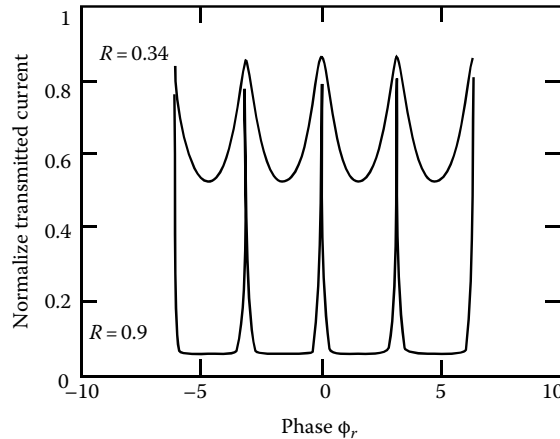


FIGURE 7.16 The normalized transmitted currents for two different values of reflectance.

7.3.5 QUANTUM TUNNELING

Using an approach similar to that for the electron-resonant device, we can easily calculate the amplitude of a wave tunneling through a barrier. We assume the energy of the electron is smaller than the barrier height $0 < E < V_0$. In the barrier, the wave vector becomes imaginary

$$k_2 = \sqrt{\frac{2m}{\hbar^2}(E - V_0)} = i\kappa_2$$

where κ_2 must be real. The scattering matrix provides

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \underline{S} \begin{pmatrix} a_1 \\ a_2 = 0 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 = 0 \end{pmatrix}$$

so that the transmitted amplitude must be

$$b_2 = S_{21}a_1$$

Where Equation 7.80

$$S = \frac{1}{T_{11}} \begin{pmatrix} T_{21} & \det T \\ 1 & -T_{12} \end{pmatrix} = \frac{t^2}{e^{-i\phi} - r^2 e^{i\phi}} \begin{pmatrix} T_{21} & \det T \\ 1 & -T_{12} \end{pmatrix}$$

with $\phi = k_2 L = i\kappa_2 L$ provides the transmitted amplitude

$$\frac{b_2}{a_1} = S_{21} = \frac{t^2}{e^{-i\phi} - r^2 e^{i\phi}}$$

where

$$\begin{aligned} t^2 &= t_{12}t_{21} & t_{12} &= \frac{2k_1}{k_1 + i\kappa_2} & t_{21} &= \frac{2i\kappa_2}{k_1 + i\kappa_2} \\ -r &= \frac{k_1 - i\kappa_2}{k_1 + i\kappa_2} & \rightarrow & & r^2 &= \frac{k_1^2 - \kappa_2^2 - i2k_1\kappa_2}{k_1^2 - \kappa_2^2 + i2k_1\kappa_2} \end{aligned}$$

We find

$$\frac{J_{\text{trans}}}{J_{\text{inc}}} = \frac{|t|^4}{4 \sin h^2(\kappa_2 L) + |t|^4} \quad (7.90)$$

where

$$|t|^4 = \left(\frac{4k_1\kappa_2}{k_1^2 + \kappa_2^2} \right)^2 \quad (7.91)$$

Figure 7.17 shows an example plot for the energy midway between 0 and V_0 so that $k_1 = \kappa_2$. The particle has a finite probability of tunneling through the barrier even though it does not have sufficient energy. Classically, one would never see such an effect.

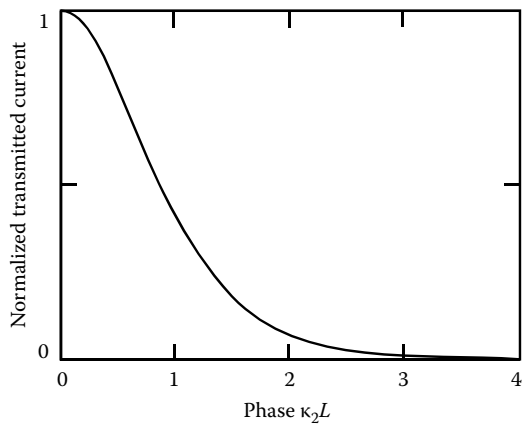


FIGURE 7.17 Example plot of normalized current for the case of $k_1 = \kappa_2$.

7.3.6 TUNNELING AND ELECTRICAL CONTACTS

Quantum tunneling has very common application to electrical contacts. At a junction between dissimilar materials, such as for pn junction or Schottky junctions (metal semiconductor), potential barriers can form. These barriers can inhibit the flow of carriers from one region to another and thereby make them highly resistive or nonlinear in the current–voltage relations. Charge carriers can efficiently tunnel through sufficiently narrow barriers and thereby produce low resistance Ohmic contacts.

Tunneling also has similar applications to thin layers of silicon dioxide (glass) separating two materials such as for metal-oxide-semiconductor (MOS) devices. Although in this case, one typically prefers to suppress the quantum tunneling in order to maintain highly resistive gates.

One can find the Wentzel-Kramers-Brillouin (WKB) approximation in many books on quantum mechanics. However, consider the following simplified argument to determine the probability of transferring a charged particle through a barrier such as the one shown in Figure 7.18. Suppose the potential energy barrier V depends on position and that $V = qv_b$ where v_b represents a voltage (for the barrier) and q the charge carried by the particle interacting with the barrier (such as in Figure 7.18). Assume $E < V(x)$ everywhere. Consider the region divided into small distances Δx over which, the potential V is relatively constant. For constant V and hence constant wave vector k , a solution to the time-independent wave equation must have the form

$$\psi(x) \sim \psi(0)e^{ikx}$$

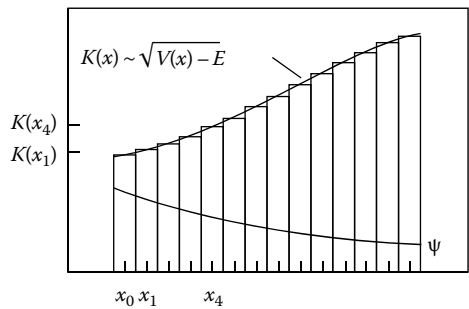


FIGURE 7.18 Example wave function due to the potential barrier that produces $K(x)$.

For $E < V$, then k will be complex and so define $k = iK$ where K is real. The solution will have the form

$$\psi(x) \sim \psi(0)e^{-Kx}$$

So the first rectangle in Figure 7.18 will produce a solution similar to

$$\psi(x_1) \sim \psi(0)e^{-K_1\Delta x}$$

where Δx represents the widths of the small rectangles and K_1 is the value evaluated at x_1 and

$$K_n = \left[\frac{2m}{\hbar^2} (V(x_n) - E) \right]^{1/2}$$

The next rectangle at x_2 produces further exponential decay from that in the rectangle at x_1 and might be written as

$$\psi(x_2) \sim \psi(x_1)e^{-K_2\Delta x} = \psi(0)e^{-K_1\Delta x}e^{-K_2\Delta x} = \psi(0)e^{-K_1\Delta x - K_2\Delta x}$$

Continuing the process and allowing Δx to approach zero produces

$$\psi(L) = \psi(0)e^{-\sum_n K_n\Delta x} \rightarrow \psi(0)e^{-\int dx K(x)}$$

where L represents the width of the barrier. Substituting for K and $V = qv_e$ into the probability of finding the particle at L

$$\psi^*\psi = \psi^*(0)\psi(0)e^{-2\int dx K(x)} = \psi^*(0)\psi(0)e^{-2\int dx \sqrt{\frac{2mq}{\hbar^2}(v_b - v_e)}}$$

where $E = qv$ is the energy of the particle. Finally then, the current can be expected to have the form

$$\left(\begin{array}{c} \text{Current} \\ \text{density at } L \end{array} \right) = q(\text{Speed}) \left(\begin{array}{c} \text{\#Electrons} \\ \text{prior to barrier} \end{array} \right) \left(\begin{array}{c} \text{Probability of} \\ \text{finding } e^- \text{ at } L \end{array} \right)$$

$$J \sim qvN e^{-2\int dx \sqrt{\frac{2mq}{\hbar^2}(v_b - v_e)}}$$

7.4 INTRODUCTION TO FREE AND NEARLY FREE QUANTUM MODELS

The time-independent Schrödinger equation provides the energy levels and states. For the case of plane wave solutions, the Laplacian ∇^2 then makes the energy eigenvalues E depend on the wave vector k . In particular, the time-independent wave equation produces a dispersion relation $E = E(\vec{k})$ from which one can obtain phase and group velocity. In addition, the dispersion curve provides the values of all allowed energy states. The free electron model uses a constant potential that produces a quadratic dispersion curve (E vs. k) and plane wave energy basis states. The nearly free electron model incorporates the periodic potential of the crystal that produces a nonquadratic dispersion curve $E(\vec{k})$ with energy gaps and basis states composed of the Bloch wave functions. The Bloch wave functions consist of the product of a plane wave portion similar to the free-electron case and a function periodic in the lattice. The present section discusses the role of periodic boundary conditions and the physical origin and implications of the bandgaps.

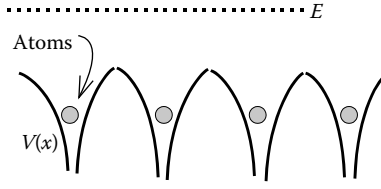


FIGURE 7.19 Example periodic potential $V(x)$ for an electron in a 1-D monatomic crystal.

7.4.1 POTENTIAL IN CUBIC MONATOMIC CRYSTAL

A crystal has a spatially periodic array of atoms bonded to one another. The directional character of the atomic orbitals dictates the bonding pattern and the resulting lattice type. The periodic arrangement of the atoms in the crystal results in a periodic crystal potential as shown for the 1-D case in Figure 7.19. That is, if $\vec{R}$ represents a lattice vector (i.e., a vector starting on one lattice site and ending on another) then the potential has the property $V(\vec{r} + \vec{R}) = V(\vec{r})$. The electrostatic potential energy for the electron decreases near the core of the atom on account of the attractive coulomb force between the positively charged core and the negatively charge electron.

The time-dependent Schrödinger equation for the situation depicted in Figure 7.19 has the form

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\vec{r},t) + V(\vec{r})\Psi(\vec{r},t) = i\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t) \quad (7.92)$$

Applying an external potential to the crystal (such as with a battery) will add an extra term to the potential V . For sufficiently large total energy $E \gg V(\vec{r})$ we might consider the variations of the periodic potential energy $V(\vec{r})$ to be negligible and thereby average the potential $V(\vec{r})$ to a constant (set to zero for simplicity).

7.4.2 FREE ELECTRON MODEL

The free electron model treats the motion of the electron in either free space or in a material when the periodic potential of a crystal can be neglected. In such a case, the Schrödinger wave equation in Equation 7.92 should be rewritten with the potential energy term $V(\vec{r})$ taken as a constant (zero for simplicity).

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\vec{r},t) = i\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t) \quad (7.93)$$

We already know the solutions to Equation 7.93 with $V=0$ must be plane waves of the form

$$\Psi = c_{\vec{k}} e^{i\vec{k}\cdot\vec{r} - i\omega t} \quad (7.94a)$$

The constant $c_{\vec{k}}$ symbolizes the normalization factor for the wave function. The terms

$$\psi_{\vec{k}}(\vec{r}) = c_{\vec{k}} e^{i\vec{k}\cdot\vec{r}} \quad (7.94b)$$

represent the energy basis states while the $e^{-i\omega t}$ always occurs for a closed system. In general, the full solution will be a time-dependent summation over the plane wave basis states resulting in a Fourier series or Fourier transform.

In order to normalize the wave functions in Equation 7.94a and b, one must choose a normalization volume. Consider first, a 1-D case where

$$\psi_k(x) = c_k e^{ik_x x}$$

The plane wave has infinite extent along x so that $x \in (-\infty, \infty)$ and the normalization would be

$$1 = \langle \psi_k | \psi_k \rangle = \lim_{L \rightarrow \infty} \int_{-L}^L dx \psi_k^*(x) \psi_k(x) \Rightarrow |c_k|^2 = \frac{1}{2L} \rightarrow 0$$

That is, such a normalization would produce a zero wave. Typically, plane waves are normalized to a finite length L such as the length of a crystal using periodic boundary conditions in the sense of $\psi_k(x+L) = \psi_k(x)$. The normalization for the 1-D case becomes, for example,

$$1 = \langle \psi_k | \psi_k \rangle = \int_0^L dx \psi_k^*(x) \psi_k(x) \Rightarrow c_k = \frac{1}{\sqrt{L}} \quad (7.95a)$$

where the phase of c_k has been ignored. For three dimensions, the normalization will have three integrals over the ranges $(0, L_x)$, $(0, L_y)$, $(0, L_z)$ and therefore the normalization constant will be

$$c_{\vec{k}} = \frac{1}{\sqrt{L_x L_y L_z}} = \frac{1}{\sqrt{V}} \quad (7.95b)$$

Note that the normalization of the basis states dictates the range of integration for the inner product and does not alter the fact that the wave has infinite extent.

Continuing with the 1-D case, the dispersion curve E versus k_x can be found by substituting Equation 7.94 into Equation 7.93

$$\frac{\hbar^2 k_x^2}{2m} = \hbar \omega_k = E \quad (7.96)$$

For zero potential energy as assumed for Equation 7.93, the energy E consists solely of kinetic energy. If Equation 7.92 had a nonzero constant potential then the dispersion curve would shift along the energy axis away from zero and E would contain a component of potential energy. Equation 7.96 represents the dispersion curve for the free particle. The free electron model produces a quadratic dispersion curve without any energy gaps as shown in Figure 7.20.

The normalization of the plane wave using periodic boundary conditions also determines allowed values of the wave vector $\vec{k}$, which for 1-D would be $k_x = 2\pi n/L_x$ where L_x represents a macroscopic scale parameter (often the size of the crystal). Consequently, the dispersion curve must

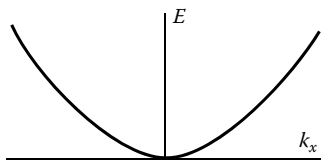


FIGURE 7.20 The dispersion curve for the quantum mechanical free electron.

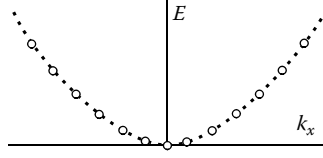


FIGURE 7.21 The free electron dispersion curve with allowed values of wave vector k_x due to periodic boundary conditions.

be augmented with information on the allowed wave vectors. This means that only certain discrete values of E will be allowed as represented by the circles in Figure 7.21. As a note, the free electron dispersion curve does not have Brillouin zones since the potential is not periodic in atomic spacing.

The allowed values of $\vec{k}$ and hence the allowed energy eigenvalues determine the solution to the time-dependent Schrödinger wave equation as a discrete summation over the basis states $\Psi(\vec{r}, t) = \sum_{\vec{k}} \beta_{\vec{k}}(t) \psi_{\vec{k}}(\vec{r}) = \sum_{\vec{k}} \beta_{\vec{k}} e^{i\vec{k} \cdot \vec{r} - i\omega_{\vec{k}} t} / \sqrt{V}$ where $\beta_{\vec{k}} = \beta_{\vec{k}}(0)$. For very large macroscopic length L , the allowed wave vectors will essentially form a continuum and the dispersion curve will be very similar to the one in Figure 7.20. For the continuum limit, the solution to the time-dependent Schrödinger wave equation will be the Fourier integral $\Psi(\vec{r}, t) = \int d^3k \beta_{\vec{k}}(t) \psi_{\vec{k}}(\vec{r}) = \int d^3k \beta_{\vec{k}} e^{i\vec{k} \cdot \vec{r} - i\omega_{\vec{k}} t} / \sqrt{V}$.

Sometimes an alternate normalization is used for the wave function. While the wave function will be assumed periodic over a macroscopic length L , the wave function is normalized according to the number of particles per unit volume N/V . The term $1/V$ in the normalization such as Equation 7.95b can be interpreted as one particle per volume so that the inner product $\langle \psi_{\vec{k}} | \psi_{\vec{k}} \rangle = 1$ shows the one particle in the region. Using the normalization constant of $\sqrt{N/V}$ would produce an inner product of $\langle \psi_{\vec{k}} | \psi_{\vec{k}} \rangle = N$ which then indicates N particles per volume.

7.4.3 NEARLY FREE ELECTRON MODEL

The nearly free electron model describes electrons moving through the periodic potential of the crystal. Figure 7.19 shows an example for a monatomic crystal with atomic spacing of a . The potential $V(\vec{r})$ must be included in the Schrödinger Equation as in Equation 7.92 and repeated here.

$$-\frac{\hbar^2}{2m} \nabla^2 \Psi(\vec{r}, t) + V(\vec{r}) \Psi(\vec{r}, t) = i\hbar \frac{\partial}{\partial t} \Psi(\vec{r}, t)$$

The time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m} \nabla^2 \psi_{\vec{k}}(\vec{r}) + V(\vec{r}) \psi_{\vec{k}}(\vec{r}) = E_{\vec{k}} \psi_{\vec{k}}(\vec{r})$$

provides the energy basis states $\psi_{\vec{k}}$ and the energy levels $E_{\vec{k}}$. The periodic potential gives rise to band structure.

The energy basis functions found as solutions to the time-independent Schrödinger equation consists of the Bloch wave functions

$$\psi_{\vec{k}} = \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{V}} u_{\vec{k}}(\vec{r}) \quad (7.97)$$

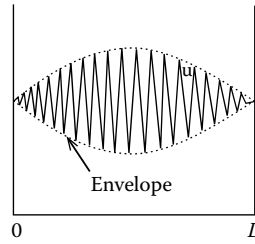


FIGURE 7.22 A Bloch wave function for an electron in a semiconductor infinitely deep well. The long wavelength for the envelope satisfies the macroscopic boundary conditions at $x=0, L$. The small wavelength provides a periodicity matching that of the crystal.

where the function $u_{\vec{k}}(\vec{r})$ has the periodicity of the lattice (c.f., Figure 7.22). That is, if $\vec{R}$ symbolizes a direct lattice vector (that begins on an arbitrary lattice point and ends on an arbitrary lattice point) then $u_{\vec{k}}(\vec{r})$ satisfies $u_{\vec{k}}(\vec{r} + \vec{R}) = u_{\vec{k}}(\vec{r})$; consequently, $u_{\vec{k}}(\vec{r})$ is invariant under lattice translations. The envelope $e^{i\vec{k}\cdot\vec{r}}/\sqrt{V}$ resembles the plane wave solutions

The structure of the Bloch wave function somewhat resembles that for an amplitude modulated radio wave. The “spatially faster” wave $u_{\vec{k}}(\vec{r})$ takes the place of the carrier wave and the envelope plays the part of the modulation on the carrier. Figure 7.22 shows an example for an electron in an infinitely deep well made from a crystal. The faster, periodic wave $u_{\vec{k}}(\vec{r})$ repeats every atomic spacing (i.e., every primitive cell). The broader envelop (actually the sine wave in the figure consists of the sum of complex exponentials) shows that the wave vector k_x must have values $n\pi/L$ (where n is an integer) in order that the wave function be zero at the boundaries. Similar to the free-electron case, $\vec{k}$ represents the electron propagation vector. The envelope function solves the free-electron Schrödinger wave equation so long as it incorporates the effective electron mass rather than the free electron mass. Therefore, we can use the solutions to the various well configurations previously developed so long as we use the envelope wave function and the effective mass. The subsequent sections will further discuss this point and the energy basis functions.

The energy eigenvalues $E_{\vec{k}}$ produce the dispersion curves shown in Figure 7.23 (for a 1-D example). Notice that the periodic potential has opened bandgaps in the dispersion curves. The figure shows both the “extended” and “reduced” zone schemes. The reduced zone scheme represents the familiar “view” of the energy bands by translating all of the bands into the first Brillouin zone (FBZ). Notice the familiar “bandgap” between the two dotted bands. Also notice for the reduced zone scheme that an extra index n must be added to the states $\psi_{n\vec{k}}$ and energy values $E_{n\vec{k}}$ in order to distinguish the state according to the band number n ; this added index is not necessary for the extended zone scheme since each k -value has only one associated state.

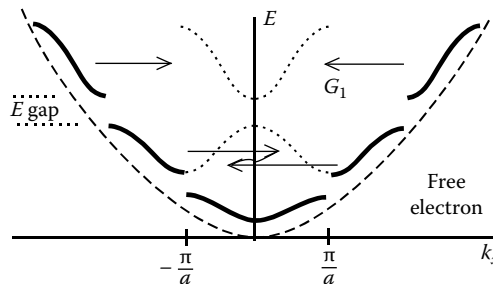


FIGURE 7.23 Comparing the dispersion curves for the free electron and the nearly free electron. The symbol a represents the interatomic spacing so that $2\pi/a$ gives the width of the FBZ. Solid curves represent the “extended zone” scheme and the dotted curves represent the “reduced zone” scheme.

To transform from the extended to reduced zones, the bands are shifted into the FBZ by adding or subtracting a reciprocal lattice vector G . The edges of the FBZ have the values $\pm G_1/2$ where $G_1/2$ is related to the interatomic spacing a by π/a . The Kronig–Penney model for the bands, for example, shows the reasons for shifting by the reciprocal vector G . As will be discussed further below, the wave vectors k_x larger than π/a produce electron wavelengths smaller than $2a$.

Even a small infinitesimal periodic potential alters the topology of the dispersion curve by opening infinitesimally small gaps. However, sufficiently small gaps have negligible physical effect. Figure 7.23 shows the gaps and the nonparabolic form of the dispersion curves. The dotted parabolic curve represents the dispersion curve for the free electron model. For small periodic potential, the gaps become smaller and the dispersion curves for the free and nearly free electron models coincide. Although not evident from the figure, the gaps become smaller with increasing energy in Figure 7.23. The name “bands” originates from the fact that the range of energy divides into allowed and disallowed regions of energy. Two types of allowed energy must be included in the nearly free dispersion curve. In order to normalize the wave function, boundary conditions must be introduced that also quantize the wave vector and energies. These allowed energies appear as allowed states in the dispersion curves. There are not any allowed energy states in the bandgaps.

The periodic boundary conditions in the volume $V = L_x L_y L_z$ produce both the normalization of $\sqrt{V}$ given in Equation 7.97 and the allowed values of $\vec{k}$. The allowed values of the wave vector $\vec{k}$ come from the macroscopic boundary conditions, which usually span a distance much larger than $L \sim 100 \text{ \AA}$. Therefore, the spacing between allowed wave vectors must be on the order of $\Delta k \sim 2\pi/L \sim 0.06$. The reciprocal lattice vectors G provide important markers for the band diagram. The set of reciprocal lattice vectors $G_n = 2n\pi/a$ (for 1-D) for a cubic monatomic crystal comes from the interatomic spacing a . For a lattice constant on the order of $a \sim 5 \text{ \AA}$, we find $\Delta G \sim 2\pi/a \sim 1$. The first value of G denotes the first Brillouin zone (FBZ) for k -space which has the width $2\pi/a$ (see Figure 7.23 for example). The wave vectors G have spacing much larger than the wave vectors k_x as can be seen from the “zoomed-in” view in Figure 7.24. We therefore see that the wave vectors $\{\vec{k}\}$ and $\{\vec{G}\}$ cannot be the same. If $\vec{b}$ is a primitive reciprocal lattice vector (along the x -direction for example) then the length of $n\vec{b}$ must increase for any integer n ; however, the length would need to decrease to produce most of the $\vec{k}$ -vectors. If we write $c\vec{b}$ where c is not integer, then the quantity $c\vec{b}$ can be a k -vector. This will become important for proving the Bloch wave function.

The bandgaps in the dispersion curves occur near the Brillouin zone edges defined by the reciprocal lattice vectors. Near the zone edges, the dispersion curve has zero slope and the electron must have negligible group velocity there. We can understand this behavior as follows. Electrons propagating through the periodic structure experience reflections due to the periodic potential. Basically, these reflections produce resonant effects similar to those discussed in connection with the electron-resonant device in Section 7.3. Near resonance, strong reflections prevent forward motion of the electron wave function. As with phonons, these resonant effects must occur near the

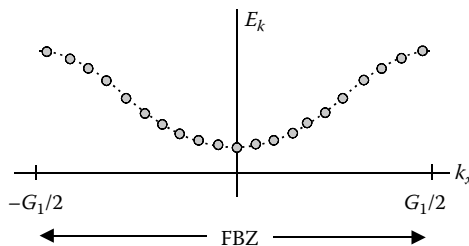


FIGURE 7.24 Zoomed-in view of lowest order band depicting the allowed k -values produced by periodic boundary conditions and their relation to the reciprocal lattice vectors.

Brillouin zone edges. For example, near π/a the forward moving plane wave has the form e^{ikx} but strong reflections produce the reverse propagating plane wave of the form e^{-ikx} where $k \sim \pi/a$. As a result, the total wave has the form of a sine or a cosine (refer to the next section) $e^{ikx} \pm e^{-ikx} \sim \sin(kx), \cos(kx)$ which represent standing waves that do not propagate.

7.4.4 BRAGG DIFFRACTION AND GROUP VELOCITY

Section 7.3 on the transfer matrix shows how certain barrier widths in conjunction with certain barrier heights produce very strong reflections. There we were considering the reflection of an envelope function from layers composed of many atoms. The present section shows how the same can occur for electrons reflecting from atoms. We will see why the bands flatten-out near values of $G/2$.

Suppose an electron enters an array of atoms as shown in Figure 7.25. For convenience, even though a single wave function (wave packet) describes the incident electron, label the two spatially separated parts of the wave function as ψ_1, ψ_2 . Initially both wave functions have the form e^{ikz} where z measures the distance along the path including the reflections. The bottom path exceeds the top one by the distance $\Delta z = 2d = 2a \cos(\theta)$. The wave along the bottom path moves the extra distance Δz compared with the portion of the wave moving along the top path. As a result, we find at the right-hand end of the path

$$\psi_1 + \psi_2 = e^{ikz} + e^{ik(z+\Delta z)} = e^{ikz}(1 + e^{ik\Delta z}) \quad (7.98)$$

If the last term in Equation 7.98 has the value $e^{ik\Delta z} = 1$, then the two waves constructively interfere on the right-hand side of the figure. Therefore, strong reflection occurs for $k\Delta z = 2\pi$. Substituting for the extra path length $\Delta z = 2d = 2a \cos(\theta)$ and setting $\theta = 0$ for normally incident electrons we find

$$k = \frac{\pi}{a} \quad (7.99)$$

Apparently we expect strong reflections every $G/2$. All of the zone boundaries occur at half multiples of reciprocal lattice vectors.

The strong reflections at the zone boundary produces zero slope in the E versus k diagram. This means the group velocity of the wave must be negligible for wave vectors near the zone boundary. This occurs because the reflections impede the forward motion of the electron. However, the electron cannot move in the reverse direction either because the lattice reflects it back from that direction as well. As a result, the electron cannot move and forms a standing wave.

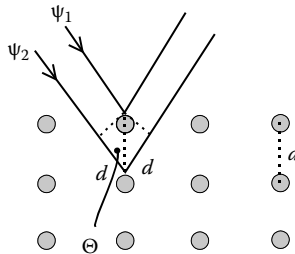


FIGURE 7.25 Electron incident on an array of atom.

7.4.5 BRIEF DISCUSSION OF ELECTRON DENSITY AND BANDGAPS

In this section, we discuss the origin of the bandgap from the point of view of stationary waves for the electrons. As just discussed for wave vector k near one of the zone boundaries, say $k \cong \pi/a$, forward propagating plane waves e^{ikx} must be strongly reflected to produce a backward propagating plane wave e^{-ikx} . Therefore, the total wave function satisfying the Schrödinger wave equation must consist of the summation of the two plane waves moving toward the right and toward the left. Let a be the lattice constant as usual. The standing waves lead to a charge distribution.

$$\psi_+ = e^{ikx} + e^{-ikx} \sim C_+ \cos(kx) = C_+ \cos\left(\frac{\pi x}{a}\right) \rightarrow \rho_{q+} = q|C_+|^2 \cos^2\left(\frac{\pi x}{a}\right) \quad (7.100a)$$

$$\psi_- = e^{ikx} - e^{-ikx} \sim C_- \sin(kx) = C_- \sin\left(\frac{\pi x}{a}\right) \rightarrow \rho_{q-} = q|C_-|^2 \sin^2\left(\frac{\pi x}{a}\right) \quad (7.100b)$$

where $\rho_{q-} = q\psi_-^* \psi_-$ and $\rho_{q+} = q\psi_+^* \psi_+$ are the charge densities associated with the cosine and sine standing waves.

Notice for the cosine wave ψ_+ , that the charge density becomes a maximum at the position of the atoms $x=na$ (where n is an integer) but for the sine wave ψ_- , the charge density becomes a minimum at the position of the atoms as shown in Figure 7.26. For the cosine wave ψ_+ , the electron experiences lower potential energy than for the sine wave because the electron (for the cosine wave) is mostly centered over the atom where it has lower potential energy. The electron represented by the sine wave ψ_- exists mostly between the atoms and therefore must have the higher potential energy. The two states ψ_+ and ψ_- live in two adjacent bands at the zone edge since they have the same wave vector but they are orthogonal. The bandgap will then be given by

$$\Delta E = \langle \psi_- | \hat{H} | \psi_- \rangle - \langle \psi_+ | \hat{H} | \psi_+ \rangle = E_- - E_+$$

where $\hat{H} = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x)$, $\hat{H}\psi_{\pm} = E_{\pm}\psi_{\pm}$ and $V(x)$ is invariant to lattice translations.

To find the bandgap ΔE , it is only necessary to evaluate the expectation value of $\hat{H}$ using the wave functions in Equation 7.100a and b. The kinetic energy terms for either ψ_+ or ψ_- produces the same value of $\hbar^2 k^2 / 2m$ and therefore cancel in ΔE . Therefore, the expectation value of the potential energy leads to the bandgap. This calculation is easiest if we represent $V(x)$ by a Fourier series over the reciprocal lattice vectors.

$$V(x) = \sum_G V_G \frac{e^{iGx}}{\sqrt{a}} = \sum_n V_n \frac{e^{in^2 \pi x / a}}{\sqrt{a}}$$

where we have assumed periodic boundary conditions over the lattice constant a . We will find only $n=1, -1$ produces nonzero terms in the final analysis. For the higher order BZs, the $n=2, 3, \dots$. The $n=0$ term is ignored because it is a DC average of the potential and only displaces the dispersion curve along the energy axis. The potential is real and symmetric which requires $V_1 = V_{-1}$. Writing ψ_+ and ψ_- in terms of complex exponentials, one can show that

$$|\Delta E| = 2|V_1|$$

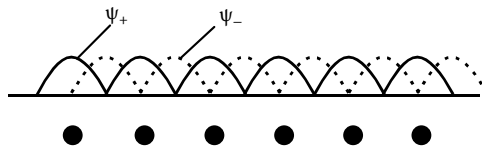


FIGURE 7.26 The two charge densities corresponding to k near the edge of the FBZ.

7.5 BLOCH FUNCTION

Previous sections and chapters have discussed electrons and holes (1) moving through semiconductor materials, (2) confined to quantum wells, and (3) scattering from barriers and wells. The discussion paid little attention to the fact that atomic cores produce a potential for the electrons and except for Section 7.4, focused primarily on the free electron model. This section begins to include the effects of the periodic potential due to the atoms in the crystal lattice. As will be seen in the next section, all of the previous examples and analysis using only the free electron model remain valid so long as the effective mass replaces the actual mass of the particle. The effect of the periodic potential appears in the effective mass of the electron. Section 7.5.1 discusses how the wave functions used in previous sections really represent the envelope wave functions for Bloch's wave functions. The envelope functions appear very similar to the modulation of a fast-varying carrier in communications such as for amplitude modulation.

7.5.1 INTRODUCTION TO BLOCH WAVE FUNCTION

Schrödinger's equation for a single electron in the periodic potential $U(\vec{r})$ can be written as

$$-\frac{\hbar^2}{2m} \nabla^2 \Psi + U\Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (7.101)$$

where m represents the actual mass of the electron (not the effective mass). We use the symbol U to emphasize the periodic nature of the potential and not to be confused with potentials V associated with wells. As discussed in Section 6.2, the potential energy U has the periodicity of the lattice if, for $\vec{R}$ a direct lattice vector, the potential has the property that

$$U(\vec{r} + \vec{R}) = U(\vec{r}) \quad (7.102)$$

We can separate variables in Equation 7.101 using $\Psi(\vec{r}, t) = \psi(\vec{r})T(t)$ to find the time-independent Schrödinger wave equation

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + U\psi = E\psi \quad (7.103)$$

As usual, we want to find the eigenfunctions and eigenvalues for the time-independent Schrödinger equation in Equation 7.103. The solutions to the time-dependent Schrödinger wave equation are the time-dependent superpositions of the energy eigenfunctions (that form a basis set). We start by examining the number of energy values before proceeding with the eigenfunctions. The allowed energy values (such as in Figure 7.28) are the eigenvalues from Equation 7.103.

As the reader might already know from previous studies, the bands describe the dynamics of the electron in the crystal. Semiconductors have at least two bands of allowed energy, namely the conduction and valence bands. An example for an indirect band appears in Figure 7.27. Later

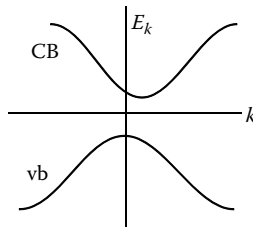


FIGURE 7.27 Generic band structure.

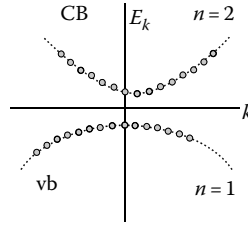


FIGURE 7.28 A zoomed-in view of the bands showing individual states.

sections will discuss bands and how they arise. What are the allowed states and how do we label them? Boundary conditions (c.f., Sections 7.13 through 7.15) impose quantization conditions on the wave vectors and hence the energy; therefore, the system supports only certain wave vectors and energies. We usually say that $\vec{k}$ specifies the state. However, a complication arises for multiple bands as in Figure 7.27. Each k value provides an allowed state for the conduction band and an allowed state for the valence band as indicated in the zoom-view in Figure 7.28. It does not matter what band an electron occupies—the states still represent plane waves. An electron promoted from the valence band to the conduction band leaves behind an empty state in the valence band. However this empty state behaves as though it were occupied by positively charge particle, namely a hole. We can think of the plane waves for the valence band as representing the holes. Regardless of our thinking on the two bands, we still must distinguish between two sets of allowed energies for each wave vector $\vec{k}$. We let the integer n represent the band such that, for example, $n = 1$ for the valence band and $n = 2$ for the conduction band. The energy eigenvalues must then be $E_{n,\vec{k}}$. The function $E_{1,\vec{k}} = E_1(\vec{k})$ represents the allowed energy states in the valence band (it gives the dispersion curve) and $E_{2,\vec{k}} = E_2(\vec{k})$ represents the allowed energy states in the conduction band. Generally, semiconductors produce more than just two bands as we will see. We label the bands by an index n . The energy eigenvalues must carry the additional subscript because we must specify the band. The energy eigenvalue $E_{n,k}$ specifies the energy of an electron in band n with wave vector k .

Having specified notation for the eigenvalues, we can now enumerate the eigenstates forming a basis for the Hilbert space. Continuing with the two-band example in Figure 7.28, we must specify an eigenfunction for each energy eigenstate. The single eigenstate corresponding to $E_{n,\vec{k}}$ can be written as $|E_{n,\vec{k}}\rangle = |n, \vec{k}\rangle = |n, k_x, k_y, k_z\rangle$. The states $|1, \vec{k}\rangle$ refer to the valence band and the states $|2, \vec{k}\rangle$ refer to those in the conduction band. In the coordinate representation, we can write $|E_{n,\vec{k}}\rangle = |n, \vec{k}\rangle \rightarrow \psi_{n\vec{k}}(\vec{r})$. Keep in mind that the index n can have more than two values since there can be more than two bands.

Now Bloch's form of the energy eigenfunctions can be stated. The energy eigenfunctions can be shown to consist of two separate functions.

$$\psi_{n\vec{k}}(\vec{r}) = \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) \quad (7.104)$$

The product contains the plane wave $e^{i\vec{k}\cdot\vec{r}}$ and a function u having the periodicity of the lattice. That is, the function u has the property that $u_{n\vec{k}}(\vec{r} + \vec{R}) = u_{n\vec{k}}(\vec{r})$. The subscripts indicate the possibility of different periodic functions $u_{n\vec{k}}(\vec{r})$ depending on the band. Equation 7.104 must be the coordinate representation of the eigenvector

$$|E_{n,\vec{k}}\rangle = |n, \vec{k}\rangle \rightarrow \psi_{n\vec{k}}(\vec{r}) = \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r})$$

In Equation 7.104, u represents the wave function for a single unit cell. Often books and the literature will leave off the normalization $\sqrt{V}$. The traveling plane wave $e^{i\vec{k}\cdot\vec{r}}$ constitutes an envelope function. The full solution to Schrödinger's wave equation requires a summation over the basis functions (the eigenfunctions ψ_{nk}).

$$\psi(x, t) = \sum_{n,k} \beta_{nk}(t) \psi_{nk}(x) = \sum_{n,k} \beta_{nk}(t) \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) \quad (7.105)$$

We therefore use the envelope function to satisfy the macroscopic boundary conditions (see Figure 7.29). For example, consider the infinitely deep well (which is not really possible using physical materials). We know the wave function must be zero at the boundaries of the infinitely deep well and this produces sinusoidal wave functions. A solution to the time-independent Schrödinger wave equation might be expected to have the form

$$X(x) = C_1 e^{ikx} u_{2,k}(x) + C_2 e^{-ikx} u_{2,-k}(x)$$

for an electron in the conduction band ($n=2$). If we assume the function u is symmetric in k then we can write

$$X(x) = (C_1 e^{ikx} + C_2 e^{-ikx}) u_{2,k}(x)$$

We therefore require the summation over the envelope wave function to be zero at the boundaries. Similar to Chapter 5 using $X(0)=0$ so that $C_1 = -C_2$, we expect

$$X(x) = (C_1 e^{ikx} + C_2 e^{-ikx}) u_{2,k}(x) \sim C_1 \sin(kx) u_{2,k}(x)$$

Subsequent sections show that the Schrödinger wave equation for a material with a macroscopic potential (such as a quantum well) only needs to include the envelop function (the $\sin(kx)$ in this case) so long as the Schrödinger equation uses the effective mass and does not include the periodic potential (c.f., Sections 7.6 and 7.10).

We might picture the energy eigenfunctions $\psi_{n\vec{k}}$ as shown in Figure 7.29; the dotted curve corresponds to the envelope. For the “standing wave” shown in the figure, we have assumed the electron is in the conduction band ($n=2$) and that the envelope function consists of a right-traveling and left-traveling plane wave (i.e., $k > 0$, and $-k$). Of course the same reasoning applies to other physical situations besides the infinitely deep well. Another example might be the wave packet traveling through a semiconductor.

Equation 7.104 can be restated as

$$\psi_{n\vec{k}}(\vec{r} + \vec{R}) = e^{i\vec{k}\cdot\vec{R}} \psi_{n\vec{k}}(\vec{r}) \quad (7.106)$$

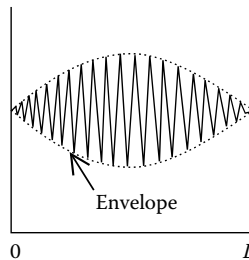


FIGURE 7.29 Example of envelope function for infinitely deep well.

where $\vec{R}$ is a direct lattice vector. We can easily demonstrate this result by using Equation 7.104

$$\psi_{n\vec{k}}(\vec{r} + \vec{R}) = \frac{e^{i\vec{k} \cdot (\vec{r} + \vec{R})}}{\sqrt{V}} u_{n\vec{k}}(\vec{r} + \vec{R}) = e^{i\vec{k} \cdot \vec{R}} \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) = e^{i\vec{k} \cdot \vec{R}} \psi_{n\vec{k}}(\vec{r}) \quad (7.107)$$

where we have used the periodicity of the function u , namely $u_{n\vec{k}}(\vec{r} + \vec{R}) = u_{n\vec{k}}(\vec{r})$.

7.5.2 PROOF OF BLOCH WAVE FUNCTION

We demonstrate the Bloch wave function

$$\psi_{n\vec{k}}(\vec{r} + \vec{R}) = e^{i\vec{k} \cdot \vec{R}} \psi_{n\vec{k}}(\vec{r}) \quad (7.108)$$

must be an energy eigenfunction for an electron traveling through a crystal with a potential periodic in the lattice. Recall from Section 7.2 that a function has the periodicity of the lattice when it is also an eigenfunction of the translation operator according to

$$T_{\vec{R}} f(\vec{r}) = f(\vec{r} + \vec{R}) = f(\vec{r}) \quad (7.109)$$

where $\vec{R} = m_1 \vec{a}_1 + m_2 \vec{a}_2 + m_3 \vec{a}_3$ represents a vector in the direct lattice.

The proof that Equation 7.109 (with the subscripts suppressed) $\psi(\vec{r} + \vec{R}) = e^{i\vec{k} \cdot \vec{R}} \psi(\vec{r})$ is an eigenfunction proceeds similar to that found in Ashcroft and Mermin's book listed in the references for this chapter. We first show any eigenvector ψ of the Hamiltonian with a periodic potential must also be an eigenvector of the translation operator. We then develop results for the translation operator acting on the eigenstates ψ . The Hamiltonian for the crystal must be invariant under translations through the lattice vectors. We therefore expect the eigenfunctions or a linear combination of eigenfunctions to be invariant under the translations at least up to a phase factor. Finally we deduce Equation 7.108.

Step 1: The Hamiltonian and translation operators have the same eigenvectors.

We first show that $[\hat{H}, \hat{T}_{\vec{R}}] = 0$ so that, according to Chapters 3 and 5, the two operators must have the same eigenvectors. Given that the translation operator shifts any function, we have for any function ϕ

$$\hat{T}_{\vec{R}} \hat{H}(\vec{r}) \phi(\vec{r}) = \hat{H}(\vec{r} + \vec{R}) \phi(\vec{r} + \vec{R}) = \hat{H}(\vec{r}) \phi(\vec{r} + \vec{R}) = \hat{H}(\vec{r}) \hat{T}_{\vec{R}} \phi(\vec{r})$$

where we have used the fact the $\hat{H}$ is invariant under lattice translations. Since this last relation holds for any function in the Hilbert space, we conclude $[\hat{H}, \hat{T}_{\vec{R}}] = 0$. Therefore we can expect the eigenfunctions of $\hat{H}$ to also be eigenfunctions of $\hat{T}_{\vec{R}}$. This step provides the link between the results of the translation operator and the results for the eigenfunctions of the Hamiltonian.

Step 2: Product of translation eigenvalues

Let $C(\vec{R})$ be the eigenvalues of the translation operator $\hat{T}_{\vec{R}}$ corresponding to the eigenvectors ψ according to

$$\hat{T}_{\vec{R}} \psi = C(\vec{R}) \psi \quad (7.110)$$

For $\vec{R}$ any vector in the direct lattice (especially a primitive vector), we can easily show

$$C(\vec{R}_1 + \vec{R}_2) = C(\vec{R}_1) C(\vec{R}_2) \quad (7.111)$$

To show the previous relation, consider

$$\hat{T}_{\vec{R}_1} \hat{T}_{\vec{R}_2} \psi(\vec{r}) = \hat{T}_{\vec{R}_1} \psi(\vec{r} + \vec{R}_2) = \psi(\vec{r} + \vec{R}_1 + \vec{R}_2) = \hat{T}_{\vec{R}_1 + \vec{R}_2} \psi(\vec{r}) = \hat{T}_{\vec{R}} \psi(\vec{r}) \quad (7.112)$$

Therefore, using Equation 7.110 in 7.112, we find

$$C(\vec{R}_1)C(\vec{R}_2)\psi(\vec{r}) = C(\vec{R}_2)\hat{T}_{\vec{R}_1} \psi(\vec{r} + \vec{R}_1) = \hat{T}_{\vec{R}_1} \hat{T}_{\vec{R}_2} \psi(\vec{r}) = \hat{T}_{\vec{R}_1 + \vec{R}_2} \psi(\vec{r}) = C(\vec{R}_1 + \vec{R}_2)\psi(\vec{r})$$

Therefore, Equation 7.111 holds.

Step 3: Translation eigenvalues and the primitive vectors

Show that for any direct lattice vector $\vec{R} = n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3$, the eigenvalue corresponding to a translation through $\vec{R}$ must be related to a product of the eigenvalues for a translation through the primitive vectors by

$$C(\vec{R}) = [C(\vec{a}_1)]^{n_1} [C(\vec{a}_2)]^{n_2} [C(\vec{a}_3)]^{n_3} \quad (7.113)$$

where n_i and $\vec{a}_i$ are integers and primitive lattice vectors

This is easy to show using step 2.

$$\begin{aligned} C(\vec{R}) &= C(n_1\vec{a}_1 + n_2\vec{a}_2 + n_3\vec{a}_3) = C\left(\underbrace{\vec{a}_1 + \cdots + \vec{a}_1}_{n_1 \text{ times}} + \underbrace{\vec{a}_2 + \cdots + \vec{a}_2}_{n_2 \text{ times}} + \underbrace{\vec{a}_3 + \cdots + \vec{a}_3}_{n_3 \text{ times}}\right) \\ &= \underbrace{C(\vec{a}_1)C(\vec{a}_1) \cdots C(\vec{a}_1)}_{n_1 \text{ times}} \underbrace{C(\vec{a}_2)C(\vec{a}_2) \cdots C(\vec{a}_2)}_{n_2 \text{ times}} \underbrace{C(\vec{a}_3)C(\vec{a}_3) \cdots C(\vec{a}_3)}_{n_3 \text{ times}} \\ &= [C(\vec{a}_1)]^{n_1} [C(\vec{a}_2)]^{n_2} [C(\vec{a}_3)]^{n_3} \end{aligned}$$

Step 4: Translation eigenvalues as complex numbers

The quantity $C(\vec{a}_i)$ is just a number and might as well be written as a complex number $C(\vec{a}_i) = e^{i2\pi\xi_i}$. It turns out that the magnitude of C must be unity (see next paragraph) so that the ξ_i must be real numbers (and not complex). As a note, if for some reason, the magnitude of C is not equal to one, we can let ξ_i be complex because then $C = e^{-2\pi\text{Im}(\xi_i)} e^{i2\pi\text{Re}(\xi_i)}$ so that the magnitude of C can be adjusted using $e^{-2\pi\text{Im}(\xi_i)}$.

We can see that the numbers $C(\vec{a}_i)$ must have unit magnitude since the translation operator must be unitary. Section 3.14 shows the translation operator can be written as an exponential to get

$$\hat{T}_{\eta} f(x) = e^{-i\eta\hat{p}} f(x) = f(x - \eta)$$

The translation operator must be unitary

$$\hat{T}_{\eta} \hat{T}_{\eta}^{\dagger} = e^{-i\eta\hat{p}} e^{i\eta\hat{p}} = e^{-i\eta\hat{p}} e^{i\eta\hat{p}} = 1$$

since η is real (an x -coordinate) and the momentum $\hat{p}$ is Hermitian. Therefore, if $\hat{T}_{\vec{a}_1} |\Phi\rangle = C(\vec{a}_1) |\Phi\rangle$ then we see

$$\langle \Phi | \Phi \rangle = \langle \Phi | \hat{I} | \Phi \rangle = \langle \Phi | \hat{T}_{\vec{a}_1}^{\dagger} \hat{T}_{\vec{a}_1} | \Phi \rangle = \left[\langle \Phi | \hat{T}_{\vec{a}_1}^{\dagger} \right] \left[\hat{T}_{\vec{a}_1} | \Phi \rangle \right] = \langle \Phi | C^*(\vec{a}_1) C(\vec{a}_1) | \Phi \rangle = |C(\vec{a}_1)|^2 \langle \Phi | \Phi \rangle$$

Canceling the inner product from both sides produces

$$|C(\vec{a}_1)|^2 = 1$$

Step 5: Traveling wave form of the translation eigenvalues

We can now show that $C(\vec{R}) = e^{i\vec{k} \cdot \vec{R}}$ where $\vec{k} = \xi_1 \vec{b}_1 + \xi_2 \vec{b}_2 + \xi_3 \vec{b}_3$, $\vec{R} = n_1 \vec{a}_1 + n_2 \vec{a}_2 + n_3 \vec{a}_3$ and the $\vec{a}_i$ and $\vec{b}_i$ are the direct and reciprocal primitive lattice vectors. It is important to realize that ξ_i are not necessarily integers and therefore $\vec{k}$ is not necessarily a reciprocal lattice vector. In fact there are many more real numbers than integers, so $\vec{k}$ is most often not a reciprocal lattice vector.

Combining the results of steps 3 and 4, we can write

$$C(\vec{R}) = [e^{i2\pi\xi_1}]^{n_1} [e^{i2\pi\xi_2}]^{n_2} [e^{i2\pi\xi_3}]^{n_3} = e^{i2\pi\xi_1 n_1 + i2\pi\xi_2 n_2 + i2\pi\xi_3 n_3} = e^{i\xi_1 \vec{b}_1 \cdot \vec{a}_1 n_1 + i\xi_2 \vec{b}_2 \cdot \vec{a}_2 n_2 + i\xi_3 \vec{b}_3 \cdot \vec{a}_3 n_3}$$

where we have used the fact that $\vec{a}_i \cdot \vec{b}_j = 2\pi\delta_{ij}$. Substituting the definition of $\vec{k}$ and $\vec{R}$, we find the required result

$$C(\vec{R}) = e^{i\vec{k} \cdot \vec{R}}$$

Step 6: Bloch's result.

Substituting the result of step 5 into

$$\hat{T}_{\vec{R}}\psi(\vec{r}) = C(\vec{R})\psi(\vec{r})$$

we find

$$\psi(\vec{r} + \vec{R}) = e^{i\vec{k} \cdot \vec{R}}\psi(\vec{r})$$

7.5.3 ORTHONORMALITY RELATION FOR BLOCH WAVE FUNCTIONS

Now to check the normalization of the Bloch wave functions (refer to Figure 7.30). These wave functions represent a type of plane wave throughout space—recall that a crystal actually has infinite size according to the definition of a lattice, which underlies the definition of the crystal. Therefore, the wave function must be normalized on a finite region of space with volume V that usually comes from periodic boundary conditions over the length L so that $V = L^3$.

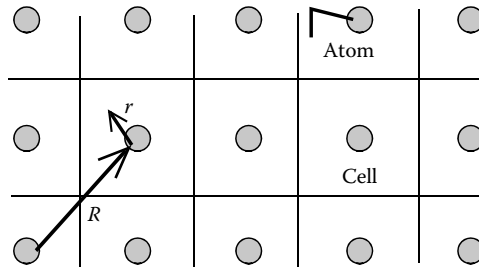


FIGURE 7.30 $\vec{R}$ indicates the center of the cell and $\vec{r}$ ranges over the interior of the cell.

We start with the definition of

$$\phi_{n\vec{k}}(\vec{r}) = \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) \quad (7.114)$$

and explicitly demonstrate the normalization for u . We want to satisfy the orthonormality relation for $|n, \vec{k}\rangle$

$$\delta_{m\vec{k}, n\vec{k}} = \langle m, \vec{k} | n, \vec{k} \rangle = \int_V d^3r \frac{e^{-i\vec{k}\cdot\vec{r}}}{\sqrt{V}} u_{m\vec{k}}^*(\vec{r}) \frac{e^{+i\vec{k}\cdot\vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) \quad (7.115)$$

The orthonormality in $\vec{k}$ mostly comes from the $e^{i\vec{k}\cdot\vec{r}}$ term since the wave vectors $\vec{k}$ and $\vec{k}$ correspond to wavelengths having the size of many unit cells whereas periodic functions $u_{n\vec{k}}(\vec{r})$ have distinct values only within the unit cell. Therefore, we expect u to be relatively independent of $\vec{k}$.

To simplify the calculation, make the substitution of $\vec{r} \rightarrow \vec{R}_j + \vec{r}$ where $\vec{R}_j$ gives the center of unit cell $\#j$ and confine $\vec{r}$ to a unit cell. Note that $u(\vec{r} + \vec{R}_j) = u(\vec{r})$ since u is periodic. This means that the integral in Equation 7.115 can be divided into a summation over each unit cell.

$$\delta_{m\vec{k}, n\vec{k}} = \sum_{j=1}^N \frac{e^{i(\vec{k}-\vec{k})\cdot\vec{R}_j}}{V} \int_{V_j} d^3r e^{i(\vec{k}-\vec{k})\cdot\vec{r}} u_{m\vec{k}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) \quad (7.116)$$

Next note that for electron wavelengths spanning many unit cells, the wave vectors k must have very small magnitude. In fact, $\vec{k} \cdot \vec{r} \sim 2\pi \frac{|\vec{r}|}{\lambda} \sim \frac{a}{L} \sim 0$ since $\vec{r}$ is now confined to a single cell, and “ a ” represents the size of the unit cell and L represents the size of the crystal. Take the exponential $e^{i(\vec{k}-\vec{k})\cdot\vec{r}}$ under the integral to be unity and Equation 7.116 becomes

$$\delta_{m\vec{k}, n\vec{k}} = \sum_{j=1}^N \frac{e^{i(\vec{k}-\vec{k})\cdot\vec{R}_j}}{V} \int_{V_j} d^3r u_{m\vec{k}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) \quad (7.117)$$

First consider different values for the wave vectors. The integral on the right is a constant independent of the particular unit cell (and hence, the subscript j). We are therefore left with the summation of the form $\sum_{j=1}^N e^{i(\vec{k}-\vec{k})\cdot\vec{R}_j}$. If one defines the angle $\theta_j = (\vec{k} - \vec{k}) \cdot \vec{R}_j$ then each term $e^{i(\vec{k}-\vec{k})\cdot\vec{R}_j}$ is a complex number of unit length as indicated in Figure 7.31. Each R_j (and there are many of these—on the order of Avogadro’s number), produces another complex number. Adding all of these complex numbers together will produce a total value of zero since for each complex number, the summation will include its negative. However when $\vec{k} = \vec{k}$, the summation produces the number of unit cells N . As a result we have justified the use of the Kronecker delta function for the wave vectors.

For the Kronecker delta function with subscripts m and n , set $\vec{k} = \vec{k}$ and examine the integral. The functions u are periodic which means that their integral must be independent of the particular unit cell V_j . Therefore, as far as the summation is concerned, the integrals are constants. We have

$$\delta_{m,n} = \frac{1}{V} \sum_{j=1}^N \int_{V_j} d^3r u_{m\vec{k}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) = \frac{N}{V} \int_{V_j} d^3r u_{m\vec{k}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) \quad (7.118)$$

Using the fact that there are N unit cells in the volume V yields $V = NV_{\text{cell}}$ and

$$\int_{V_{\text{cell}}} d^3r u_{m\vec{k}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) = V_{\text{cell}} \delta_{m,n} \quad (7.119)$$

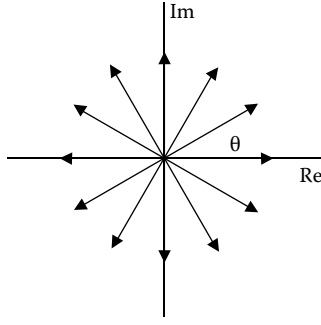


FIGURE 7.31 Sum of the complex numbers produce zero.

We can go further by making an approximation namely that the functions u are approximately independent of k . Then we have

$$\int_{V_{\text{cell}}} d^3 r u_{m\vec{k}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) \cong V_{\text{cell}} \delta_{m,n} \quad (7.120)$$

We assume that for a given $\vec{k}$, the function $u_{n,\vec{k}}$ form a complete set (where n runs over all of the bands). We can see the normalization factor V_{cell} must be correct by using the case of the periodic potential going to zero since then $u \rightarrow 1$ and the integral then produces V_{cell} .

If desired, one can consider changing the normalization of the u to eliminate V_{cell} . Making the replacement

$$u_{n\vec{k}} \rightarrow \sqrt{V_{\text{cell}}} u_{n\vec{k}} \quad (7.121a)$$

we then have the orthonormality relation

$$\int_{V_{\text{cell}}} d^3 r' u_{m\vec{k}}^* u_{n\vec{k}} = \delta_{mn} \quad \text{or} \quad \int_V d^3 r' u_{m\vec{k}}^* u_{n\vec{k}} = \delta_{mn} \quad (7.121b)$$

However, such a normalization is not normally used.

7.6 INTRODUCTION TO EFFECTIVE MASS AND BAND CURRENT

This section explores the dynamics of electrons in energy bands giving rise to current within a semiconductor. A phenomenological argument demonstrates the origin of the effective mass. The effective mass equation for electron dynamics incorporates Newton's laws without explicitly including the forces exerted by the crystal periodic potential. The effective mass can be easily related to the curvature of the band. Some complication arises for three-dimensional (3-D) crystals when the band curvature depends on direction since then a tensor effective mass must be used instead of the scalar one. The discussion next addresses the ability of the bands to support charge transport (current) within a semiconductor material.

7.6.1 MASS, MOMENTUM, AND NEWTON'S SECOND LAW

The electron can be most conveniently pictured as a wave packet moving through the crystal (c.f. Appendix F with a discussion of the superposition of plane waves). The superposition includes

k -states within a narrow range centered on a nonzero wave vector. For the present situation, imagine that a small range of states within the conduction (or valence) enter into a superposition to produce the wave packet. However, keep in mind that these wave vectors k refer to the envelope portion of the Bloch wave function. The applied forces really change the motion of the “modulation” impressed on the “carrier.”

Now one can provide a phenomenological argument to recast Newton’s second law $F = ma$ into one involving only the applied force F and thereby circumvent the complexity introduced by the crystal forces (through the periodic potential). For this simplification, we must use an effective mass for the electron (and hole). Further, we expect the Schrödinger equation to incorporate the effective mass when we neglect the crystal periodic potential because of the close relation between the Heisenberg formulation using time-dependent dynamical variables and the classical Hamiltonian formulation of mechanics.

To begin, represent an electron moving through a crystal by a traveling wave packet (Figure 7.32). We might picture this packet as having a Gaussian-shaped envelope. The group velocity can be written from the dispersion relation $\omega(k)$ by

$$v_g = \frac{\partial \omega}{\partial k} \quad (7.122)$$

which can also be related to the E - k dispersion relation $E(k)$ using $E = \hbar\omega$

$$v_g = \frac{1}{\hbar} \frac{\partial E}{\partial k} \quad (7.123)$$

The group velocity essentially represents the average motion of the components of the wave packet. We assume a very narrow packet (in k -space or ω -space) with k and ω representing average values.

Assume an external agent applies a force F to this electron in the crystal. For example, a battery might be connected across the crystal so as to apply an electric field to the electron. Let x represent the center spatial coordinate of the wave packet. The work done on the electron changes its energy E according to

$$dE = F dx = F \frac{dx}{dt} dt = F v_g dt \quad (7.124)$$

We can solve for the force applied to the wave packet.

$$F = \frac{1}{v_g} \frac{dE}{dt} = \frac{1}{v_g} \frac{dE}{dk} \frac{dk}{dt} \quad (7.125)$$

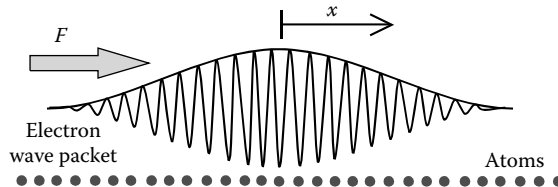


FIGURE 7.32 Electron represented as a wave packet. Applying a force to the electron causes the center of the wave packet to move through distance x .

Substituting Equation 7.123 for the group velocity provides

$$F = \frac{1}{v_g} \frac{dE}{dk} \frac{dk}{dt} = \frac{1}{v_g} \hbar v_g \frac{dk}{dt} = \frac{d(\hbar k)}{dt} \quad (7.126)$$

Therefore we draw two important conclusions from Equation 7.126. First the externally applied force F causes a change in the momentum of the particle. Second, the momentum of the particle is still defined by $p = \hbar k$.

Now we would also like to relate the applied force to the effective mass m_e . The applied force should change the group velocity. We look for a formula very similar to Newton's second law $F = m_e a$ (where m_e represents the effective mass). The rate of change of the speed of the wave packet (modulation) can be determined

$$a = \frac{dv_g}{dt} = \frac{d}{dt} \left(\frac{1}{\hbar} \frac{dE}{dk} \right) = \frac{dk}{dt} \frac{d}{dk} \left(\frac{1}{\hbar} \frac{dE}{dk} \right) = \frac{1}{\hbar} \frac{d^2 E}{dk^2} \frac{dk}{dt} = \frac{1}{\hbar^2} \frac{d^2 E}{dk^2} \frac{d(\hbar k)}{dt} \quad (7.127)$$

However Equation 7.126 already describes the effect of the force F . Therefore Equation 7.127 can be rewritten as

$$F = \left(\frac{1}{\hbar^2} \frac{d^2 E}{dk^2} \right)^{-1} \frac{dv_g}{dt} = m_e a = \frac{d(m_e v_g)}{dt} = \frac{dp}{dt} \quad (7.128)$$

Equation 7.128 provides very important information. First the momentum of the wave packet must be given by

$$p = m_e v_g \quad (7.129)$$

The external force changes the momentum according to

$$F = m_e \frac{dv_g}{dt} \quad (7.130)$$

Most importantly, curvature of the band occupied by the electron provides the effective mass according to

$$m_e = \left(\frac{1}{\hbar^2} \frac{d^2 E}{dk^2} \right)^{-1} \quad (7.131)$$

Near the edge of a band (i.e., the CB minimum or VB maximum of the bands shown in Figure 7.33), the E - k relation is approximately parabolic.

$$E = ck^2 \text{ (reduced zone diagram)}$$

$$E = c \left(k - \frac{G_n}{2} \right)^2 \text{ (extended zone diagram)}$$

where G_n is one of the reciprocal lattice vectors. In either case, the effective mass is given by

$$m_e = \frac{\hbar^2}{2c}$$

Therefore near the bottom of the conduction band or top of the valence band, the electrons have a constant effective mass over a range of k -values. According to Equation 7.131 an upward curvature

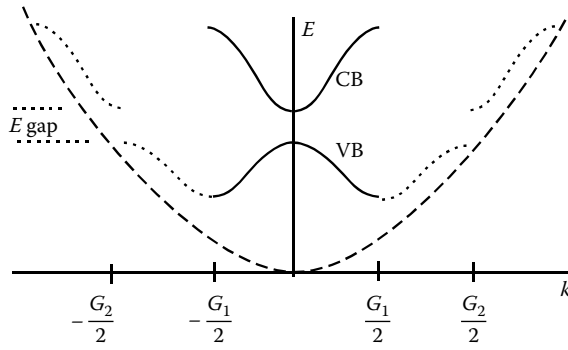


FIGURE 7.33 Reduced and extended zone diagram.

(such as for the conduction band) produces a positive effective mass while a downward curvature (such as for the valence band) produces a negative effective mass.

7.6.2 ELECTRON AND HOLE CURRENT

Now we discuss how conduction band electrons and valence band holes produce current. We start with a 1-D crystal but the extension to 3-D crystals will be easy.

Assume a 1-D crystal has N atoms. For the time being, also assume that each atom has two electrons that it can contribute to a crystal. This actual crystal has finite size as opposed to the crystal required by the mathematical definition. Boundary conditions on finite regions of space produce discrete sets of allowed wave vectors $\{k_i\}$. Recall that these allowed wave vectors occur within the first Brillouin zone (FBZ). The spacing between the allowed vectors must be much smaller than the magnitude of the reciprocal lattice vectors that define the edges of the FBZ. For the present section, the length of the crystal must be $L = Na$ where a represents the lattice spacing (and $2\pi/a$ represents the first reciprocal lattice vector). If we assume periodic boundary conditions on the macroscopic length L , then the allowed wave vectors must be

$$k = \frac{2\pi n}{L} = \frac{2\pi n}{Na} \quad (7.132)$$

(see Sections 6.8 and 7.13). Figure 7.34 shows the largest suitable wavelength satisfying the periodic boundary conditions; the period for the boundary conditions has the same size as the crystal. Sometimes people imagine that the crystal “wraps around” to form a torus. In this case, the periodic boundary conditions force the wave to fit exactly once or an integer number of times around the circumference.

Regardless of the method of visualizing the periodic boundary conditions, we find that each band has exactly N available states. The 25 states shown in either band in Figure 7.35 correspond to 30 atoms each with a valence electron contributed to bonding. This can easily be understood as follows. The spacing between each mode can be found from Equation 7.132 to be

$$\Delta k = \frac{2\pi}{Na} \rightarrow \frac{\# \text{States}}{k\text{-Length}} = \frac{1}{\Delta k} = \frac{Na}{2\pi}$$

The total number of states in the FBZ, which has width $w = 2\pi/a$, must be

$$\text{Number/band} = \frac{\# \text{States}}{k\text{-Length}} * w = \frac{Na}{2\pi} * \frac{2\pi}{a} = N$$

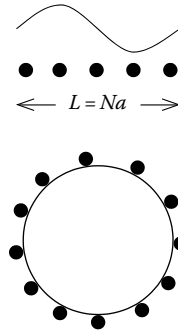


FIGURE 7.34 Top: a multiple number of wavelengths must fit in the distance Na . The wave repeats every Na in distance. Bottom: periodic BCs are sometimes pictured as waves that exactly fit around a circle circumference measuring Na .

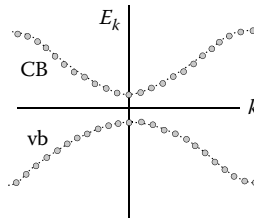


FIGURE 7.35 Each band has N states (ignores electron spin).

For N -atoms with two electrons, the lowest two bands ($2N$ available states) must be completely filled at 0 K. We cite the temperature of 0 K because at the temperature, the electrons cannot absorb enough thermal energy to make a transition across the gap, and the bands must remain full. Of course we assume that there are not any other available forms of energy either (such as light).

Now apply an electric field to the crystal. Normally, with nearly free electrons in the crystal, current would flow. We can show that empty and full bands do not contribute anything to current flow. An empty band does not have any electrons that can flow and therefore does not produce any current (i.e., simple). Now let us consider a full band (say vb in Figure 7.35). Each state in the band corresponds to a velocity. States with wave vector k_i correspond to velocity v_i , and so states with wave vector $-k_i$ corresponds to $-v_i$. Let n be the number of electrons per unit volume ($V=AL$) and let A represent the cross sectional area of the crystal. The current can be written as

$$I = JA = A \sum_{i=1}^N (-e)n_i v_i \tag{7.133}$$

where

n_i represents the number of electrons (per volume) in state i

J represents the current density

Notice that the summation extends over all the states in the band because electrons fill all of the states. The number of electrons n_i (per state per crystal volume) can be written as

$$n_i = \frac{\eta_i}{AL} \tag{7.134}$$

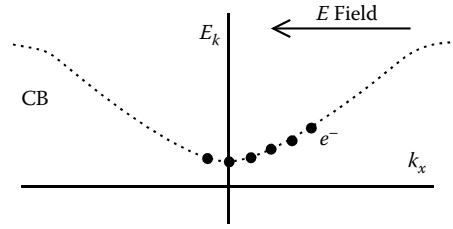


FIGURE 7.36 Electric field shifts electron distribution. Scattering events maintain the steady state.

where we take $\eta_i = 1$ to indicate exactly one electron in the state (the Fermi–Dirac distribution will show that values between 0 and 1 can be obtained for nonzero temperatures). The current can now be written as

$$I = JA = A \sum_{i=1}^N (-e)n_i v_i = \left(\frac{-e}{L}\right) \sum_{i=1}^N v_i = 0 \quad (7.135)$$

since for every state k with speed v_k there exists a state $-k$ with speed $v_{-k} = -v_k$. Therefore, full bands contribute nothing to the current.

Now consider a partially filled conduction band. For a current provided by Equation 7.135, where the summation extends over only the filled states, one might erroneously find zero current by thinking that for every electron moving with $+v$, there exists another moving with $-v$. What's wrong?!? The answer: we have not applied an electric field! Applying an electric field should cause the wave vector to change according to Equation 7.126. At any given instant of time, the bands should be occupied similar to Figure 7.36. Now summing over all occupied states produces the result.

$$I = \left(\frac{-e}{L}\right) \sum_{i \text{ filled}} v_i \neq 0 \quad (7.136)$$

Of course Figure 7.36 exaggerates the shift in k ; the shift does not need to be all that large to produce significant current. Of importance for the discussion of holes below, the summation over all states in a band (including both filled and empty) produces zero as given by

$$\sum_{\text{all states in a band}} v_i = 0 \quad (7.137)$$

At this point, we make a comment on Equation 7.126, rewritten as

$$k(t) = \frac{1}{\hbar} \int dt F \quad (7.138)$$

According to this equation, the particle wave vector should grow without bound so long as the applied force continues to act on the particle. This would require the charge distribution for a partially filled band, such as in Figure 7.36, to continue moving toward the right (larger k) leaving behind more empty states near $k=0$. However, Equations 7.126 and 7.138 do not account for mobility-limiting collisions. Once we include the collisions in these two equations, the distribution will reach the steady state configuration depicted in Figure 7.36.

Equation 7.138 must apply to the full band as well and the effects of collisions must also be included. Consider the full band as the conduction approaches steady state. Electrons moving to

larger values of k leave behind empty states that can be filled by the electrons with smaller k . At the right-hand Brillouin zone edge, an electron must move past the edge toward the right. All the electrons in the band must shift to states with larger k . At the left-hand Brillouin zone edge, an electron from a state with smaller k shifts into the vacant state at that Brillouin zone edge. The process can be equivalently thought of as requiring an electron passing through the right-hand Brillouin zone boundary to reappear at the left-boundary. In this way, the band remains full and cannot conduct current.

Finally, consider a partially empty valence band. We want to find the current due to the remaining electrons in the valence band (a field must be applied!). Similar to previous equations we can write

$$I_e = \left(\frac{-e}{L}\right) \sum_{e^- \text{ in vb}} v_i \quad (7.139)$$

However, this can be related to the motion of holes. The holes in the valence band occur either because the atoms in the crystal absorb energy to promote electrons to the conduction band or the semiconductor has p-type doping. Either way, we know the number of holes and consequently, the number of electrons remaining in the band ($\#e^- = N - \#h^+$). For Equation 7.139, the total number of states can be divided into filled and empty. First in view of Equation 7.137, one can write a current associated with the entire vb as

$$I = \left(\frac{-e}{L}\right) \sum_{\text{all vb states}}^N v_i = 0$$

where note that the summation extends over all vb states and not only the filled ones. Now divide the summation into two summations over filled and empty states.

$$0 = \left(\frac{-e}{L}\right) \sum_{\text{all states}}^N v_i = \left(\frac{-e}{L}\right) \sum_{e^- \text{ in vb}} v_i + \left(\frac{-e}{L}\right) \sum_{\text{empty}} v_i \Rightarrow \left(\frac{-e}{L}\right) \sum_{e^- \text{ in vb}} v_i = \left(\frac{+e}{L}\right) \sum_{\text{empty}} v_i \quad (7.140)$$

Therefore, the current flow in the partially filled band can be attributed to either the motion of electrons or the motion of holes. The hole current comes from the summation over the “empty” states. Equation 7.140 shows that the holes behave as though they have positive charge ($+e$). Also, we should note that electrons have negative effective mass in the vb whereas holes have positive effective mass in the vb. Applying an electric field causes holes to move in the opposite direction from the electrons because the hole acts as a positive charge.

7.7 3-D BAND DIAGRAMS AND TENSOR EFFECTIVE MASS

Band diagrams characterize the effect of the crystal geometry on the behavior of the electrons within the semiconductor. This section discusses 3-D bands and the tensor form of the effective mass. The band-edge diagrams that plot energy versus position provide convenient pictures for device operation such as for diodes. Later sections show the band-edge diagrams owe their existence to the fact that electrons and holes occupy very narrow ranges of energy near the lower and upper edges of the band, respectively. The effective density of states approximation allows us to essentially reduce dispersion curves to two discrete levels.

7.7.1 E - k DIAGRAMS FOR 3-D CRYSTALS

Our work so far with bands has shown the electron energy E plotted against the wave vector k . Both positive and negative k -values appear on the same plot (also see Section 1.3 in Chapter 1). Actual

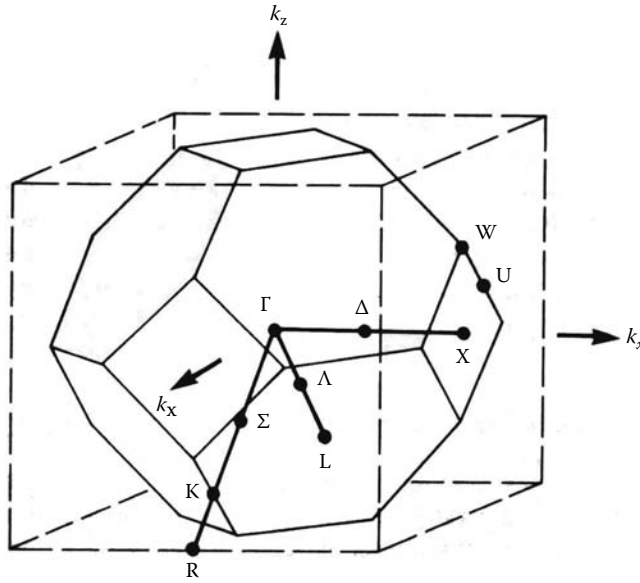


FIGURE 7.37 Zinc blende FBZ. (From Blakemore, J.S., *J. Appl. Phys.*, 53, R123, 1982. With permission.)

band diagrams do not show the negative k -values because bands exhibit symmetry in k and there is no need to show redundant information. Instead, actual diagrams show the bands along two different directions.

Figure 7.37 shows the FBZ for materials with the zinc blende crystal structure. The Γ point sits at the center of the zone; that is, the Γ point corresponds to $k = 0$ for the maximum of the valence band. The line $\Gamma \rightarrow X$ represents wave vectors along the $\langle 100 \rangle$ direction (easy to remember: x stands for the x -direction). The line $\Gamma \rightarrow L$ (L for diagonaL) represents the $\langle 111 \rangle$ direction. Figure 7.38 shows the band diagram for GaAs which crystallizes in the zinc blende configuration. As mentioned, the horizontal axis represents two different directions in this type of diagram. The band diagrams in Figure 7.38 do not show the $-k$ direction because the bands would look the same as the $+k$ direction. The $-k$ looks the same as $+k$ because the lattice has inversion symmetry (if R is a lattice vector then so is $-R$).

Figure 7.38 shows a direct bandgap at Γ . The minimum in the conduction band and maximum in the valence band look fairly symmetrical about the Γ point for small values of k . Notice the formation of L and X valleys in the conduction band. Under certain conditions, electrons can scatter from the Γ valley into these other valleys. The group velocity and effective mass of the electron in these other valleys must be different from that in the Γ valley. We expect any electrons scattered into these side valleys to decay back to the Γ minimum after a period of time. Actually, because the L valley has low energy for GaAs, a significant number of GaAs electrons can have sufficient thermal energy to populate the L valley.

The valence band structure consists of the heavy-hole, light hole, and split-off bands. The heavy-hole band gives rise to larger effective masses for the holes than does the light-hole band. The split-off band gives roughly the same effective mass as does the light-hole band owing to the somewhat similar curvatures. The diagram shows nearly degenerate light- and heavy-hole bands near the Γ point; that is, they have roughly the same energy at $\Gamma = 0$. Both the heavy and light-hole bands can contribute to current flow and absorption/emission processes. As a point of interest, people sometimes add strain (i.e., strain the lattice—a force applied to the atoms in the lattice) to the GaAs

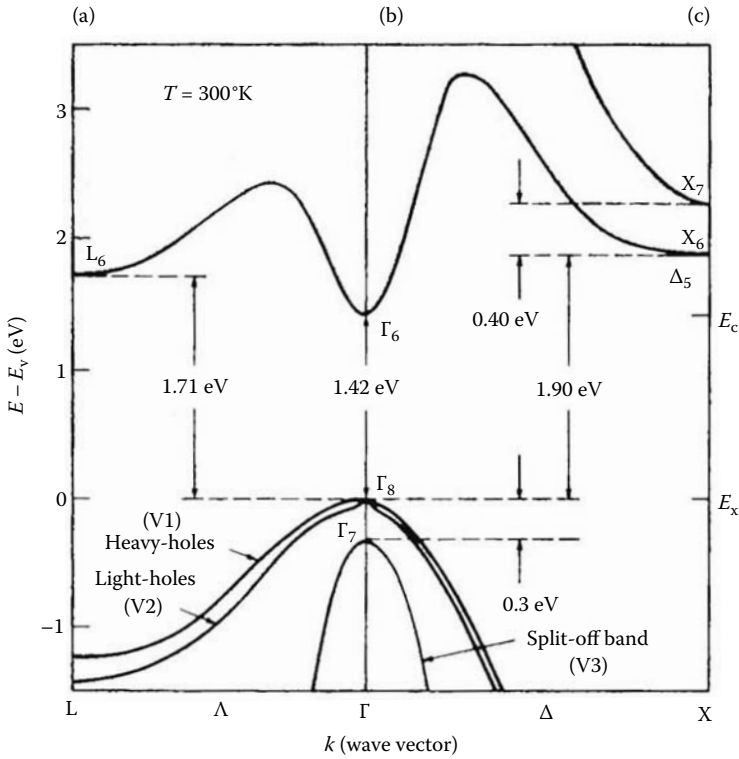


FIGURE 7.38 GaAs band diagram. (From Blakemore, J.S., *J. Appl. Phys.*, 53, R123, 1982. With permission.)

lattice by adding Indium. The band-edge shifts to longer wavelengths. More importantly, devices can be made more efficient because the curvature of the heavy-hole valence band can be made the same as the curvature of the conduction band. Also, the light-hole band moves further away (in energy) from the heavy-hole band (and no longer necessarily participates in optical and electronic processes).

7.7.2 EFFECTIVE MASS FOR THREE-DIMENSIONAL BAND STRUCTURE

Before discussing the 3-D case, let us first discuss the 1-D case. A 1-D crystal with a direct bandgap (for example, a 1-D version of GaAs) has a dispersion relation (E vs. k) of the form

$$E - E_c = \frac{\hbar^2 k_x^2}{2m^*} \quad (7.141)$$

near the bottom of the band. In general, we might find a conduction band for a direct-bandgap semiconductor having the form

$$E - E_c = A k_x^2$$

Indirect bandgaps have dispersion relations for the conduction band (near the minimum) of the form

$$E - E_c = A(k_x - k_{ox})^2 \quad (7.142)$$

where k_{ox} gives the center of the parabola in k -space. Obviously, the direct bandgap dispersion relation comes from setting $k_{ox} = 0$. We see that two wave vectors give the same energy E according to

$$k_x = k_{ox} \pm \sqrt{\frac{\hbar^2}{2m^*}(E - E_c)} \quad (7.143)$$

We can calculate the effective mass using the relation from Section 7.6

$$\frac{1}{m^*} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_x^2} \quad (7.144)$$

Taking the second derivative of Equation 7.143, we find

$$\frac{1}{m^*} = \frac{2A}{\hbar^2} \quad (7.145)$$

The effective mass in Equation 7.145 does not depend on the exact location of the center k_{ox} because of the parabolic form of the dispersion curve and the fact that we take two derivatives. The effective mass does not depend on whether $k < k_{ox}$ or $k > k_{ox}$ near the extremum point. Therefore, the rate of change of the group velocity due to an applied force F

$$\frac{dv}{dt} = \frac{1}{m^*} F \quad (7.146)$$

For a two-dimensional (2-D) crystal, we expect the dispersion curve to have the form

$$E - E_c = \frac{\hbar^2 k^2}{2m^*} = \frac{\hbar^2}{2m^*} [(k_x - k_{ox})^2 - (k_y - k_{oy})^2] \quad (7.147)$$

near the extremum point where k_{ox}, k_{oy} denote the center of the paraboloid in k -space (see Figure 7.39). Circles describe a level contour over which $E - E_c$ has a fixed value. However, Equation 7.147 describes a crystal with symmetric dispersion curves since it has a constant effective mass (independent of k -direction) as a coefficient. One would need different coefficients for the two terms in brackets for nonsymmetrical bands. Rather than discuss the general 2-D case, let us go on to the 3-D one.

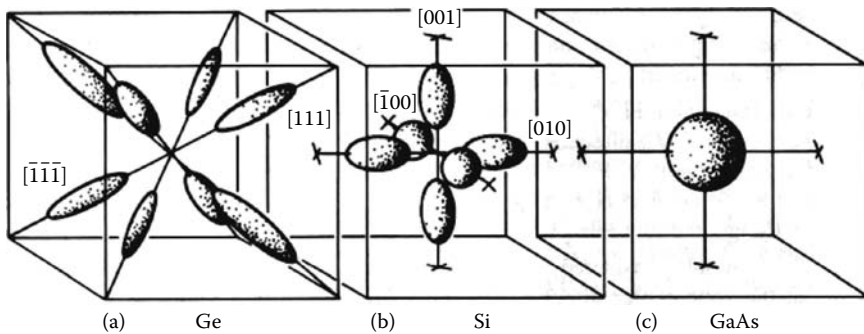


FIGURE 7.39 Constant energy surfaces for (a) Ge, (b) Si, and (c) GaAs. (From Sze, S.M., *Physics of Semiconductor Devices*, 2nd edn., John Wiley & Sons, New York, 1981. With permission.)

The 3-D crystal has a dispersion relation of the form (near the minimum)

$$E = E_c + A(k_x - k_{ox})^2 + B(k_y - k_{oy})^2 + C(k_z - k_{oz})^2 \quad (7.148)$$

where the symbol E_c represents the extremum value of the band, and k_{ox} , k_{oy} , k_{oz} give the location (in k -space) of the extremum. The surface has the form of an ellipsoid for surfaces of constant energy. If $A = B = C$ then the surface must be a sphere with effective mass independent of direction. Consider the case of nonsymmetric bands. We would find the effective mass for motion along the three directions have the form

$$(m_x)^{-1} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_x^2} = \frac{2A}{\hbar^2}, \quad (m_y)^{-1} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_y^2} = \frac{2B}{\hbar^2}, \quad (m_z)^{-1} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_z^2} = \frac{2C}{\hbar^2} \quad (7.149)$$

Apparently the mass and acceleration depend on the direction of an applied force. Depending on the shape of the energy bands, the effective mass can be larger along one direction than another. We need to discuss the effective mass for more than three directions since the particle can move in any direction through the crystal.

It seems strange to talk about the “effective mass in a given direction” since mass should be a scalar quantity measuring inertia. First of all, we really mean that the effective mass depends on wave vector k because the band curvature depends on k . To say that the mass depends on the direction of travel just means that it depends on the values of k_x , k_y , k_z that determine the direction of travel. That dependence on direction must be ultimately related to the form of the periodic potential. Next, one should note that the effective mass really provides a proportionality constant between the applied force and the acceleration. Therefore, saying the effective mass depends on the direction of motion really means that the relation between the applied force and the acceleration depends on the direction of the applied force.

Apparently the relation between force and acceleration in a crystal should be written as a tensor product. Some people write Newton’s second law in the form of a dyadic equation

$$\vec{F} = \overleftrightarrow{m} * \vec{a} \quad (7.150)$$

The mass in this case is a dyad which represents a tensor. Let us leave off the asterisk $*$ for convenience. Technically a dyad like m can be written as

$$\overleftrightarrow{m} = m_{xx}\tilde{x}\tilde{x} + m_{xy}\tilde{x}\tilde{y} + \cdots$$

where $\tilde{x}, \tilde{y}, \tilde{z}$ represent basis vectors. Please refer to Section 3.16 for a review of dyads. The dyad provides a representation for a second rank tensor, both of which can be represented by a 3×3 matrix.

$$\begin{pmatrix} m_{xx} & m_{xy} & m_{xz} \\ m_{yx} & m_{yy} & m_{yz} \\ m_{zx} & m_{zy} & m_{zz} \end{pmatrix}$$

We can now demonstrate the 3-D effective mass and its relation to the band curvature that has the form

$$\overleftrightarrow{m}^{-1} = \frac{1}{\hbar^2} \nabla_k \nabla_k E = \frac{1}{\hbar^2} \left(\tilde{x} \frac{\partial}{\partial k_x} + \tilde{y} \frac{\partial}{\partial k_y} + \tilde{z} \frac{\partial}{\partial k_z} \right) \left(\tilde{x} \frac{\partial}{\partial k_x} + \tilde{y} \frac{\partial}{\partial k_y} + \tilde{z} \frac{\partial}{\partial k_z} \right) E \quad (7.151)$$

where the gradient in k -space ∇_k has the form

$$\nabla_k = \tilde{x} \frac{\partial}{\partial k_x} + \tilde{y} \frac{\partial}{\partial k_y} + \tilde{z} \frac{\partial}{\partial k_z}$$

and neither the dot nor cross product appears between the two gradients in Equation 7.151 which can be written as matrix elements

$$(m^{-1})_{ij} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_i \partial k_j} \quad (7.152)$$

To demonstrate Equation 7.151, let $\vec{F}$ be an arbitrary applied force. The energy supplied by the applied force can be written as

$$dE = \vec{F} \cdot d\vec{r} = \vec{F} \cdot \frac{d\vec{r}}{dt} dt = \vec{F} \cdot \vec{v}_g dt \quad (7.153a)$$

where the vector form of the group velocity can be written as

$$\vec{v}_g = \nabla_k \omega = \frac{1}{\hbar} \nabla_k E \quad (7.153b)$$

Therefore the rate of change of particle energy must be

$$\frac{dE}{dt} = \vec{F} \cdot \frac{1}{\hbar} \nabla_k E \quad (7.153c)$$

Now working with Newton's second law using the dyadic notation for the effective mass

$$\vec{F} = \overleftrightarrow{m} \cdot \vec{a} = \overleftrightarrow{m} \cdot \frac{d\vec{v}_g}{dt} = \overleftrightarrow{m} \cdot \frac{d}{dt} \frac{1}{\hbar} \nabla_k E \quad (7.154a)$$

However, the change in a function G with k can be written as

$$dG = (\nabla_k G) \cdot d\vec{k} \quad (7.154b)$$

In our case, taking the function G to be

$$G = \frac{1}{\hbar} \nabla_k E \quad (7.154c)$$

Therefore Equation 7.154a becomes

$$\begin{aligned} \vec{F} &= \overleftrightarrow{m} \cdot \frac{d\vec{G}}{dt} = \overleftrightarrow{m} \cdot \frac{1}{dt} d\vec{G} = \overleftrightarrow{m} \cdot \nabla_k \vec{G} \cdot d\vec{k} \frac{1}{dt} = \overleftrightarrow{m} \cdot \frac{1}{\hbar} \nabla_k \nabla_k E \cdot \frac{d\vec{k}}{dt} = \overleftrightarrow{m} \cdot \frac{1}{\hbar^2} \nabla_k \nabla_k E \cdot \frac{d\vec{p}}{dt} \\ &= \left(\overleftrightarrow{m} \cdot \frac{1}{\hbar^2} \nabla_k \nabla_k E \right) \cdot \vec{F} \end{aligned}$$

For arbitrary $\vec{F}$ we conclude that the operator gives

$$\vec{1} = \vec{m} \cdot \frac{1}{\hbar^2} \nabla_k \nabla_k E \quad (7.155a)$$

Therefore, as discussed in Section 3.16, we surmise

$$\vec{m}^{-1} = \frac{1}{\hbar^2} \nabla_k \nabla_k E \quad (7.155b)$$

as required for the demonstration. The examples below show how to calculate the effective mass.

An average effective mass often appears in formulas such as for the density of states. The average usually appears as a geometric average such as $\langle m \rangle = (m_1 m_2 m_3)^{1/3}$.

Example 7.5

As an example if $\vec{a} = a\hat{x}$ then $\vec{F} = \tilde{x}m_{xx}a + \tilde{y}m_{yx}a + \tilde{z}m_{zx}a$

Example 7.6

Find the effective mass m_{ij} for the isotropic band

$$E = A\hbar^2 k^2 = A\hbar^2 (k_x^2 + k_y^2 + k_z^2)$$

SOLUTION

Using Equation 7.152, namely $(m^{-1})_{ij} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_i \partial k_j}$ one finds $(m^{-1})_{ij} = 2A\delta_{ij}$. Therefore the effective mass must be $m = 1/2A$, independent of direction.

Example 7.7

Find the effective mass m_{ij} for the band

$$E = \hbar^2 (Ak_x^2 + Bk_y^2 + Ck_z^2)$$

SOLUTION

We use Equation 7.152

$$(m^{-1})_{ij} = \frac{1}{\hbar^2} \frac{\partial^2 E}{\partial k_i \partial k_j}$$

to find $m_{11}^{-1} = 2A$, $m_{22}^{-1} = 2B$, $m_{33}^{-1} = 2C$ and the others are zero. The inverse effective mass matrix and the effective mass matrix must be

$$\underline{m}^{-1} = \begin{pmatrix} 2A & 0 & 0 \\ 0 & 2B & 0 \\ 0 & 0 & 2C \end{pmatrix} \Rightarrow \underline{m} = \begin{pmatrix} \frac{1}{2A} & 0 & 0 \\ 0 & \frac{1}{2B} & 0 \\ 0 & 0 & \frac{1}{2C} \end{pmatrix} = \begin{pmatrix} m_x & 0 & 0 \\ 0 & m_y & 0 \\ 0 & 0 & m_z \end{pmatrix}$$

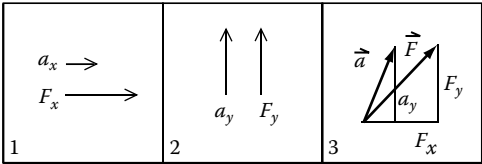


FIGURE 7.40 Although acceleration and force are linearly related for each direction, the vector force and acceleration are not parallel when the effective mass depends on direction of motion.

Example 7.8

Using the last example, show the relation between force and acceleration.

SOLUTION

Consider just a 2-D case for simplicity. We have $\vec{F} = \overleftrightarrow{m} \cdot \vec{a}$ or equivalently

$$\begin{pmatrix} F_x \\ F_y \end{pmatrix} = \begin{pmatrix} m_x & 0 \\ 0 & m_y \end{pmatrix} \begin{pmatrix} a_x \\ a_y \end{pmatrix}$$

The linear relations between the force and acceleration become

$$F_x = m_x a_x \quad \text{and} \quad F_y = m_y a_y$$

Because the effective mass can be different for motion along different directions, identical forces can produce two different accelerations as illustrated in the left two panes of Figure 7.40. We can combine these two panes to produce the vector diagram in the third pane. Notice the force and acceleration vectors are no longer parallel.

7.7.3 INTRODUCTION TO BAND-EDGE DIAGRAMS

The band-edge diagrams (spatial diagrams) can be found from the normal E - k band diagrams (dispersion curves). Recall that a dispersion curve has axes of E versus k and does not give any indication or information on how the energy depends on the position variable x . In fact, there must exist one dispersion curve for each value of x (we assume just one spatial dimension) in the material. We group the states near the bottom of the E - k conduction band together to form the conduction band c for the band-edge diagram (see Figure 7.41). Similarly, we group the top-most hole states in the E - k diagram to produce v for the band-edge diagram.

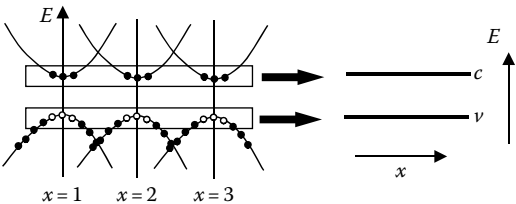


FIGURE 7.41 The states within an energy kT of the bottom of the conduction band or the top of the valence band form the levels in the band-edge diagram.

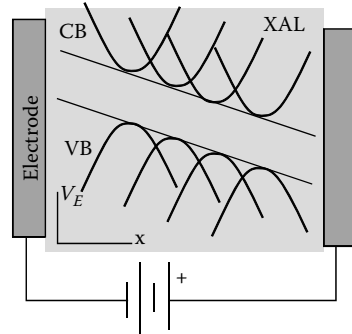


FIGURE 7.42 Band bending between parallel plates connected to a battery.

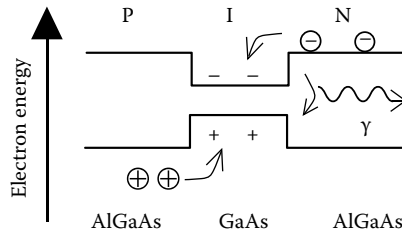


FIGURE 7.43 Band-edge diagram for heterostructure with a single quantum well.

Now consider band bending. Imagine a semiconductor material embedded between two electrodes, which are attached to a battery as shown in Figure 7.42. The electric field points from right to left inside the material. An electron placed inside the material would move toward the right under the action of the electric field. We must add energy to move an electron closer to the left-hand electrode (since it is negatively charged and naturally repels electrons). This means that all electrons have higher energy near the left-hand electrode and lower energy near the right-hand electrode.

For the situation depicted in Figure 7.42, all of the electrons have higher energy near the left-hand electrode. The term “all electrons” refers to conduction and valence band electrons. This means that near the left electrode, the E - P diagrams must be shifted upward to higher energy levels. Once again grouping the states at the bottom of the conduction bands across the regions, we find a band edge. Similarly, the tops of the valence bands produce the valence band-edge diagram. So to say that the CB (for example) bends, we are actually saying that the dispersion curves are displaced in energy for each adjacent point in x . By the way, we will see that the entire conduction band can be represented by the thin line representing the conduction band edge by using the effective density of states approximation—more on this later. Now we see that the electric field between the plates causes the electron energy to be larger on the left and smaller on the right. An electron placed in the crystal moves to the right to achieve the lowest possible energy. Stated equivalently, the electron moves opposite to the electric field toward the right-hand plate.

The band-edge diagrams allow one to understand a large number of optoelectronic components such as PIN photodetectors and semiconductor lasers. Figure 7.43 shows an example of heterostructure for a quantum well laser or LED. Electrons drop into the conduction band (CB) well and holes drop into the valence band (vb) well. These carriers can recombine and produce photons. The GaAs forms the well region while AlGaAs forms the barriers owing to its larger bandgap.

7.8 KRONIG–PENNEY MODEL FOR NEARLY FREE ELECTRONS

The Kronig–Penney model predicts band structure and effective mass for electrons and holes in a semiconductor as a result of a periodic potential. The model approximates the actual electrostatic potential with a series of square wells and square barriers.

7.8.1 MODEL

Figure 7.44 shows the electrostatic potential energy $V(x)$ of the electron in the crystal due to the atomic cores with lattice constant a . The Kronig–Penney model approximates the atomic potential energy curves with a series of wells and barriers as shown. Near the position of the atoms, the Kronig–Penney potential forms the bottom of a quantum well with the minimum value of $V=0$ and width ξ . The barriers separating the wells have height V_0 and width η . The lattice constant must then have the value $a = \xi + \eta$. We see the vector $a\hat{z}$ must be a direct lattice vector. For now, assume the energy of the electron E is larger than the barrier height V_0 .

Before continuing, we should discuss the overall goal of the model. As usual we want to find the allowed energy eigenvalues E and the corresponding energy eigenfunctions $|n, k\rangle$ such that $\hat{H}|n, k\rangle = E_{nk}|n, k\rangle$. The full solution to the time-dependent Schrödinger wave equation then has the form

$$|\Psi(t)\rangle = \sum_{n,k} \beta_{n,k}(t) |n, k\rangle \text{ or equivalently } \Psi(x, t) = \sum_{n,k} \beta_{n,k}(t) \psi_{n,k}(x) \quad (7.156)$$

The Sturm–Liouville problem and the associated boundary conditions lead to quantized wave vectors k and energy values E , and it leads to the dispersion relation $E = E(k)$. For the Kronig–Penney model, we let the wave vector k correspond to waves spanning many unit cells (i.e., the wavelength must be much larger than the lattice constant a). We use the Bloch eigenfunctions $|n, k\rangle \rightarrow \psi_{n,k}(x) = e^{ikx} u_{n,k}(x)$ as solutions to the time-independent Schrödinger equation in order to solve the Sturm–Liouville problem. However, examining the topology of the potential, we see that different regions (well and barrier) lead to different forms of the functions u similar to the finitely deep well in Chapter 5. Each region of the well has an associated wave vector κ (different from k) related to the difference between the potential V and the energy E of the electron. Therefore, the eigenfunctions must be specified in parts according to

$$\psi_{n,k}(x) = \begin{cases} \psi_{n,k,[w\xi]} & \text{wells} \\ \psi_{n,k,[b\eta]} & \text{barriers} \end{cases} \text{ or dropping the } n, k \psi(x) = \begin{cases} \psi_{[w\xi]} & \text{wells} \\ \psi_{[b\eta]} & \text{barriers} \end{cases}$$

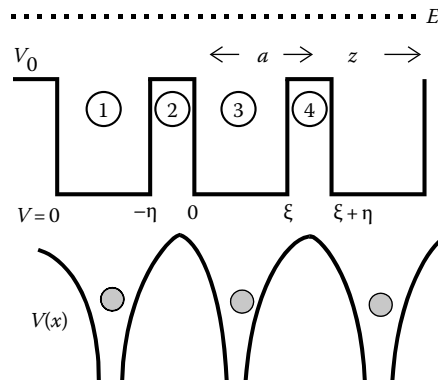


FIGURE 7.44 Periodic potential. Minimum corresponds to location of atom.

The new subscripts $[w\xi]$ and $[b\eta]$ do not refer to the Bloch wave function. We include these two subscripts as a convenient reminder that “w” refers to the well having width ξ and “b” refers to the barrier having width η . They do not refer to two sequences of values but they do label the region of space. In order to find the eigenfunctions and dispersion curves, we must match boundary conditions across the interfaces in Figure 7.44. The eigenfunctions for each region will be the sum of sines and cosines and we will need to find four coefficients A, B, C, D . In the process, we will find the dispersion curve $E = E(k)$. We still want to know the allowed quantized energy values and the eigenfunctions. For this, we would need to specify macroscopic boundary conditions that determine the allowed k and therefore the allowed E through the dispersion relations. We realize that the gaps in the dispersion curve appear very similar to those developed for the phonon curves.

Now proceed to solving the Sturm–Liouville problem for an electron in the periodic potential $V(z)$. Schrödinger’s time-independent equation can be written as

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(z)}{\partial z^2} + V(z)\psi(z) = E\psi(z) \quad (7.157)$$

where m represents the free electron mass (and not the effective mass). For $E > V_0$, we expect plane wave solutions (i.e., sines and cosines) for all regions of space. The solutions for the barrier regions must be

$$\psi_{[b\eta]} = C \sin(\kappa_{[b\eta]}z) + D \cos(\kappa_{[b\eta]}z) \quad (7.158)$$

where the wave vector for the barrier region is given by

$$\kappa_{[b\eta]} = \sqrt{\frac{2m}{\hbar^2}(E - V_0)} \quad (7.159)$$

The subscripts in Equations 7.158 and 7.159 do not refer to the Bloch wave function. We include two subscripts as a convenient reminder that “b” refers to the barrier and “ η ” refers to its width. They do not refer to two sequences of values but they do label the region of space. Assume the solutions in the vicinity of atoms (i.e., for all regions similar to $0 < z < \xi$, the wells) have the form

$$\psi_{[w\xi]} = A \sin(\kappa_{[w\xi]}z) + B \cos(\kappa_{[w\xi]}z) \quad (7.160)$$

where the wave vector is given by

$$\kappa_{[w\xi]} = \sqrt{\frac{2m}{\hbar^2}E} \quad (7.161)$$

Again, notice that the subscripts in Equations 7.160 and 7.161 have been chosen to remind the reader of “w” for well and “ ξ ” for the width of the well. As an important note, notice that the wave vectors $\kappa_{[w\xi]}$, $\kappa_{[b\eta]}$ refer to subregions of the unit cell over the distance $(0, a)$ whereas k in Bloch’s theorem (Equation 7.156) refers to a wavelength that can span many unit cells.

We must subject the solutions in Equations 7.158 and 7.160 to boundary conditions. The interface between regions 2 and 3 provides

$$\begin{aligned} \psi_{[b\eta]}(0) &= \psi_{[w\xi]}(0) \\ \left. \frac{\partial \psi_{[b\eta]}(z)}{\partial z} \right|_{z=0} &= \left. \frac{\partial \psi_{[w\xi]}(z)}{\partial z} \right|_{z=0} \end{aligned} \quad (7.162)$$

Next consider the interface between regions 3 and 4 at $z=\xi$. The two remaining boundary conditions take into account both continuity and periodicity. First, the wave functions must be continuous across the barrier

$$\psi_{[w\xi]}(\xi) = \psi_{[b\eta]}(\xi) \quad (7.163)$$

Using the direct lattice vector $\vec{R} = -a\hat{z}$, Bloch's relation $\psi(\vec{r} + \vec{R}) = e^{i\vec{k}\cdot\vec{R}}\psi(\vec{r})$ can be written as $\psi(z-a) = e^{-ika}\psi(z)$. *Note:* this is where the wave vector k makes its appearance in preparation for Equation 7.168 below. Therefore $\psi_{[b\eta]}(z)$ in Equation 7.163 can be written as $\psi_{[b\eta]}(z-a) = e^{-ika}\psi_{[b\eta]}(z)$ or, upon substituting $z=\xi$ we find

$$\psi_{[b\eta]}(\xi - a) = e^{-ika}\psi_{[b\eta]}(\xi) \Rightarrow \psi_{[b\eta]}(\xi) = e^{ika}\psi_{[b\eta]}(\xi - a) \quad (7.164)$$

Combining Equations 7.164 and 7.163 and using $-\eta = \xi - a$, we find

$$\begin{aligned} \psi_{[w\xi]}(\xi) &= e^{ika}\psi_{[b\eta]}(-\eta) \\ \left. \frac{\partial \psi_{[w\xi]}(z)}{\partial z} \right|_{z=\xi} &= e^{ika} \left. \frac{\partial \psi_{[b\eta]}(z)}{\partial z} \right|_{z=-\eta} \end{aligned} \quad (7.165)$$

In summary, the four boundary conditions are

$$\begin{aligned} \psi_{[w\xi]}(0) &= \psi_{[b\eta]}(0) & \psi_{[w\xi]}(\xi) &= e^{ika}\psi_{[b\eta]}(-\eta) \\ \left. \frac{\partial \psi_{[w\xi]}(z)}{\partial z} \right|_{z=0} &= \left. \frac{\partial \psi_{[b\eta]}(z)}{\partial z} \right|_{z=0} & \left. \frac{\partial \psi_{[w\xi]}(z)}{\partial z} \right|_{z=\xi} &= e^{ika} \left. \frac{\partial \psi_{[b\eta]}(z)}{\partial z} \right|_{z=-\eta} \end{aligned} \quad (7.166)$$

Now we apply the boundary conditions. Substituting Equations 7.158 and 7.160, namely

$$\psi_{[b\eta]} = C \sin(\kappa_{[b\eta]}z) + D \cos(\kappa_{[b\eta]}z) \quad \text{and} \quad \psi_{[w\xi]} = A \sin(\kappa_{[w\xi]}z) + B \cos(\kappa_{[w\xi]}z)$$

into the boundary conditions provides

$$\begin{aligned} B &= D \quad \text{and} \quad A\kappa_{[w\xi]} = C\kappa_{[b\eta]} \\ A \sin(\kappa_{[w\xi]}\xi) + B \cos(\kappa_{[w\xi]}\xi) &= e^{ika}[-C \sin(\kappa_{[b\eta]}\eta) + D \cos(\kappa_{[b\eta]}\eta)] \\ A\kappa_{[w\xi]} \cos(\kappa_{[w\xi]}\xi) - B\kappa_{[w\xi]} \sin(\kappa_{[w\xi]}\xi) &= e^{ika}[C\kappa_{[b\eta]} \cos(\kappa_{[b\eta]}\eta) + D\kappa_{[b\eta]} \sin(\kappa_{[b\eta]}\eta)] \end{aligned}$$

Eliminating C and D in the last two equations provides two simultaneous equations

$$\left[\sin(\kappa_{[w\xi]}\xi) + \frac{\kappa_{[w\xi]}}{\kappa_{[b\eta]}} e^{ika} \sin(\kappa_{[b\eta]}\eta) \right] A + [\cos(\kappa_{[w\xi]}\xi) - e^{ika} \cos(\kappa_{[b\eta]}\eta)] B = 0 \quad (7.167a)$$

$$[\kappa_{[w\xi]} \cos(\kappa_{[w\xi]}\xi) - \kappa_{[w\xi]} e^{ika} \cos(\kappa_{[b\eta]}\eta)] A + [-\kappa_{[w\xi]} \sin(\kappa_{[w\xi]}\xi) - \kappa_{[b\eta]} e^{ika} \sin(\kappa_{[b\eta]}\eta)] B = 0 \quad (7.167b)$$

Next solve the simultaneous Equation 7.167a and b. We have a set of equations of the form

$$\underline{M} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We want nontrivial solutions for the coefficients A and B . If the matrix $\underline{M}$ can be inverted then A and B can be uniquely determined to be zero. Therefore, we want the matrix $\underline{M}$ to be noninvertible so that A and B can assume nonzero values. This requires that the determinant of the matrix $\underline{M}$ must be zero $\det \underline{M} = 0$. The requirements on the determinant of $\underline{M}$ determines the dispersion relation $E = E(k)$ and the ranges of forbidden energy. Applying the determinant condition to Equation 7.167a and b provides

$$\left[\sin(\kappa_{[w\xi]}\xi) + \frac{\kappa_{[w\xi]}}{\kappa_{[b\eta]}} e^{ika} \sin(\kappa_{[b\eta]}\eta) \right] \left[-\kappa_{[w\xi]} \sin(\kappa_{[w\xi]}\xi) - \kappa_{[b\eta]} e^{ika} \sin(\kappa_{[b\eta]}\eta) \right] \\ + \left[\cos(\kappa_{[w\xi]}\xi) - e^{ika} \cos(\kappa_{[b\eta]}\eta) \right] \left[\kappa_{[w\xi]} \cos(\kappa_{[w\xi]}\xi) - \kappa_{[b\eta]} e^{ika} \cos(\kappa_{[b\eta]}\eta) \right] = 0$$

After a lot of straightforward algebra, we find

$$\underbrace{\cos(\kappa_{[w\xi]}\xi) \cos(\kappa_{[b\eta]}\eta) - \frac{\kappa_{[b\eta]}^2 - \kappa_{[w\xi]}^2}{2\kappa_{[w\xi]}\kappa_{[b\eta]}} \sin(\kappa_{[w\xi]}\xi) \sin(\kappa_{[b\eta]}\eta)}_{F(E)} = \cos(ka) \quad (7.168)$$

This last equation relates the values of the wave vector k to the energy E through Equations 7.159 and 7.161: $\kappa_{[w\xi]} = \sqrt{\frac{2m}{\hbar^2}E}$ and $\kappa_{[b\eta]} = \sqrt{\frac{2m}{\hbar^2}(E - V_0)}$ (see the chapter problems). As a note, if we had started with $E < V_0$ then making the replacement $\kappa_{[b\eta]} \rightarrow i\kappa_{[b\eta]}$ produces hyperbolic cosines and sines in Equation 7.168.

Once having obtained Equation 7.168, we need to find the dispersion curve with the resulting energy band structure. Afterwards, we show how electron reflection from the interfaces corresponds to the Brillouin zone edges. We then solve a full quantum well problem and show the macroscopic boundary conditions, eigenfunctions and energy levels.

7.8.2 BANDS

Given a wave vector k , we need to find the corresponding energy $E = E(k)$ from Equation 7.168. The right side of $F(E)$ depends only on k while the left side implicitly depends on only E without any reference to k . Normally macroscopic boundary conditions over regions larger than 100 Å lead to discrete values of k . The allowed k values are very close together and for now, we assume the values of k form a continuum. Later, we will apply the macroscopic conditions. We choose a value for k and substitute into $\cos(ka)$ on right-hand side of Equation 7.168. The function $\cos(ka)$ can assume all values in the range $[-1, +1]$. Therefore the allowed values of energy E are those that put $F(E)$ in the range $[-1, +1]$ (refer to Figure 7.45). As a comment, notice that the equations still use the free electron mass.

To find the allowed energy as a function of k , proceed as follows. Take $k = 0$, then substitute in the right-hand side of Equation 7.168 to find $\cos(ka) = 1$. The left-hand side requires $F(E) = 1$. Find the values of E satisfying $F(E) = 1$. For $k = 0$, label the distinct values of E as $E = E_{1,k=0}, E_{2,k=0}, \dots$. Similarly, pick any value k and find a range of values from $F(E) = \cos(kL)$. The values can be labeled as $E = E_{1,k}, E_{2,k}, \dots$. Each value of k leads to a sequence of acceptable E as indicated in Figure 7.46. Notice that the subscripts have the same form as used for the Bloch wave function.

The “reduced zone representation” in Figure 9.46 shows that the bands (i.e., the values of k) only range within the FBZ. The width of the FBZ corresponds to the magnitude of the primitive reciprocal

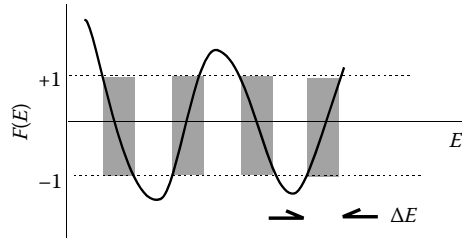


FIGURE 7.45 Kronig–Penney model produces bands.

lattice vector $G_1 = 2\pi/a$. Why do all the bands appear within the FBZ? The right-hand side of Equation 7.168 has $\cos(ka)$. If we allow k to change by a reciprocal lattice vector G (which is a multiple of the primitive reciprocal lattice vectors) then, recalling $Ga = 2\pi$ from Chapter 6, we find

$$\cos[(k + G)a] = \cos(ka + 2\pi) = \cos(ka) \quad (7.169)$$

So, the energy levels are sensitive to k only to within a reciprocal lattice vector!

Figure 9.47 shows how the “reduced zone representation” can be “unfolded” into the “extended zone representation” using translations in the reciprocal lattice. The figure also represents the solution of Equation 7.168. As k ranges over the real numbers, the allowed energy ranges (graph of E vs. k) forms the curved lines that approximate the parabola representing the free electron. Notice the energy gaps in the bands. The gaps correspond to the portions of Figure 7.45 with $F(E)$ outside the range of $(-1,1)$. The solid curved lines approximating the free-electron parabola make up the “extended zone representation.” Figures 7.46 and 7.47 provide two different plots of the dispersion relation $E(k)$ or sometimes written as $\omega(k)$.

As we will see later in the chapter, the interaction of the mass with periodic potential produces an effective mass. In actuality, the mass of the electron remains the same as the free space value. However, when we apply a force to the electron (such as with an electric field), we want to know the speed of the electron as calculated from Newton’s second law. The total force on the electron consists not only of the externally applied force but also those forces exerted by the crystal. We can ignore the crystal forces in Newton’s law so long as we replace the actual mass with an effective mass (which really represents the effects of the crystal potential). In this way, Newton’s law can be used without needing to consider the added complexity imposed by the crystal.

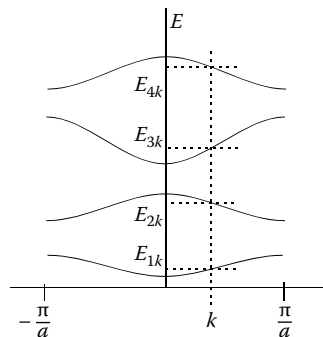


FIGURE 7.46 Shaded region indicates allowed energy bands. The band width is indicated by ΔE .

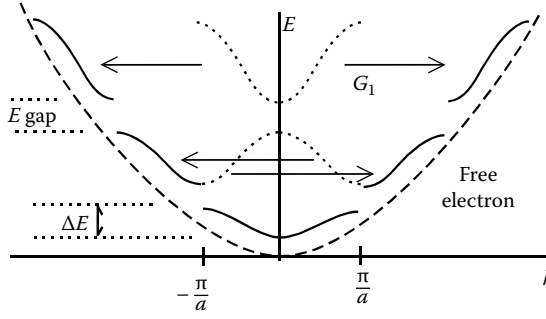


FIGURE 7.47 The extended zone representation.

Sections 7.7 and 7.8 show that the effective mass must be inversely proportional to the curvature of the band according to

$$m^* = \frac{1}{\hbar^2} \left(\frac{\partial^2 E}{\partial k^2} \right)^{-1} \quad (7.170)$$

Therefore, Figure 7.47 indicates the effective mass can vary appreciably throughout the bands. The effective mass for the electron and hole in GaAs near the bottom of the conduction band and the top of the valence band have the values 0.067 and 0.05 times the vacuum electron mass, respectively.

7.8.3 BANDWIDTH AND PERIODIC POTENTIAL

Finally in this section, we show how the magnitude of the periodic potential affects the bandwidth (ΔE in Figure 7.45), the gap size and the effective mass. We can see this most easily by using $E \ll V_0$. In this case, we let $\kappa_{b\eta} = i\kappa'$ to find

$$\cos(ka) = \frac{\kappa'^2 - \kappa_{w\xi}^2}{2\kappa'\kappa_{w\xi}} \sin h(\kappa'\eta) \sin[\kappa_{w\xi}(a - \eta)] + \cos h(\kappa'\eta) \cos[\kappa_{w\xi}(a - \eta)] \quad (7.171)$$

where

$$\kappa_{[w\xi]} = \sqrt{\frac{2m}{\hbar^2} E} \quad \text{and} \quad \kappa' = \sqrt{\frac{2m}{\hbar^2} (V_0 - E)} \quad (7.172)$$

To further simplify, let the barrier width approach zero $\eta \rightarrow 0$ and use the fact that

$$\sin h(\kappa'\eta) \rightarrow \kappa'\eta \quad \cos h(\kappa'\eta) \rightarrow 1 \quad (7.173)$$

Combining Equations 7.171 through 7.173 and using $E \ll V_0$ (essentially the barrier becomes a delta function), we find

$$\cos(ka) = \frac{U}{\kappa_{w\xi} a} \sin(\kappa_{w\xi} a) + \cos(\kappa_{w\xi} a) = F' \quad (7.174)$$

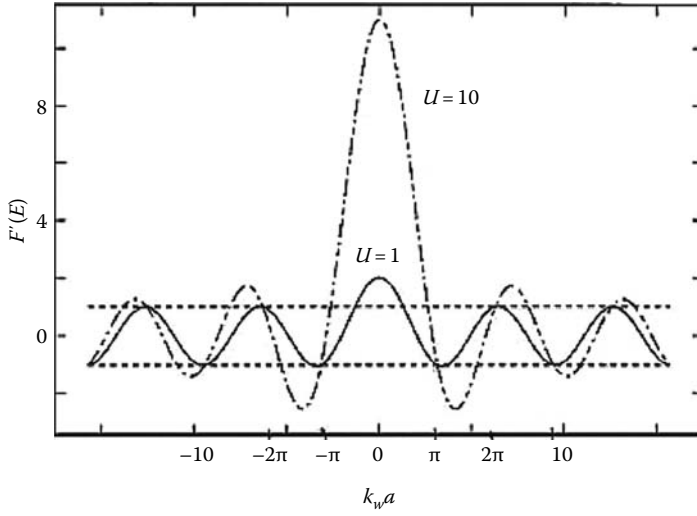


FIGURE 7.48 A plot of $F(E)$ vs. $\kappa_{[w\xi]}a$. (After Tiwari, S., *Compound Semiconductor Device Physics*, Academic Press, Boston, MA, 1992. With permission.)

where $U = m\alpha V_0/\hbar^2$. Figure 7.48 shows a plot of F' versus $\kappa_w \xi a$. As shown, larger values of U (i.e., larger values of potential) produce larger bandgaps (region above or below $+1$ or -1). Further the regions of allowed energy (the bandwidth ΔE) must decrease. Because the change in potential does not affect the FBZ width, the shape of the bands must change and so must the effective mass of the electron and hole.

7.9 TIGHT BINDING APPROXIMATION

The tight binding approximation (TBA) provides a model for the origin of bands without the simplification required by the Kronig–Penney model. The TBA model does not require us to approximate the crystal periodic potential as a sequence of square wells and barriers. In the nearly free electron model, the electron experiences relatively weak periodic potential. The TBA describes situations for which the atomic potential is quite large and the electron wave function is mostly localized about the atomic core. We show how the model gives rise to bandgaps. However, it can be extended to include real calculations for real crystals as discussed for example, Ashcroft and Mermin and their references.

7.9.1 INTRODUCTION

The model starts with a single atom having the single electron Hamiltonian

$$\hat{H}_0 = -\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r}) \quad (7.175)$$

with energy eigenvectors and eigenvalues satisfying

$$\hat{H}_0|\phi_a\rangle = E_a|\phi_a\rangle \quad (7.176)$$

(the index a stands for atom and labels the states of the atom).

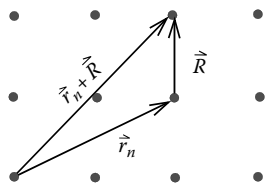


FIGURE 7.49 Atoms at lattice sites.

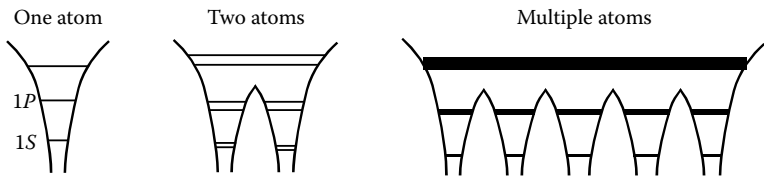


FIGURE 7.50 A single atom has single levels. Two atoms form two levels. Atoms in a lattice form energy bands.

Next, the model places N -identical atoms at lattice sites given by the lattice vectors $\vec{r}_n$ as shown in Figure 7.49 (where n ranges from 1 to N). As discussed in previous sections, the atomic core produces a periodic potential throughout the lattice. The electrons in the lowest energy levels E_a closest to the core interact mostly with their own core and very little with neighboring atoms. For example, Figure 7.50 shows a cartoon representation of the single atom with its discrete levels. Two noninteracting atoms each have identical energy levels E_a . Taken together, they degenerate. Bringing these two atoms together causes the levels to split in energy and form two distinct but closely spaced states as illustrated. The electrons nearest the cores experience the least amount of splitting since the outer electrons tend to shield them from neighboring atoms. The levels with higher energies correspond to the electrons with the smallest binding energy since inner electrons screen the core field. These valence electron wave functions tend to overlap between nearest neighbors and result in the greatest level splitting. If N atoms form a crystal then the resulting bands have N closely spaced levels in each band—essentially, each atom contributes one state to each band. The number of states agrees with that determined from the density of states calculations.

In the tight binding model, the wave function ϕ for an electron belonging to an atom forming part of a crystal mostly remains localized about the atomic core—hence the name “tight binding.” The outer-electron wave functions only slightly overlap between neighboring atoms. In fact, we assume that only nearest neighbor atoms have overlapping electron wave functions (see Figure 7.51). In the following calculations, we let the lattice vector $\vec{R}$ point to one of the nearest neighbor atoms as shown in Figure 7.49. As we have discussed previously, there appears to be two types of orthogonality between wave functions. The first type comes from the fact that Hermitian operators, such as the Hamiltonian $\vec{H}_0$ in Equation 7.175, produce orthonormal wave functions $|\phi_a\rangle$ satisfying $\langle\phi_a|\phi_b\rangle = \delta_{ab}$. The second type of orthogonality has to do with spatially separated wave functions

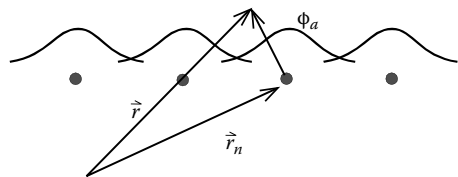


FIGURE 7.51 Overlapping wave functions from neighboring sites.

such as for $\phi_a(\vec{r} - \vec{r}_m)$ and $\phi_a(\vec{r} - \vec{r}_n)$. Both have the same index a to indicate identical wave functions but centered on two different atoms. Because of the small amount of overlap even between nearest neighbor atoms, we make the approximation.

$$\langle \phi_a(\vec{r} - \vec{r}_m) | \phi_a(\vec{r} - \vec{r}_n) \rangle = \int_{\text{all space}} d^3r \phi_a^*(\vec{r} - \vec{r}_m) \phi_a(\vec{r} - \vec{r}_n) \cong \delta_{mn} \quad (7.177)$$

As a note, the translation operators are unitary $\hat{T}_{\vec{r}_m}^\dagger = \hat{T}_{\vec{r}_m}^{-1}$ with $\hat{T}_{\vec{r}_m}^{-1} = \hat{T}_{-\vec{r}_m}$ and they commute $[\hat{T}_{\vec{r}_m}, \hat{T}_{\vec{r}_n}] = 0$ regardless of the chosen lattice vectors.

7.9.2 BLOCH WAVE FUNCTIONS

The tight binding approximation (TBA) assumes the full Hamiltonian for an electron in the crystal has an eigenvalue equation of form

$$\hat{\mathcal{H}}|\psi_{\vec{k}}\rangle = \mathcal{E}_{\vec{k}}|\psi_{\vec{k}}\rangle \quad (7.178a)$$

where $\psi_{\vec{k}}(\vec{r})$ represents the eigenfunction having the Bloch form and $\mathcal{E}_{\vec{k}}$ produces the electron dispersion curve. Technically the Bloch wave function should be subscripted with a to indicate the contributing atomic level (i.e., the band number). A number of authors view the full Hamiltonian $\hat{\mathcal{H}}$ as a type of perturbation to the atomic Hamiltonian $\hat{H}_0$ when they write $\hat{\mathcal{H}} = \hat{H}_0 + \Delta U$. We start with another approach that explicitly writes the full Hamiltonian and full potential as

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m}\nabla^2 + \mathcal{V}(\vec{r}) \quad \mathcal{V}(\vec{r}) = \sum_n V(\vec{r} - \vec{r}_n) \quad (7.178b)$$

We will later find the full Hamiltonian can be viewed as a perturbation to the atomic one. One of the eigenfunction solutions to Equation 7.178a, a type of Wannier function, has the Bloch form and consist of a linear combination of atomic orbitals (LCAO)

$$\psi_{\vec{k}}(\vec{r}) = \sum_{n=1}^N e^{i\vec{k}\cdot\vec{r}_n} \phi_a(\vec{r} - \vec{r}_n) \quad (7.179)$$

The full eigenfunction has a summation over N different atoms but using the same orbital a for each one.

We can demonstrate that Equation 7.179 provides a Bloch wave function with the property

$$\psi(\vec{r} + \vec{R}) = e^{i\vec{k}\cdot\vec{R}}\psi(\vec{r})$$

for any direct lattice vector $\vec{R}$. We find

$$\psi_{\vec{k}}(\vec{r} + \vec{R}) = \sum_{\vec{r}_n} e^{i\vec{k}\cdot\vec{r}_n} \phi_a(\vec{r} + \vec{R} - \vec{r}_n) = e^{i\vec{k}\cdot\vec{R}} \sum_{\vec{r}_n} e^{i\vec{k}\cdot(\vec{r}_n - \vec{R})} \phi_a(\vec{r} - (\vec{r}_n - \vec{R})) = e^{i\vec{k}\cdot\vec{R}} \psi_{\vec{k}}(\vec{r})$$

where we recognize the difference $\vec{r}_m = \vec{r}_n - \vec{R}$ as another lattice vector in the summation.

7.9.3 DISPERSION RELATION AND BANDS

We now calculate the energy eigenvalues $\mathcal{E}_{\vec{k}}$ in Equation 7.178a that provide the bands.

$$\langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle = \mathcal{E}_{\vec{k}} \langle \psi_{\vec{k}} | \psi_{\vec{k}} \rangle \quad \rightarrow \quad \mathcal{E}_{\vec{k}} = \frac{\langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle}{\langle \psi_{\vec{k}} | \psi_{\vec{k}} \rangle} \quad (7.180)$$

which explicitly retains the wave function inner product since the wave function is not normalized to one.

First work with the denominator. The inner product can be rewritten using Equation 7.179 and the approximate normalization in Equation 7.177.

$$\langle \psi_{\vec{k}} | \psi_{\vec{k}} \rangle = \sum_{m,n=1}^N e^{i\vec{k} \cdot (\vec{r}_n - \vec{r}_m)} \langle \phi_a(\vec{r} - \vec{r}_m) | \phi_a(\vec{r} - \vec{r}_n) \rangle = \sum_{m,n=1}^N e^{i\vec{k} \cdot (\vec{r}_n - \vec{r}_m)} \delta_{mn} = N \quad (7.181)$$

Next work with the numerator of Equation 7.180. Substituting Equation 7.179

$$\begin{aligned} \langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle &= \left\langle \sum_{m=1}^N e^{i\vec{k} \cdot \vec{r}_m} \phi_a(\vec{r} - \vec{r}_m) \left| \hat{\mathcal{H}} \right| \sum_{n=1}^N e^{i\vec{k} \cdot \vec{r}_n} \phi_a(\vec{r} - \vec{r}_n) \right\rangle \\ &= \sum_{m,n=1}^N e^{i\vec{k} \cdot (\vec{r}_n - \vec{r}_m)} \langle \phi_a(\vec{r} - \vec{r}_m) | \hat{\mathcal{H}} | \phi_a(\vec{r} - \vec{r}_n) \rangle \end{aligned}$$

Divide the summation into a diagonal and nondiagonal part as in

$$\langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle = \sum_{n=1}^N \langle \phi_a(\vec{r} - \vec{r}_n) | \hat{\mathcal{H}} | \phi_a(\vec{r} - \vec{r}_n) \rangle + \sum_{\substack{m,n \\ m \neq n}} e^{i\vec{k} \cdot (\vec{r}_n - \vec{r}_m)} \langle \phi_a(\vec{r} - \vec{r}_m) | \hat{\mathcal{H}} | \phi_a(\vec{r} - \vec{r}_n) \rangle \quad (7.182)$$

We assume only nearest neighbors contribute to the off diagonal terms and require $\vec{r}_m - \vec{r}_n$ to be a lattice vector $\vec{R} = \vec{r}_n - \vec{r}_m$ to the nearest neighbor. For a cubic lattice for example, there would be six such vectors $\vec{R}$. Equation 7.182 becomes

$$\langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle = \sum_{n=1}^N \langle \phi_a(\vec{r} - \vec{r}_n) | \hat{\mathcal{H}} | \phi_a(\vec{r} - \vec{r}_n) \rangle + \sum_{n=1}^N \sum_{\vec{R} \neq \vec{r}_n} e^{i\vec{k} \cdot \vec{R}} \langle \phi_a(\vec{r} - \vec{r}_n + \vec{R}) | \hat{\mathcal{H}} | \phi_a(\vec{r} - \vec{r}_n) \rangle \quad (7.183)$$

Dividing the potential into two parts

$$\mathcal{V}(\vec{r}) = \sum_{\eta=1}^N V(\vec{r} - \vec{r}_{\eta}) = V(\vec{r} - \vec{r}_n) + \sum_{\eta \neq n} V(\vec{r} - \vec{r}_{\eta}) \quad (7.184)$$

Note the full Hamiltonian can now be written as

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m} \nabla^2 + \mathcal{V}(\vec{r}) = -\frac{\hbar^2}{2m} \nabla^2 + V(\vec{r} - \vec{r}_n) + \sum_{\eta \neq n} V(\vec{r} - \vec{r}_{\eta}) = \hat{H}_0 + \sum_{\eta \neq n} V(\vec{r} - \vec{r}_{\eta}) \quad (7.185)$$

This last equation shows the TBA considers the full Hamiltonian to be a perturbation on the Hamiltonian for the atomic orbitals. Now Equation 7.183 becomes

$$\begin{aligned} \langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle &= \sum_{n=1}^N \langle \phi_a(\vec{r} - \vec{r}_n) | \left\{ \hat{H}_0 + \sum_{\eta \neq n} V(\vec{r} - \vec{r}_\eta) \right\} | \phi_a(\vec{r} - \vec{r}_n) \rangle \\ &+ \sum_{n=1}^N \sum_{\vec{R} \neq \vec{r}_n} e^{i\vec{k} \cdot \vec{R}} \langle \phi_a(\vec{r} - \vec{r}_n + \vec{R}) | \left\{ \hat{H}_0 + \sum_{\eta \neq n} V(\vec{r} - \vec{r}_\eta) \right\} | \phi_a(\vec{r} - \vec{r}_n) \rangle \end{aligned} \quad (7.186)$$

Noting $\hat{H}_0 | \phi_a \rangle = E_a | \phi_a \rangle$ from Equation 7.176, regardless of the coordinate origin because of the translational symmetry, we find

$$\begin{aligned} \langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle &= \sum_{n=1}^N E_a + \sum_{n=1}^N \langle \phi_a(\vec{r} - \vec{r}_n) | \left\{ \sum_{\eta \neq n} V(\vec{r} - \vec{r}_\eta) \right\} | \phi_a(\vec{r} - \vec{r}_n) \rangle \\ &+ \sum_{n=1}^N \sum_{\vec{R} \neq \vec{r}_n} e^{i\vec{k} \cdot \vec{R}} \langle \phi_a(\vec{r} - \vec{r}_n + \vec{R}) | \left\{ \sum_{\eta \neq n} V(\vec{r} - \vec{r}_\eta) \right\} | \phi_a(\vec{r} - \vec{r}_n) \rangle \end{aligned} \quad (7.187)$$

where the orthonormality in Equation 7.177 has been used

$$\langle \phi_a(\vec{r} - \vec{r}_n) | \phi_a(\vec{r} - \vec{r}_n) \rangle = 1 \quad \text{and} \quad \langle \phi_a(\vec{r} - \vec{r}_n + \vec{R}) | \phi_a(\vec{r} - \vec{r}_n) \rangle = 0$$

The last two summations in Equation 7.187 have translational symmetry and must give N identical terms. We might as well translate to the origin where $\vec{r}_n = 0$. Equation 7.187 becomes

$$\langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle = NE_a + N \sum_{\eta \neq 0} \langle \phi_a(\vec{r}) | V(\vec{r} - \vec{r}_\eta) | \phi_a(\vec{r}) \rangle + N \sum_{\vec{R} \neq 0} e^{i\vec{k} \cdot \vec{R}} \sum_{\eta \neq 0} \langle \phi_a(\vec{r} + \vec{R}) | V(\vec{r} - \vec{r}_\eta) | \phi_a(\vec{r}) \rangle$$

The total energy $\mathcal{E}_{\vec{k}}$ in Equation 7.180 can now be rewritten using this last equation and the normalization for $\psi_{\vec{k}}$ in Equation 7.181

$$\mathcal{E}_{\vec{k}} = \frac{\langle \psi_{\vec{k}} | \hat{\mathcal{H}} | \psi_{\vec{k}} \rangle}{\langle \psi_{\vec{k}} | \psi_{\vec{k}} \rangle} = \frac{NE_a - NA - N \sum_{\vec{R} \neq 0} B e^{i\vec{k} \cdot \vec{R}}}{N} \quad (7.188a)$$

or

$$\mathcal{E}_{\vec{k}} = E_a - A - \sum_{\vec{R} \neq 0} B e^{i\vec{k} \cdot \vec{R}} \quad (7.188b)$$

where $\{\vec{R}\}$ is the lattice vector from the atom at the origin to the nearest neighbors and

$$-A = \sum_{\eta \neq 0} \langle \phi_a(\vec{r}) | V(\vec{r} - \vec{r}_\eta) | \phi_a(\vec{r}) \rangle \quad -B = \sum_{\eta \neq 0} \langle \phi_a(\vec{r} + \vec{R}) | V(\vec{r} - \vec{r}_\eta) | \phi_a(\vec{r}) \rangle \quad (7.188c)$$

In the equation $\mathcal{E}_{\vec{k}} = E_a - A - \sum_{\vec{R} \neq 0} B e^{i\vec{k} \cdot \vec{R}}$, the constant A shifts the atomic energy E_a to lower values and the coefficient B determines the bandwidth.

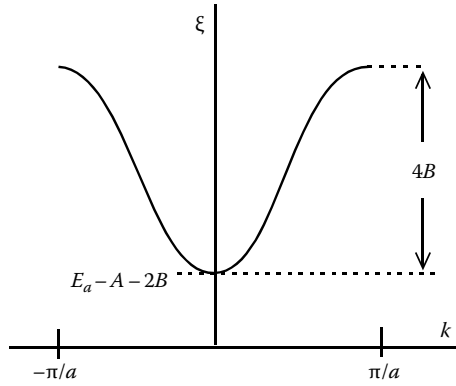


FIGURE 7.52 Band for Example 7.9.

Example 7.9

Find the band diagram for a 1-D crystal with lattice constant a .

SOLUTION

The vectors $\vec{R}$ to the nearest neighbors must be $\vec{R}_1 = a\hat{x}$ and $\vec{R}_2 = -a\hat{x}$. Equation 7.188b provides

$$\mathcal{E}_{\vec{k}} = E_a - A - \sum_{\vec{R} \neq 0} B e^{i\vec{k} \cdot \vec{R}} = E_a - A - B \{e^{ik_x a} + e^{-ik_x a}\}$$

So that

$$\mathcal{E}_{\vec{k}} = E_a - A - 2B \cos(k_x a)$$

The solution appears in Figure 7.52. The band has the shape of the cosine. A similar solution holds for the cubic crystal with 3-D k -vectors having components k_x, k_y, k_z . The bandwidth is $4B$ with a minimum energy at $E_a - A - 2B$. Notice also that multiple bands can only come from multiple values of the parameters A and B . These come from Equation 7.188c for the various values of a signifying the various atomic orbitals.

Example 7.10

Find the effective mass for the previous example near $k = 0$

SOLUTION

The effective mass can be calculated as

$$m_e^{-1} = \frac{1}{\hbar^2} \left(\frac{\partial^2 \mathcal{E}}{\partial k^2} \right)_{k=0} = \frac{2Ba^2}{\hbar^2}$$

We therefore see that the bandwidth must always be related to the effective mass.

7.10 INTRODUCTION TO EFFECTIVE MASS EQUATION

The electron can be treated as a free-electron for many purposes so long as the effective mass is used in the dynamical equations. The term “effective mass equation” refers to a Schrödinger wave equation that uses the effective mass but not the periodic crystal potential. The results from previous sections in this chapter implicitly use the effective mass equation without discussing it. The Bloch wave functions consist of the product of a plane wave and a periodic function; these wave functions comprise the eigenfunctions of the Hamiltonian consisting of kinetic and lattice potential terms. This section introduces the envelope approximation whereby adding an external macroscopic potential (spanning distances large compared with the unit cell size) requires the wave function to be a linear combination over the plane wave part but not over the periodic function. In short, the macroscopic potential only affects the envelope wave function. The added potential affects neither the effective mass nor the periodic portion of the block function. We do not prove the effective mass equation, but rather show the equivalence between it and the full Schrödinger equation with the periodic lattice potential.

7.10.1 THESIS

Previous sections show that the eigenfunctions and eigenvalues

$$\phi_{n,\vec{k}}(\vec{r}) = \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{\Omega}} u_{n,\vec{k}}(\vec{r}) \quad E_{n,\vec{k}} \quad (7.189)$$

satisfy the time-independent Schrödinger wave equation

$$-\frac{\hbar^2}{2m} \nabla^2 \phi(\vec{r}) + V_L(\vec{r}) \phi(\vec{r}) = E_{n,\vec{k}} \phi(\vec{r}) \quad (7.190a)$$

where V_L is the portion of the potential with the periodicity of the direct lattice and Ω represents the macroscopic volume associated with the periodic boundary conditions. The collection of eigenvalues $E_{n,\vec{k}} = E_n(\vec{k})$ gives the dispersion relation for band n . The allowed values of the wave vector $\vec{k}$ come from the macroscopic boundary conditions, $e^{i\vec{k}\cdot\vec{r}}$ is the envelope function, $u_{n,\vec{k}}(\vec{r})$ is the periodic solution for each unit cell, and m represents the free mass of the electron. Equation 7.190a describes a single electron in a periodic potential. In the next section, it will be convenient to replace Equation 7.190a with more compact notation

$$\hat{H}_0 |n, \vec{k}\rangle = E_{n,\vec{k}} |n, \vec{k}\rangle \quad (7.190b)$$

where

$$\hat{H}_0 = -\frac{\hbar^2}{2m} \nabla^2 + V_L(\vec{r}) \quad \text{and} \quad |n, \vec{k}\rangle = \phi_{n,\vec{k}} \quad (7.190c)$$

As usual, the basis functions $\langle \vec{r} | n, \vec{k} \rangle \equiv \phi_{n,\vec{k}}(\vec{r})$ must be orthonormal according to

$$\langle m, \vec{k} | n, \vec{k} \rangle = \delta_{m,n} \delta_{\vec{k},\vec{k}} \quad (7.191)$$

Even though u_{nk} repeats itself from one unit cell to the next, the inner product in this last equation must be over the larger distances L because of the wave vector k (we will see an example

later in this section). The reader should remember that “o” on $\hat{H}_o$ indicates the “original” Hamiltonian $\hat{H}_o$ since it includes only the lattice potential. The ket vectors in Equation 7.190c represent the eigenvectors for this simplest Hamiltonian. The circle on the symbol ϕ resembles the “o” on $\hat{H}_o$ to help remember which vector goes with which Hamiltonian. The solution to the time-dependent Schrödinger wave equation

$$-\frac{\hbar^2}{2m}\nabla^2\phi(\vec{r},t) + V_L(\vec{r})\Phi(\vec{r},t) = i\hbar\frac{\partial}{\partial t}\Phi(\vec{r},t) \quad (7.192)$$

consists of a sum over the eigenfunctions in Equation 7.189

$$\Phi(\vec{r},t) = \sum_{n,\vec{k}} C_{n,\vec{k}} \phi_{n,\vec{k}}(\vec{r}) e^{-itE_{n,\vec{k}}/\hbar} \quad (7.193)$$

In this section we wish to demonstrate the “envelop function approximation.” Suppose we apply a macroscopic potential $V_E(\vec{r},t)$ to a crystal from an external source. Actually, $V_E(\vec{r},t)$ can be any potential not related to the periodic lattice potential. Figure 7.53 shows an example that causes electrons and holes to start moving just after applying it. Assume for now that the potential $V_E(\vec{r})$ is independent of time. For example, it might be the built-in field in a junction. The word “macroscopic” refers to any variation that occurs over distances large compared with the size of the unit cells. The Schrödinger wave equation can be written as

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\vec{r},t) + \{V_L(\vec{r}) + V_E(\vec{r},t)\}\Psi(\vec{r},t) = i\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t) \quad (7.194a)$$

The effective mass equation eliminates the potential V_L at the expense of changing the free mass m to the effective mass m_e . The effective mass equation becomes

$$-\frac{\hbar^2}{2m_e}\nabla^2\Psi^{(e)}(\vec{r},t) + V_E(\vec{r})\Psi^{(e)}(\vec{r},t) = i\hbar\frac{\partial}{\partial t}\Psi^{(e)}(\vec{r},t) \quad (7.194b)$$

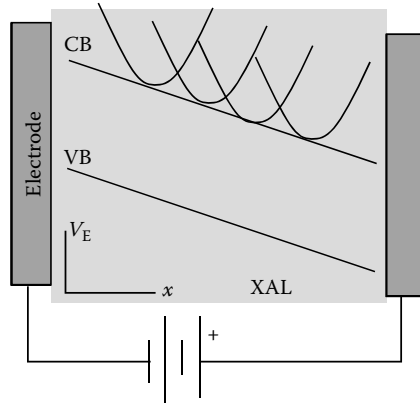


FIGURE 7.53 The band edge (slanted straight line) indicates the minimum of the cb or maximum of the valence band as a function of position.

where $\Psi^{(e)}(\vec{r}, t)$ is the solution wave function. We will show that the solution to full Schrödinger wave Equation 7.194a can be approximated by

$$\Psi(\vec{r}, t) = \sum_{n, \vec{k}} C_{n, \vec{k}} \phi_{n, \vec{k}}(\vec{r}) e^{-itE_{n, \vec{k}}/\hbar} = u_{n, \vec{k}}(\vec{r}) \sum_{\vec{k}} C_{n, \vec{k}} e^{i\vec{k} \cdot \vec{r} - itE_{n, \vec{k}}/\hbar} \quad (7.195a)$$

and that it is equivalent to the solution of 7.194b namely

$$\Psi^{(e)}(\vec{r}, t) = \sum_{\vec{k}} C_{n, \vec{k}} e^{i\vec{k} \cdot \vec{r} - itE_{n, \vec{k}}/\hbar} \quad (7.195b)$$

This is the envelop approximation. The same procedure does not work for the valence band of GaAs since the light-hole and heavy-hole bands are degenerate and the electron wave function will be a mixture of states from these two bands.

7.10.2 DISCUSSION OF THE SINGLE-BAND EFFECTIVE-MASS EQUATION

To show the equivalence between the single-band effective-mass equation for the envelope wave function and the full Schrödinger equation that includes the periodic potential, we will need four Hamiltonians. Let $\hat{H}_0$ be the original simplest Hamiltonian that includes only the periodic potential of the lattice. Let the Hamiltonian $\hat{H}$ include both the simplest Hamiltonian $\hat{H}_0$ and the macroscopic potential V_E . Assume the effective mass (or envelope) Hamiltonian $\hat{H}_e$ includes the macroscopic potential but not the lattice potential (the effective mass m_e in $\hat{H}_e$ already accounts for the lattice potential). Assume the “plane wave” Hamiltonian $\hat{H}_{eo}$ consists of the effective Hamiltonian without the macroscopic potential. The following list summarizes the various Hamiltonians.

Original:

$$\hat{H}_0 |n, \vec{k}\rangle = E_{n, \vec{k}} |n, \vec{k}\rangle \quad \hat{H}_0 = -\frac{\hbar^2}{2m} \nabla^2 + V_L(\vec{r}) \quad |n, \vec{k}\rangle = \phi_{n, \vec{k}} \equiv \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} u_{n, \vec{k}}(\vec{r}) \quad (7.196a)$$

Full:

$$\hat{H} |\psi(t)\rangle = i\hbar \partial_t |\psi(t)\rangle \quad \hat{H} = -\frac{\hbar^2}{2m} \nabla^2 + (V_L + V_E) \quad \hat{H} = \hat{H}_0 + V_E \quad (7.196b)$$

Plane wave:

$$\hat{H}_{eo} |\vec{k}\rangle = E_{\vec{k}} |\vec{k}\rangle \quad \hat{H}_{eo} = -\frac{\hbar^2}{2m_e} \nabla^2 \quad \langle \vec{r} | \vec{k} \rangle = \psi_{\vec{k}}(\vec{r}) = \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} \quad (7.196c)$$

Effective:

$$\hat{H}_e |\psi^e(t)\rangle = i\hbar \partial_t |\psi^e(t)\rangle \quad \hat{H}_e = -\frac{\hbar^2}{2m_e} \nabla^2 + V_E \quad \hat{H}_e = \hat{H}_{eo} + V_E \quad (7.196d)$$

Notice that Equation 7.196c and d agree except for the macroscopic potential V_E . The plane waves in Equation 7.196c have been normalized to a macroscopic volume Ω .

The objective consists of showing that for systems including large numbers of primitive cells (i.e., those with boundary conditions covering distances L larger than a good number of primitive cells) the solutions to the effective wave Equation 7.196d can be used in place of the full wave Equation 7.196b.

$$\hat{H}|\psi(t)\rangle = i\hbar\partial_t|\psi(t)\rangle \leftrightarrow \hat{H}_e|\psi^e(t)\rangle = i\hbar\partial_t|\psi^e(t)\rangle \quad (7.197)$$

We want to reduce the full Hamiltonian and wave function to the effective ones using specific assumptions about the potential. Assume that the electron lives in a single band—the conduction band for GaAs (i.e., $n = 2$). As mentioned previously, the effective mass equation cannot be used for the GaAs valence bands because the LH and HH bands are degenerate at $k = 0$. The demonstration starts by showing the eigenfunctions for the full Hamiltonian $\hat{H}$ reduce to those for the effective Hamiltonian $\hat{H}_e$.

We first expand the solution to the full wave equation in terms of the Bloch wave functions (since they form a basis set). Next we obtain a matrix equation to replace the full Schrödinger wave equation.

Expanding the solution to the full wave equation in terms of the Bloch wave functions, which are eigenvectors of the original simplest Hamiltonian $\hat{H}_0$

$$|\psi(t)\rangle = \sum_{n\vec{k}} C_{n\vec{k}}(t) |n, \vec{k}\rangle \quad (7.198)$$

Substituting into the full Hamiltonian in Equation 7.196b provides

$$\hat{H} \sum_{n\vec{k}} C_{n\vec{k}}(t) |n, \vec{k}\rangle = i\hbar\partial_t \sum_{n\vec{k}} C_{n\vec{k}}(t) |n, \vec{k}\rangle \quad (7.199)$$

A matrix equation can be obtained by operating on the left with a general bra $\langle m, \vec{k}|$

$$\sum_{n\vec{k}} C_{n\vec{k}}(t) \langle m, \vec{k} | \hat{H} | n, \vec{k} \rangle = i\hbar \sum_{n\vec{k}} \dot{C}_{n\vec{k}}(t) \langle m, \vec{k} | n, \vec{k} \rangle \quad (7.200)$$

Next using $\hat{H} = \hat{H}_0 + V_E$ and the fact that the Bloch eigenvectors are orthonormal yields

$$\sum_{n\vec{k}} C_{n\vec{k}}(t) \langle m, \vec{k} | \hat{H}_0 + V_E | n, \vec{k} \rangle = i\hbar \dot{C}_{m\vec{k}} \quad (7.201)$$

We also note that

$$\langle m, \vec{k} | \hat{H}_0 | n, \vec{k} \rangle = E_{n,\vec{k}} \langle m, \vec{k} | n, \vec{k} \rangle = E_{n,\vec{k}} \delta_{m\vec{k}, n\vec{k}} \quad (7.202)$$

Let us use new notation for V_E because of all the subscripts running around $V_E \rightarrow V$. Equation 7.201 becomes

$$\sum_{n\vec{k}} C_{n\vec{k}}(t) (E_{n,\vec{k}} \delta_{m\vec{k}, n\vec{k}} + V_{m\vec{k}, n\vec{k}}) = i\hbar \dot{C}_{m\vec{k}} \quad (7.203)$$

or

$$C_{m\vec{k}}(t) E_{m\vec{k}} + \sum_{n\vec{k}} C_{n\vec{k}}(t) V_{m\vec{k}, n\vec{k}} = i\hbar \dot{C}_{m\vec{k}} \quad (7.204)$$

However, we only work with the conduction band and should only have $n = m = 2$ in the equations. Examining Equation 7.204, we see that the subscript m does not occur in a summation contrary to the situation for the subscript n .

Suppose V does not connect states in the valence band with those in the conduction band. In particular, the motion of the electron in the conduction band does not depend on the presence of the valence band. Therefore assume

$$V_{m\vec{k},n\vec{k}} = \langle m, \vec{k} | V | n, \vec{k} \rangle = V_{\vec{k},\vec{k}} \delta_{mn} \quad (7.205)$$

where $V = V_E$. For nondegenerate valence bands, the same considerations apply to motion in the valence bands (indium added to GaAs strains the crystal, lowers the LH valence band and thereby removes the degeneracy at $k = 0$). Without recombination processes, we usually assume electrons in the conduction band do not depend on those in the valence band. The discussion in the next section examines the issue of V in Equation 7.205 more carefully. Equation 7.204 becomes

$$C_{m\vec{k}}(t)E_{m\vec{k}} + \sum_{\vec{k}} C_{m\vec{k}}(t) V_{m\vec{k},m\vec{k}} = i\hbar \dot{C}_{m\vec{k}} \quad (7.206)$$

Notice the potential V_L periodic in the lattice does not appear but E carries its imprint. This last equation has the same band index m in all terms which means an electron starting in band $\#m$ must remain in band $\#m$. In this case the solution does not allow electrons to transition from the CB to the vb. The full wave function solution to the full Schrödinger wave equation in Equation 7.196b must not include any of the valence band states. However, the potential connects different k -states. An applied field, for example, can accelerate an electron or hole and thereby change its first wave vector into a different one.

We drop the dependence on the index m in Equation 7.206 by assuming the electron remains in a single band (say CB) and is not influenced by other bands. We essentially reverse the procedure leading to Equation 7.206. Denote the plane waves given by $|\vec{k}\rangle$; that is, the coordinate representation is given by $\langle \vec{r} | \vec{k} \rangle = \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{\Omega}}$, which essentially represents the plane wave part of the Bloch wave function without the periodic part u .

$$\sum_{\vec{k}} C_{\vec{k}}(t)(E_{\vec{k}}\delta_{\vec{k}\vec{k}} + V_{\vec{k}\vec{k}}) = i\hbar \dot{C}_{\vec{k}}(t)$$

where we drop the subscript m from Equation 7.206 for simplicity. Reinserting the bras and kets produces

$$\sum_{\vec{k}} C_{\vec{k}}(t) \langle \vec{k} | (\hat{H}_{eo} + V) | \vec{k} \rangle = i\hbar \sum_{\vec{k}} \dot{C}_{\vec{k}}(t) \langle \vec{k} | \vec{k} \rangle$$

where $\hat{H}_{eo}$ has the mass parameter m_e to produce the $E_{\vec{k}}$ that differ from the free-electron case. This last expression must be true for all plane wave projectors $\langle \vec{k} |$. We therefore conclude

$$(\hat{H}_{eo} + V) \sum_{\vec{k}} C_{\vec{k}}(t) |\vec{k}\rangle = i\hbar \frac{\partial}{\partial t} \sum_{\vec{k}} C_{\vec{k}}(t) |\vec{k}\rangle \quad (7.207)$$

We recognize the envelope wave function expanded in plane waves

$$|\psi^e\rangle = \sum_{\vec{k}} C_{\vec{k}}(t) |\vec{k}\rangle \quad (7.208)$$

Finally, combining the last two equations produces

$$(\hat{H}_{e0} + V)|\psi^e\rangle = i\hbar \frac{\partial}{\partial t} |\psi^e\rangle \rightarrow \hat{H}_e |\psi^e\rangle = i\hbar \frac{\partial}{\partial t} |\psi^e\rangle$$

as required. This procedure shows the effective mass equation and the envelope approximation. The effective mass in $\hat{H}_e$ ensures the proper energy $E_{\vec{k}} = E(k)$ that carries the imprint of the periodic potential and differs from that of the free electron.

The assumption that the macroscopic potential is diagonal in the band index (Equation 7.205) shows that full Hamiltonian (without the effective mass) reduces to the effective Hamiltonian (with the effective mass) so long as the full wave function (with the periodic Bloch function) is replaced with the envelope wave function (without the periodic Bloch function). In this section, we have discussed how the full Hamiltonian reduces to the effective Hamiltonian. The same E_k appears as the eigenvalue for both $\hat{H}_{e0}$ in Equation 7.196c and $\hat{H}_0$ in Equation 7.196a. However, the $\hat{H}_{e0}$ has the effective mass m_e that adjusts the value of the energy so that the same E_k can be used. The appearance of the effective mass will be demonstrated in the section on $\vec{k} \cdot \vec{p}$ theory.

7.10.3 ENVELOPE APPROXIMATION

One can see the reason why the envelope approximation works as given in Equation 7.195. Consider the equivalence between the following two equations

$$\Psi(\vec{r}, t) = \sum_{n, \vec{k}} C_{n, \vec{k}} \phi_{n, \vec{k}}(\vec{r}) e^{-itE_{n, \vec{k}}/\hbar} = u_{n, \vec{k}}(\vec{r}) \sum_{\vec{k}} C_{n, \vec{k}} e^{i\vec{k} \cdot \vec{r} - itE_{n, \vec{k}}/\hbar} \quad (7.195a)$$

$$\Psi^{(e)}(\vec{r}, t) = \sum_{\vec{k}} C_{n, \vec{k}} e^{i\vec{k} \cdot \vec{r} - itE_{n, \vec{k}}/\hbar} \quad (7.195b)$$

The first summation in Equation 7.195a represents the full wave function for the Hamiltonian that includes the periodic potential. The second summation shows the envelope approximation. The wave function in Equation 7.195b corresponds to the Hamiltonian with the effective mass and without the periodic potential. The present section will discuss the equivalence.

First assume the full wave function Ψ does not contain a mixture of valence and conduction band states (no sum over the band index); this eliminates the summation over n in the first term of Equation 7.195a.

Now one can show why the periodic function u can be removed from Equation 7.195a. To do this, we need a result from the next section on $\vec{k} \cdot \vec{p}$ theory that shows the lowest order perturbation to the Hamiltonian for the periodic Bloch function u is the term $\vec{k} \cdot \vec{p}$. Near the bottom of the conduction band where $k \sim 0$, the perturbation $\vec{k} \cdot \vec{p}$ is small and we therefore expect the periodic part of the Bloch wave function u to only weakly depend on k near $k=0$. Therefore, $u_{n, \vec{k}} \cong u_{n, \vec{0}}$ as required to remove u from the summation.

We might expect that $u_{n, \vec{k}} \cong u_{n, \vec{0}}$ by making a Fourier transformation of the partial differential equation for the periodic function u . Because u has the periodicity of the lattice, the Fourier summation must be over the reciprocal lattice vectors $\vec{G}$ as discussed in Section 6.4. The solution for the function u must depend on both $\vec{G}$ and $\vec{k}$ where $|\vec{G}| \gg |\vec{k}|$ as discussed for wave vectors confined to the FBZ. Therefore in the formula for u , we can neglect the $\vec{k}$ to lowest order to find $u_{n, \vec{k}} \cong u_{n, \vec{0}}$.

Armed with the key assumption that u is relatively independent of wave vector $\vec{k}$, we find

$$\psi(\vec{r}, t) = \sum_{\vec{k}} C_{n\vec{k}}(t) \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} u_{n\vec{k}}(\vec{r}) \cong u_{n\vec{k}}(\vec{r}) \sum_{\vec{k}} C_{n\vec{k}}(t) \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} = u_{n\vec{k}}(\vec{r}) \sum_{\vec{k}} C_{\vec{k}}(t) \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} = u_{n\vec{k}}(\vec{r}) \psi^e(\vec{r}, t) \quad (7.209)$$

Equation 7.209 provides an alternate statement of the Bloch wave function. Notice only the envelope part depends on time. Traveling waves therefore involve only the motion of the envelope.

There are some problems with continuity of the effective wave function ψ_E across boundaries. The literature discusses whether to require the first derivative of the effective wave function to be continuous across a boundary or not. Some schemes require the current density to be continuous and therefore include the effective mass. However, where the effective mass does not change much with material composition, it does not matter whether differentiates it or not.

7.10.4 DIAGONAL MATRIX ELEMENTS OF V_E

In this section consider the conditions under which $V_{m\vec{k},n\vec{k}} = V_{\vec{k},\vec{k}}\delta_{mn}$; that is, V does not connect different bands. Start with the definition of the matrix element

$$V_{m\vec{k},n\vec{k}} = \langle m\vec{k} | V | n\vec{k} \rangle = \int_{\Omega} d^3r \frac{e^{-i\vec{k}\cdot\vec{r}}}{\sqrt{\Omega}} u_{m\vec{k}}^*(\vec{r}) V(\vec{r}) \frac{e^{i\vec{k}\cdot\vec{r}}}{\sqrt{\Omega}} u_{n\vec{k}}(\vec{r}) \quad (7.210)$$

where the integral covers the large volume Ω (volume of the crystal) which leads to the allowed values of $\vec{k}$. We follow a procedure similar to finding the normalization of the Bloch wave functions in Section 7.6. Assume that the volume Ω includes N unit cells. Let $\vec{R}_j$ be the lattice vector pointing to the center of unit cell $\#j$ (see Figure 7.54). Rather than having $\vec{r}$ range over all space, let us make the new definition $\vec{r} \rightarrow \vec{R}_j + \vec{r}$ so that $\vec{r}$ ranges only within a single unit cell. The vector $\vec{R}_j$ picks the unit cell and $\vec{r}$ picks the point within the cell. Equation 7.210 can be written as

$$V_{m\vec{k},n\vec{k}} = \frac{1}{\Omega} \sum_{j=1}^N \int_{\Omega_j} d^3r e^{-i\vec{k}\cdot(\vec{R}_j+\vec{r})} u_{m\vec{k}}^*(\vec{R}_j+\vec{r}) V(\vec{R}_j+\vec{r}) e^{i\vec{k}\cdot(\vec{R}_j+\vec{r})} u_{n\vec{k}}(\vec{R}_j+\vec{r}) \quad (7.211)$$

where Ω_c is the volume of each unit cell (i.e., $\Omega = N\Omega_c$), and the direct lattice vector $\vec{R}_j$ points to the center of unit cell $\#j$ (see Figure 7.54). Because the functions u have the periodicity of the lattice, Equation 7.211 can be written as

$$V_{m\vec{k},n\vec{k}} = \frac{1}{\Omega} \sum_{j=1}^N \int_{\Omega_j} d^3r e^{-i\vec{k}\cdot(\vec{R}_j+\vec{r})} u_{m\vec{k}}^*(\vec{r}) V(\vec{r}+\vec{R}_j) e^{i\vec{k}\cdot(\vec{R}_j+\vec{r})} u_{n\vec{k}}(\vec{r}) \quad (7.212)$$

If the wave vectors $k \sim 2\pi/L$ correspond to large distances L (larger than the size of the primitive cells) and if $|\vec{r}| \ll L$ (r in the integral is confined to a given unit cell), then we have $\vec{k} \cdot \vec{r} \sim 0$ and $e^{\pm i\vec{k}\cdot\vec{r}} \sim 1$. If the wave vector k is close to the Brillouin zone edge ($k \sim \pi/a$ where a is the atomic

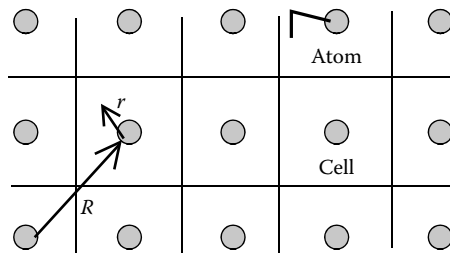


FIGURE 7.54 The vector R indicates the center of the cell and r ranges over the interior of the cell.

spacing for a simple cubic lattice) then for $|\vec{r}| \sim a$ the $\vec{k} \cdot \vec{r} \sim \pi$ and the approximation breaks down. The electrons should remain close to the bottom of the conduction band for our approximation to work. That is, the sum over k in the equations for Section 7.10.2 should not include k close to the FBZ edge.

Continuing with Equation 7.212, factor out the exponentials that depend on $\vec{R}_j$ and assume that the external potential $V = V_E$ varies slowly over the unit cell so that $V(\vec{r} + \vec{R}_j) \sim V(\vec{R}_j)$; therefore it can be factored out of the integral.

$$V_{m\vec{\kappa},n\vec{k}} = \frac{1}{\Omega} \sum_{j=1}^N e^{i(\vec{k}-\vec{\kappa}) \cdot \vec{R}_j} V(\vec{R}_j) \int_{\Omega_j} d^3r u_{m\vec{\kappa}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) \quad (7.213)$$

This last integral in Equation 7.213 can be evaluated by making use of a similar procedure to that just employed. Section 7.5 shows the orthonormality

$$\delta_{m\vec{\kappa},n\vec{k}} = \langle m, \vec{\kappa} | n, \vec{k} \rangle = \int_{\Omega} d^3r \frac{e^{-i\vec{\kappa} \cdot \vec{r}}}{\sqrt{\Omega}} u_{m\vec{\kappa}}^*(\vec{r}) \frac{e^{+i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} u_{n\vec{k}}(\vec{r}) \quad (7.214)$$

and demonstrates for u relatively independent of k

$$\int_{\Omega_{\text{cell}}} d^3r u_{m\vec{\kappa}}^*(\vec{r}) u_{n\vec{k}}(\vec{r}) \cong \Omega_{\text{cell}} \delta_{m,n} \quad (7.215)$$

Returning to Equation 7.213, we therefore find

$$V_{m\vec{\kappa},n\vec{k}} = \frac{1}{\Omega} \sum_{j=1}^N e^{i(\vec{k}-\vec{\kappa}) \cdot \vec{R}_j} V(\vec{R}_j) \Omega_{\text{cell}} \delta_{m,n} \quad (7.216)$$

Considering each small cell volume $\Omega_j = \Omega_{\text{cell}}$ as a differential volume d^3R we find

$$V_{m\vec{\kappa},n\vec{k}} = \int_{\Omega} d^3R \frac{e^{-i\vec{\kappa} \cdot \vec{R}}}{\sqrt{\Omega}} V(\vec{R}) \frac{e^{-i\vec{k} \cdot \vec{R}}}{\sqrt{\Omega}} \delta_{mn} = \langle \vec{\kappa} | V | \vec{k} \rangle \delta_{mn} = V_{\vec{\kappa},\vec{k}} \delta_{mn} \quad (7.217)$$

Based on the above calculations, the potential $V = V_E$ does not connect different bands because V varies over large scales compared with the unit cells and because the functions u only weakly depend on k .

7.10.5 SUMMARY

The Schrödinger wave equation for the heterostructure can be written as

$$-\frac{\hbar^2}{2m} \nabla^2 \Psi + (V + V_L) \Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (7.218)$$

where m denotes the free mass of the electron. The wave function has the form

$$|\Psi(t)\rangle = \sum_{\vec{k}} \beta_{n\vec{k}}(t) |n, \vec{k}\rangle = \sum_{\vec{k}} \beta_{n\vec{k}}(0) |n, \vec{k}\rangle e^{-iE_{n\vec{k}}t/\hbar} \quad (7.219)$$

where the eigenfunctions have the form

$$|n, \vec{k}\rangle \sim \psi(\vec{r}) = \frac{1}{\sqrt{V}} e^{i\vec{k} \cdot \vec{r}} u_{n,\vec{k}}(\vec{r}) \quad (7.220)$$

and we confine our attention to the conduction band. A similar expression can be used for the valence bands so long as the light and heavy-hole bands have sufficient separation in energy (nondegenerate bands). The basis functions for the Hilbert space of envelope functions

$$\phi_{\vec{k}}(\vec{r}) = \frac{1}{\sqrt{V}} e^{i\vec{k} \cdot \vec{r}} \quad (7.221a)$$

satisfy the orthonormality relation

$$\langle \phi_{\vec{k}} | \phi_{\vec{k}'} \rangle = \delta_{\vec{k}\vec{k}'} \quad (7.221b)$$

The Bloch functions $u_{n,\vec{k}}$ are periodic on the crystal so that the values of $u_{n,\vec{k}}$ repeat from one unit cell to the next. We have included a normalization factor in the Bloch function $u_{n,\vec{k}}$ so that they satisfy an inner product over the unit cell of the form.

$$\langle u_{n\vec{k}} | u_{m\vec{k}} \rangle_{\text{uc}} = \int_{\text{uc}} dV u_{n\vec{k}}^* u_{m\vec{k}} = \delta_{mn} \quad (7.222a)$$

We consider only the conduction band ($n=2$) and define $u_{2,\vec{k}} = u_{\vec{k}}$. So that

$$\langle u_{2\vec{k}} | u_{2\vec{k}} \rangle_{\text{uc}} \equiv \langle u_{\vec{k}} | u_{\vec{k}} \rangle_{\text{uc}} = \int_{\text{uc}} dV u_{\vec{k}}^* u_{\vec{k}} = 1 \quad (7.222b)$$

where “uc” restricts the integration over any unit cell and we represent the conduction band by $n=2$.

The general vector in the space spanned by the basis set

$$|n, \vec{k}\rangle \sim \psi_{n,\vec{k}}(\vec{r}) = \frac{1}{\sqrt{V}} e^{i\vec{k} \cdot \vec{r}} u_{n,\vec{k}}(\vec{r}) \quad (7.223a)$$

has the form

$$\Psi(\vec{r}, 0) = \sum_{\vec{k}} \beta_{\vec{k}} \psi_{n,\vec{k}}(\vec{r}) = \sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}} u_{n,\vec{k}}(\vec{r}) \quad (7.223b)$$

The envelope approximation notes that $u_{n,\vec{k}}(\vec{r})$ must be relatively independent of the wave vector $\vec{k}$ since it corresponds to a wavelength having the size of many unit cells whereas has distinct values only within the unit cell. Therefore, writing $u_{n,\vec{k}}(\vec{r}) \sim u_{n,0}(\vec{r}) \equiv u_n(\vec{r})$, we can write

$$\Psi(\vec{r}, 0) = \sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}} u_{n,\vec{k}}(\vec{r}) \cong \left[\sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}}(\vec{r}) \right] u_n(\vec{r}) = F(\vec{r}) u_n(\vec{r}) \quad (7.223c)$$

The envelope function $F(\vec{r})$ resides in the Hilbert space spanned by the envelope basis set $\{\phi_{\vec{k}}(\vec{r})\}$. We therefore see that the solution to the Schrödinger wave equation must have the form of modulated carrier.

7.11 INTRODUCTION TO $\vec{k} \cdot \vec{p}$ BAND THEORY

The $\vec{k} \cdot \vec{p}$ approximation finds widespread application in semiconductor theory especially for optoelectronics. It provides a method of deducing the periodic Bloch function u in the Bloch wave function $\frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{V}} u_{n,\vec{r}}(\vec{r})$. The $\vec{k} \cdot \vec{p}$ theory allows one to calculate the band structure $E_n(\vec{k})$ near the band edge (bottom of the conduction band or the top of the valence bands). The theory can be applied to single or to multiply degenerate bands. For the $\vec{k} \cdot \vec{p}$ theory, two approaches are common. The first approach applies perturbation theory while the second one solves an equation for a determinant.

The present sections develops the $\vec{k} \cdot \vec{p}$ theory by first substituting the Bloch wave function into the Schrödinger equation with the lattice potential and the free electron mass. The envelope portion can be removed from the equation leaving behind a new Schrödinger equation for the periodic wave function $u_{n,\vec{k}}$ and a $\vec{k} \cdot \vec{p}$ term. For electrons near the band edge (CB minimum or vb maximum), the wave vector is small and the $\vec{k} \cdot \vec{p}$ term can be treated as a type of perturbation. The theory provides the dispersion curves $E_n(\vec{k})$ and the periodic wave function $u_{n,\vec{k}}$ so long as we know the band energy $E_n(\vec{k} = 0)$ (the minimum and maximum band energy for direct bandgaps) and the wave function $u_{n,\vec{k}=0}$. The $u_{n,\vec{k}=0}$ are eigenfunctions of a Hermitian operator and therefore form a basis set for all functions periodic across the unit cells (periodic in the lattice). We can therefore expand each $u_{n,\vec{k}}$ (for fixed $\vec{k} \neq 0$) in terms of the $u_{n,\vec{k}=0}$ (as a summation over the index n). The $u_{n,\vec{k}}$ also appear as eigenfunctions of a Hermitian operator and can be used as a basis in the indices n and $\vec{k}$. The second order energy term for the perturbation produces the effective mass. Often, higher order corrections to the function $u_{n,\vec{k}}$ can be ignored when the direct bandgap $E_g = E_2(\vec{k} = 0) - E_1(\vec{k} = 0)$ is sufficiently large. One then finds the usual envelop approximation whereby the wave propagating in a crystal appears as a summation over the envelope wave function and not over the periodic function $u_{n,\vec{k}=0}$. We generalize the development to degenerate valence bands.

The present section discusses nondegenerate bands. The next section discusses the degenerate case. It accounts for the conduction band, the light- and heavy-hole bands and the split-off band due to spin-orbit coupling.

7.11.1 BRIEF REMINDER ON BLOCH WAVE FUNCTION

As a reminder, the Bloch wave functions have the following orthonormality relations.

$$\delta_{mn} \delta_{\vec{k}\vec{k}'} = \left\langle m\vec{k} \left| n\vec{k}' \right. \right\rangle = \int_V d^3r \left(\frac{e^{i\vec{k}' \cdot \vec{r}}}{\sqrt{V}} u_{n\vec{k}'}(\vec{r}) \right)^* \left(\frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{V}} u_{m\vec{k}}(\vec{r}) \right) = \int_V d^3r \frac{e^{i(\vec{k}-\vec{k}') \cdot \vec{r}}}{V} u_{m\vec{k}}^* u_{n\vec{k}'}$$

As discussed in Section 7.5, the integral can be simplified if we recall that the extended states sufficiently far from the Brillouin zone edge have small wave vectors. We used the properties that (1) the electron wavelength is large compared with the size of the unit cell and (2) the function u is periodic over the lattice. We found

$$\frac{1}{V_{uc}} \int_{V_{uc}} d^3r' u_{m\vec{k}}^* u_{n\vec{k}} = \delta_{mn} \quad (7.224a)$$

where V_{uc} represents the volume of the unit cell. Notice that this last integral can also be written as

$$\frac{1}{V} \int_V d^3r' u_{m\vec{k}}^* u_{n\vec{k}} = \delta_{mn} \quad (7.224b)$$

where $V_{uc} \rightarrow V$ because of the repetitive nature of u . The $\vec{k} \cdot \vec{p}$ theory shows that u is relatively independent of k so that the envelope function (the exponential) carries most of the orthonormality over the k variable.

For simplicity of notation, one can normalize the function u so that the integral does not require the extra V_{uc} factor. Making the replacement

$$u_{n\vec{k}} \rightarrow \sqrt{V_{uc}} u_{n\vec{k}} \quad \text{or} \quad u_{n\vec{k}} \rightarrow \sqrt{V} u_{n\vec{k}} \quad (7.224c)$$

which produces the orthonormality relation with

$$\int_{V_{uc}} d^3 r' u_{m\vec{k}}^* u_{n\vec{k}} = \delta_{mn} \quad \text{or} \quad \int_V d^3 r' u_{m\vec{k}}^* u_{n\vec{k}} = \delta_{mn} \quad (7.224d)$$

When the normalization of V_{uc} is needed, simply follow through the calculations and replace u and the inner products as necessary. We also assume that for a given $\vec{k}$, the function $u_{n,\vec{k}}$ form a complete set (where n runs over all of the bands).

7.11.2 $\vec{k} \cdot \vec{p}$ EQUATION FOR PERIODIC BLOCH FUNCTION

We must first find a partial differential equation for the periodic Bloch function u . The Bloch functions satisfy the time-independent Schrödinger equation that includes a kinetic term and the crystal potential V_L .

$$\left[-\frac{\hbar^2}{2m_o} \nabla^2 + V_L \right] \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) = E_{n\vec{k}} \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{V}} u_{n\vec{k}}(\vec{r}) \quad (7.225)$$

where

$E_{n\vec{k}} = E_n(\vec{k})$ gives the dispersion relation for band n
 V refers to the macroscopic size of the crystal

We can evaluate the derivatives to find

$$e^{i\vec{k} \cdot \vec{r}} \left[-\frac{\hbar^2}{2m_o} \left(-k^2 u_{n\vec{k}} + 2i\vec{k} \cdot \nabla u_{n\vec{k}} + \nabla^2 u_{n\vec{k}} \right) + V_L u_{n\vec{k}} \right] = E_{n\vec{k}} e^{i\vec{k} \cdot \vec{r}} u_{n\vec{k}}(\vec{r}) \quad (7.226)$$

where the exponential has been factored out. The exponential cancels from each side. Further, we make the replacement $\nabla u_{n\vec{k}} = \frac{i}{\hbar} \hat{p} u_{n\vec{k}}$. We will later see that $\langle u_{m\vec{k}} | \hat{p} | u_{n\vec{k}} \rangle$ refers to the motion of an electron near a nucleus (for example, valence bands).

$$\frac{\hbar^2}{2m_o} \left(k^2 u_{n\vec{k}} + \frac{2}{\hbar} \vec{k} \cdot \hat{p} u_{n\vec{k}} + \frac{1}{\hbar^2} \hat{p}^2 u_{n\vec{k}} \right) + V_L u_{n\vec{k}} = E_{n\vec{k}} u_{n\vec{k}} \quad (7.227)$$

Regrouping provides

$$\left(H_o + \frac{\hbar}{m_o} \vec{k} \cdot \hat{p} \right) u_{n\vec{k}} = \left(E_n(\vec{k}) - \frac{\hbar^2 k^2}{2m_o} \right) u_{n\vec{k}} \quad (7.228a)$$

where

$$H_o = \frac{\hat{p}^2}{2m_o} + V_L \quad (7.228b)$$

One can note the overall form of Equation 7.228a. The eigenvalue on the right side consists of the difference between dispersion curves. The first dispersion curve has the bandgaps while the second one corresponds to the free electron. That is, the eigenvalue must represent the difference between the two curves shown in Figure 7.23 or 7.33. Equation 7.228a by virtue of the second term on the left hand side, shows that the $k \cdot p$ term gives rise to the changes in the E_n dispersion curve. Notice that $k \sim 0$, the eigenvalue would have a similar form but slightly different values for the $E_n(k)$

Let us examine specific terms in Equation 7.228a. The second term on the left serves as a perturbation since long wavelengths produce small k ; this is especially true at the band edges (i.e., extrema) near $k = 0$ where many devices operate. Notice that the free mass of the electron appears in Equation 7.228a. For $k = 0$, we find

$$H_0 u_{n0} = E_n(0) u_{n0} \quad (7.229)$$

and so $E_n(0)$ gives the position of the band edge. Keep in mind that we are looking for the dispersion curves given by $E_n(\vec{k})$ which includes the effects of the $\vec{k} \cdot \vec{p}$ term. Defining the energy

$$W_n(\vec{k}) = E_n(\vec{k}) - \frac{\hbar^2 k^2}{2m_0} \quad (7.230a)$$

The perturbation theory expands around $k = 0$ so that the lowest order term must be at $k = 0$

$$W_n^{(0)}(0) = E_n(0) \quad (7.230b)$$

Once we find the first and second order terms, we can write

$$W_n(\vec{k}) \cong W_n^{(0)} + W_n^{(1)} + W_n^{(2)} = E_n(0) + W_n^{(1)} + W_n^{(2)} \quad (7.230c)$$

and therefore substituting Equation 7.230a into Equation 7.230c provides

$$E_n(\vec{k}) = E_n(0) + \frac{\hbar^2 k^2}{2m_0} + W_n^{(1)} + W_n^{(2)} \quad (7.230d)$$

We see that the band retains its basic parabolic shape because of the k^2 term. The effective mass must come from the last two terms in the last equation.

7.11.3 NONDEGENERATE BANDS

This section shows how the effective mass arises in band theory rather than the quasiphenomenological description given in previous sections. We will also see that the periodic portion of the Bloch wave function has weak dependence on the wave vector k .

Equation 7.228a can be treated using perturbation theory from Chapter 5. We will take the perturbing potential in Equation 7.228 as

$$\hat{V} = \frac{\hbar}{m_0} \vec{k} \cdot \hat{p} \quad (7.231)$$

Section 5.9 shows the correction terms to W in Equation 7.230c can be written as

$$W_m = E_m(0) + \mathcal{V}_{mm} + \sum_{n \neq m} \frac{|\mathcal{V}_{mn}|^2}{E_m(0) - E_n(0)} \quad (7.232)$$

where the matrix elements are taken using the zeroth order correction to the basis functions u , namely u_{n0} (at $\vec{k} = 0$) and we divide by the zeroth order correction to the energy. Substituting 7.231 and 7.230a, we find

$$E_m(\vec{k}) = E_m(0) + \frac{\hbar^2 k^2}{2m_0} + \mathcal{V}_{mm} + \sum_{n \neq m} \frac{|\mathcal{V}_{mn}|^2}{E_m(0) - E_n(0)} \quad (7.233)$$

where $E_m(0)$ must represent the band at $k=0$ —for GaAs CB this must be the minimum.

One can evaluate the matrix elements in Equation 7.233. Consider the diagonal terms $\mathcal{V}_{mm}$. First, recall the matrix elements in the perturbation expansion use the zeroth order basis function found by setting $k=0$, i.e., $u_{m0}(\vec{r})$. One can show that $\mathcal{V}_{mm}=0$ so long as the crystal has a center of symmetry, which means $u_{m0}(-\vec{r}) = \pm u_{m0}(\vec{r})$. This can be demonstrated using two methods. The first and most elegant method defines the parity operator $\hat{P}$ to give the new wave function corresponding to the replacement $\vec{r} \rightarrow -\vec{r}$; that is $\hat{P}|u_{n0}\rangle = \pm|u_{n0}\rangle$. A similarity transformation gives the results of the interchange $\vec{r} \rightarrow -\vec{r}$ for operators. For example, in the case of 1-D, we find

$$\hat{P}^+ p_x \hat{P} = \hat{P}^+ \frac{\hbar}{i} \frac{d}{dx} \hat{P} = \frac{\hbar}{i} \frac{d}{d(-x)} = -\frac{\hbar}{i} \frac{d}{dx} = -p_x$$

Similar statements can be made regarding the y- and z-components. Therefore, we can calculate the matrix elements

$$\langle u_{n0} | \vec{p} | u_{n0} \rangle = [\pm |u_{n0}\rangle]^+ \vec{p} [\pm |u_{n0}\rangle] = [\hat{P}|u_{n0}\rangle]^+ \vec{p} [\hat{P}|u_{n0}\rangle] = \langle u_{n0} | \hat{P}^+ \vec{p} \hat{P} | u_{n0} \rangle = -\langle u_{n0} | \vec{p} | u_{n0} \rangle$$

Therefore we conclude

$$\langle u_{n0} | \vec{p} | u_{n0} \rangle = 0 \quad \text{and} \quad \mathcal{V}_{mm} = 0 \quad (7.234)$$

The second method just writes the integral for $\langle u_{n0} | \vec{p} | u_{n0} \rangle$ and substitutes $-x$ for x to get the same results.

The second set of matrix elements $\mathcal{V}_{mn}$ are not necessarily zero, which gives rise to the effective mass. The energy eigenvalues can be written by substituting Equation 7.231 into 7.233.

$$E_m(\vec{k}) = E_m(0) + \frac{\hbar^2 k^2}{2m_0} + \left(\frac{\hbar}{m_0} \right)^2 \sum_{n \neq m} \frac{|\vec{k} \cdot \hat{p}_{nm}|^2}{E_m(0) - E_n(0)} \quad (7.235)$$

Breaking the vectors into components, and defining the indices α, β to take on the values 1, 2, 3 to represent the x-, y-, z-components, we can write

$$k^2 = k_x^2 + k_y^2 + k_z^2 = \sum_{\alpha} k_{\alpha} k_{\alpha} = \sum_{\alpha, \beta} k_{\alpha} k_{\beta} \delta_{\alpha\beta} \quad (7.236a)$$

where $\delta_{\alpha\beta}$ is the Kronecker delta. We can also rewrite $|\vec{k} \cdot \hat{p}_{nm}|^2$ in terms of components.

$$|\vec{k} \cdot \hat{p}_{nm}|^2 = (\vec{k} \cdot \hat{p}_{nm})^* (\vec{k} \cdot \hat{p}_{nm}) = \left(\sum_{\alpha} k_{\alpha} \hat{p}_{nm}^{(\alpha)} \right)^* \left(\sum_{\beta} k_{\beta} \hat{p}_{nm}^{(\beta)} \right) = \sum_{\alpha, \beta} k_{\alpha} \hat{p}_{nm}^{(\alpha)*} k_{\beta} \hat{p}_{nm}^{(\beta)} \quad (7.236b)$$

where $\hat{p}_{nm}^{(1)} = \langle u_{n0} | p_x | u_{m0} \rangle = \langle u_{n0} | \frac{\hbar}{i} \frac{\partial}{\partial x} | u_{m0} \rangle$ and so on, and the wave vectors k are real. Next recall that the complex conjugate of a Hermitian matrix element has the same effect as taking the adjoint so that $\hat{p}_{nm}^{(\alpha)*} = \hat{p}_{mn}^{(\alpha)}$. Therefore the last equation becomes

$$|\vec{k} \cdot \hat{p}_{nm}|^2 = \sum_{\alpha, \beta} \hat{p}_{mn}^{(\alpha)} \hat{p}_{nm}^{(\beta)} k_\alpha k_\beta \quad (7.236c)$$

Combining Equations 7.236 with 7.235 produces

$$E_m(\vec{k}) - E_m(0) = \sum_{\alpha, \beta} \left[\frac{\hbar^2 \delta_{\alpha\beta}}{2m_0} + \left(\frac{\hbar}{m_0} \right)^2 \sum_{n \neq m} \frac{p_{mn}^{(\alpha)} p_{nm}^{(\beta)}}{E_m(0) - E_n(0)} \right] k_\alpha k_\beta \quad (7.237)$$

Now we can find the effective mass. Section 7.7 shows the effective mass tensor can be defined as follows

$$E_m(\vec{k}) - E_m(0) = \frac{\hbar^2}{2} \sum_{\alpha, \beta} \left(\frac{1}{m^*} \right)_{\alpha\beta} k_\alpha k_\beta \quad (7.238a)$$

where the indices α, β take on the values 1, 2, 3 which symbolize the x -, y -, z -directions. Comparing Equations 7.237 and 7.238 we find

$$\left(\frac{1}{m^*} \right)_{\alpha\beta} = \left[\frac{\delta_{\alpha\beta}}{m_0} + \frac{2}{m_0^2} \sum_{n \neq m} \frac{p_{mn}^{(\alpha)} p_{nm}^{(\beta)}}{E_m(0) - E_n(0)} \right] \quad (7.238b)$$

where we assume that the electron occupies band m . Those bands with energy larger than the one under consideration tend to make the effective mass larger than the free mass. Those bands with energy smaller than the one under consideration tend to decrease the effective mass.

Finally, we demonstrate how the periodic Bloch function u depends on the wave vector k . We can verify that it depends on k only in first order perturbation theory. Section 5.10 shows the wave function $u_{n,\vec{k}}$ to first order approximation can be expressed as

$$|u_{n\vec{k}}\rangle \cong |u_{n0}\rangle - \sum_{m \neq n} \frac{\mathcal{V}_{mn}}{E_m(0) - E_n(0)} |u_{m0}\rangle \quad (7.239a)$$

where the matrix elements V_{nm} are taken using the zeroth order wave functions

$$\mathcal{V}_{mn} = \langle u_{m0} | \hat{\mathcal{V}} | u_{n0} \rangle \quad \text{where} \quad \hat{\mathcal{V}} = \frac{\hbar}{m_0} \vec{k} \cdot \hat{p} \quad (7.239b)$$

This time the first order correction is not zero. Therefore we see that the periodic part of the extended states depend on k only through the first order correction. For example, suppose the electron is in the conduction band (in a two-band semiconductor) so that $m = 2$. For this example, the only other band is the valence band giving $n = 1$. For a wide bandgap semiconductor, $E_g = |E_n(0) - E_m(0)|$ is quite large and we expect the second term on the right-hand side of Equation 7.239a to be small. Of particular significance is fact that the function $u_{n\vec{k}}$ can be expanded in terms of the function u_{n0} as

$$u_{n\vec{k}} = \sum_m a_m(\vec{k}) u_{m0} \quad (7.240)$$

where $a_m(\vec{k})$ represents the expansion coefficients in Equation 7.239a. In general, 7.240 contains all order corrections and not just the first few.

In summary, the $\vec{k} \cdot \vec{p}$ theory modifies the parabolic bands found for free electrons. We could have made an expansion around other k -vectors rather than zero. We then expect to find the actual band structure around this different vector. The $\vec{k} \cdot \vec{p}$ theory predicts an effective mass for the electron as in Equation 2.38b. The effective mass can be anisotropic as described by Equation 7.155b in the discussion on tensor mass.

7.11.4 $\vec{k} \cdot \vec{p}$ THEORY FOR TWO NONDEGENERATE BANDS

Now consider the $\vec{k} \cdot \vec{p}$ theory for two nondegenerate bands using a determinant method. This section show the dispersion curve for one of the bands and the effective mass. Let the set $u_{n,\vec{k}=0}$ denote the basis set and require $u_{n\vec{k}} = \sum_m a_m(\vec{k}) u_{m0}$ to hold.

The Schrödinger equation for the periodic part of the Bloch wave function $u_{n\vec{k}}$ appears in Equation 7.228a

$$\left(H_0 + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar}{m_0} \vec{k} \cdot \hat{p} \right) u_{n\vec{k}} = E_n(\vec{k}) u_{n\vec{k}} \quad (7.241a)$$

where

$$H_0 = \frac{\hbar^2 \hat{p}^2}{2m_0} + V_L \quad (7.241b)$$

and V_L represents the periodic lattice potential and

$$\hat{H}_0 |u_{n0}\rangle = E_n(0) |u_{n0}\rangle \quad (7.241c)$$

As shown in Equation 7.240, the functions $u_{n0}(\vec{r})$ form a complete set so that any of the functions $u_{n\vec{k}}(\vec{r})$ can be expanded in terms of them. We can write

$$u_{n\vec{k}} = \sum_m a_m(\vec{k}) u_{m0} \quad (7.242)$$

Substituting this last expression into Equation 7.241a provides

$$\sum_m \left(E_n(0) |u_{m0}\rangle + \frac{\hbar^2 k^2}{2m_0} |u_{m0}\rangle + \frac{\hbar}{m_0} \vec{k} \cdot \hat{p} |u_{m0}\rangle \right) a_m = \sum_m E_n(\vec{k}) a_m |u_{m0}\rangle$$

Operate with $\langle u_{n0} |$ to find

$$\sum_m \left(E_n(0) \delta_{mn} + \frac{\hbar^2 k^2}{2m_0} \delta_{mn} + \frac{\hbar}{m_0} \vec{k} \cdot \vec{p}_{nm} \right) a_m = E_n(\vec{k}) a_n \quad (7.243)$$

which has the form of an eigenvalue equation.

For simplicity, consider two bands with the index 1 for the valence band and index 2 for the conduction band. Assume we are looking for the conduction band for $n=2$. Equation 7.243 can be written as a matrix equation with the Kronecker delta terms providing the diagonal elements.

$$\begin{bmatrix} E_2(0) + \frac{\hbar^2 k^2}{2m_0} - E_2(\vec{k}) & \frac{\hbar}{m_0} \vec{k} \cdot \vec{p}_{21} \\ \frac{\hbar}{m_0} \vec{k} \cdot \vec{p}_{12} & E_1(0) + \frac{\hbar^2 k^2}{2m_0} - E_2(\vec{k}) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 0 \quad (7.244)$$

where $\vec{p}_{11} = 0 = \vec{p}_{22}$ as previously discussed under Equation 7.233. We want the eigenvalues $E_2(\vec{k})$ which gives the dispersion curve for the conduction band.

For simplicity, assume the valence band maximum occurs at $E_1(\vec{k} = 0) = 0$ and the bandgap energy is $E_g = E_2(0) - E_1(0) = E_2(0)$. We require the determinant in Equation 7.244 to yield zero. Defining

$$E'_{2,k} = E_2(\vec{k}) - \frac{\hbar^2 k^2}{2m_0} \quad (7.245)$$

the determinant becomes

$$(E_g - E'_{2,k})(-E'_{2,k}) = \frac{\hbar^2}{m_0^2} (\vec{k} \cdot \vec{p}_{12})(\vec{k} \cdot \vec{p}_{21}) \quad (7.246)$$

We assume that $E'_{2,k} \ll E_g$ near the bottom of the conduction band. Solving Equation 7.246 for $E_2(\vec{k}) = E_{2,\vec{k}}$ provides

$$E_{2,k} = \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar^2}{m_0^2} \frac{(\vec{k} \cdot \vec{p}_{12})(\vec{k} \cdot \vec{p}_{21})}{E_g} \quad (7.247)$$

If we assume isotropic bands, the dispersion curve reduces to the simple form

$$E_{2,k} = \frac{\hbar^2 k^2}{2m_{\text{eff}}}$$

The development clearly shows the origins of the effective mass. It depends on the matrix element of the momentum between bands. Only nearest bands are important because of the factor of E_g .

7.12 INTRODUCTION TO $\vec{k} \cdot \vec{p}$ THEORY FOR DEGENERATE BANDS

The $\vec{k} \cdot \vec{p}$ theory accounts for the effective mass of the electron, the band shape, and the periodic wave function. The Kane version of the $\vec{k} \cdot \vec{p}$ theory describes these quantities for the case of degenerate bands. The coupling between bands produces a perturbation. However, the usual form of the Kane approximation breaks down and does not accurately predict the effective mass of the heavy-hole electrons since the calculation does not include a sufficient number of bands. The present section should provide enough details to enable the reader to use the references that summarize the results of the model and apply the Luttinger–Kane model. The first-time reader can safely skip over the topics without loss of continuity with ensuing sections.

7.12.1 SUMMARY OF CONCEPTS AND PROCEDURE

The Kane model starts with the same Hamiltonian as for the nondegenerate $\vec{k} \cdot \vec{p}$ theory $\hat{H}_{\text{ND}}$ and then includes the spin-orbit interaction $\hat{H}_{\text{LS}}$ that splits off one of the valence bands (refer to Chuang's book or Yu's book and their references). The procedure looks for eigenfunctions $\psi_{n\vec{k}}$ and eigenvalues $E_{n\vec{k}}$ satisfying the eigenvalue equation

$$\hat{H}\psi_{n\vec{k}} = E_{n\vec{k}}\psi_{n\vec{k}} \quad (7.248)$$

with n representing the band index and $\hat{H} = \hat{H}_{\text{ND}} + \hat{H}_{\text{LS}}$. Each $E_{n\vec{k}}$ gives a different dispersion curve (for different band index n). As with the nondegenerate $\vec{k} \cdot \vec{p}$ theory, we write the wave function $\psi_{n\vec{k}}$ in the Bloch form

$$\psi_{n\vec{k}} = \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} u_{n\vec{k}} \quad (7.249)$$

where note the use of (Ω) to represent the crystal normalization volume rather than $\mathcal{V}$ so as not to cause confusion with potential.

Substituting Equation 7.249 into Equation 7.248 produces an eigenvalue equation for the periodic functions $u_{n\vec{k}}$.

$$\hat{\mathcal{H}}|u_{n\vec{k}}\rangle = E_{n\vec{k}}|u_{n\vec{k}}\rangle \quad (7.250)$$

In addition to the finding the dispersion curves $E_{n\vec{k}}$, look for the periodic part of the Bloch function. We will find the $u_{n\vec{k}}$ are essentially linear combinations of the sp orbitals described in Section 6.1 regarding the physical origin of bonding.

To solve the eigenvalue Equation 7.250, convert it to a matrix equation by assuming an initial basis set. Essentially, one then finds a new basis that makes the Hamiltonian diagonal. The initial basis set consists of the periodic part of the Bloch wave functions near the band edges u_{n0} (i.e., near the band extremum); these will be taken as certain linear combinations of the s and p orbitals.

$$|u_{n\vec{k}}\rangle = \sum_{\beta} a_{\beta\vec{k}}^{(n)} |u_{\beta 0}\rangle \quad (7.251)$$

Notice that the expansion coefficients must depend on the wave vector. The summation combines the s and p orbitals as represented by the band edge functions. Substituting Equation 7.251 into Equation 7.250 and operating with $\langle u_{\alpha 0}|$ produces the matrix equation

$$[\mathcal{H}_{\alpha\beta}][a_{\beta}] = E[a_{\beta}] \quad (7.252)$$

The determinant of $\underline{\mathcal{H}} - E\underline{1}$ (where $\underline{1}$ represents a unit matrix) produces the set of eigenvalues E . Each such eigenvalue produces a dispersion curve. In this case, we find the conduction band (CB), the light-hole (LH) band, the heavy-hole (HH) band, and the split-off (SO) band as shown in Figure 7.55.

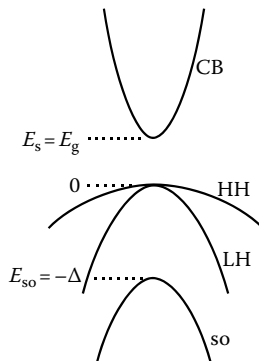


FIGURE 7.55 The four most important bands in GaAs.

Substituting each different eigenvalue E back into Equation 7.252 produces different expansion coefficients $\{a_\beta\}$. Each different set of coefficients produces a different eigenfunction $u_{n\vec{k}}$ using Equation 7.251.

The Kane model shown in this section only uses the four bands CB, LH, HH, SO. We will find an effective mass for each band except for the heavy-hole band. More bands need to be included in the model in order to correctly provide an effective mass for the HH band. The problem with the HH band can be handled using the Luttinger–Kane model. Refer to Chang’s and Yu’s book in the references.

7.12.2 HAMILTONIAN FOR KANE’S MODEL

The Hamiltonian for Kane’s version of the $\vec{k} \cdot \vec{p}$ model (i.e., for degenerate bands) has essentially the same form as for the nondegenerate case but with the addition of a spin-orbit interaction.

$$\hat{H} = \frac{\hat{p}^2}{2m_0} + V(\vec{r}) + \frac{\hat{S} \cdot \nabla V \times \hat{p}}{2m_0^2 c^2} \quad (7.253)$$

where m_0 is the free mass, the spin and Pauli operators are

$$\hat{S} = \hbar \hat{\sigma} / 2 \quad \hat{\sigma} = \tilde{x} \hat{\sigma}_x + \tilde{y} \hat{\sigma}_y + \tilde{z} \hat{\sigma}_z \quad (7.254)$$

and

$$\underline{\sigma}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \underline{\sigma}_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \underline{\sigma}_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (7.255)$$

The term $\hat{S} \cdot \nabla V \times \hat{p}$ comes from the interaction of the magnetic field produced by the electron spin with a magnetic field indirectly related to the Coulomb field (electric field) produced by the atomic nucleus. The protons in the nucleus produce an electric field. The field extends to points external to the nucleus but with reduced magnitude due to screening by any orbiting electrons. Consider a classical spinning electron moving near the atom. And consider a reference frame (i.e., coordinate system) moving with the same translational motion as the electron so that an observer in this frame sees only the spinning electron motion and not its translational motion. The observer sees the charged nucleus move. Therefore, the observer sees that the moving nuclear charge must produce a magnetic field. Furthermore, the observer finds the spinning electron produces a magnetic dipole field. These two magnetic fields interact and alter the energy of the spinning electron in its orbit.

The interaction energy between the electron magnetic dipole and the magnetic field produced by the translational motion of the nucleus must be similar to that discussed in Section 5.6 due to a spinning electron in an external magnetic field.

$$\text{Energy} \sim -\vec{B}_{\text{nuc}} \cdot \vec{B}_{\text{elect}} \quad (7.256)$$

The magnetic field related to the “moving” electric field produced by the nucleus has the form $\vec{B}_{\text{nuc}} \sim \vec{v} \times \vec{E}$. The magnetic field due to the electron spin has the form $\vec{B}_{\text{elect}} \sim -\hat{S}$. Further, the electric field to lowest order approximation can be represented as the radial derivative of the potential.

$$\vec{E} = -\nabla V \sim -\hat{r} \frac{\partial V}{\partial r} = -\vec{r} \frac{1}{r} \frac{\partial V}{\partial r}$$

Combining these last expressions with Equation 7.256 produces

$$\text{Energy} \sim -\vec{B}_{\text{nuc}} \cdot \vec{B}_{\text{elect}} \sim \vec{v} \times \vec{E} \cdot \hat{S} \sim -\frac{1}{r} \frac{\partial V}{\partial r} \vec{v} \times \vec{r} \cdot \hat{S} \sim \frac{1}{r} \frac{\partial V}{\partial r} \hat{L} \cdot \hat{S} \quad (7.257)$$

since the orbital angular momentum has the form $\vec{L} = \vec{r} \times \vec{p}$. Equation 7.257 shows the origin of the name “spin-orbit interaction.” This interaction produces the “split-off valence band.”

We use the interaction Hamiltonian in Equation 7.257 in a slightly different form. Leaving the electric field in terms of the gradient and returning to the momentum rather than the angular momentum, we have

$$\hat{H}_{\text{LS}} = \frac{\hbar}{4m_0^2 c^2} \hat{\sigma} \cdot \nabla V \times \hat{p} \quad (7.258a)$$

where $\hat{S} = \hbar \hat{\sigma} / 2$. The full Hamiltonian becomes

$$\hat{H} = \hat{H}_0 + \hat{H}_{\text{LS}} = \frac{\hat{p}^2}{2m_0} + V + \frac{\hbar}{4m_0^2 c^2} \hat{\sigma} \cdot \nabla V \times \hat{p} \quad (7.258b)$$

where V refers to the atomic potential and must be the same for all sites in the crystal.

7.12.3 EIGEN EQUATION FOR PERIODIC BLOCH STATES

As discussed in Section 7.12.1, the Bloch eigenstates satisfy the eigenvector equation using the Hamiltonian in Equation 7.258b

$$\left[\frac{\hat{p}^2}{2m_0} + V + \frac{\hbar}{4m_0^2 c^2} \hat{\sigma} \cdot \nabla V \times \hat{p} \right] \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} u_{n\vec{k}} = E_{n\vec{k}} \frac{e^{i\vec{k} \cdot \vec{r}}}{\sqrt{\Omega}} u_{n\vec{k}} \quad (7.259)$$

where m_0 represents the free mass of the electron. Carrying out the differentiation for the coordinate representation of the momentum operator, and then dividing out the envelope portion of the wave function, we find

$$\left[\frac{\hat{p}^2}{2m_0} + V + \vec{k} \cdot \hat{p} + \frac{\hbar}{4m_0^2 c^2} \nabla V \times \hat{p} \cdot \hat{\sigma} + \frac{\hbar^2}{4m_0^2 c^2} \nabla V \times \vec{k} \cdot \hat{\sigma} \right] u_{n\vec{k}} = E_{n\vec{k}} u_{n\vec{k}} \quad (7.260a)$$

where

$$E_{n\vec{k}} = E_{n\vec{k}} - \frac{\hbar^2 k^2}{2m_0} \quad (7.260b)$$

Assume the crystal momentum $\hbar \vec{k}$ is small compared with the orbital momentum of the electron and therefore drop the fifth term in Equation 7.260a

$$\left[\frac{\hat{p}^2}{2m_0} + V + \vec{k} \cdot \hat{p} + \frac{\hbar}{4m_0^2 c^2} \nabla V \times \hat{p} \cdot \hat{\sigma} \right] u_{n\vec{k}} = E_{n\vec{k}} u_{n\vec{k}} \quad (7.260c)$$

It is customary to assume the electron travels along the z -axis so that $\vec{k} = k\hat{z}$ and then

$$\left[\frac{\hat{p}^2}{2m_0} + V + k\hat{p}_z + \frac{\hbar}{4m_0^2 c^2} \nabla V \times \hat{p} \cdot \hat{\sigma} \right] u_{n\vec{k}} = E_{n\vec{k}} u_{n\vec{k}} \quad (7.260d)$$

For later convenience, we define the Hamiltonian

$$\hat{\mathcal{H}} = \frac{\hat{p}^2}{2m_0} + V + k\hat{p}_z + \frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} \cdot \hat{\sigma} = \hat{H}_0 + k\hat{p}_z + \frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} \cdot \hat{\sigma} \quad (7.261)$$

where $\hat{H}_0$ must be the Hamiltonian for the atomic orbital. Changing the direction of travel necessarily changes the form of the eigenvectors. Any weak dependence of the eigenfunctions u on k must come from the third term $k\hat{p}_z$ on the right-hand side of this last equation.

As discussed in Section 7.12.1, we must start with a basis set in order to write Equation 7.260d in the form of a matrix. We will use the complete set of functions $\{u_{n0}\}$ centered on $k=0$ where n represents the band index. The next section shows the connection between these functions and the atomic orbitals. We will use only four bands “CB, HH, LH, SO” so that $n=1 \dots 4$. The eigenfunction $u_{n\vec{k}}$ must be a linear combination of the functions $\{u_{n0}\}$

$$|u_{n\vec{k}}\rangle = \sum_{\beta} a_{\beta\vec{k}}^{(n)} |u_{\beta 0}\rangle \quad (7.262)$$

We need to find the energy eigenvalues and then the coefficients $a_{\beta\vec{k}}^{(n)}$ in order to find the eigenvectors $|u_{n\vec{k}}\rangle$. The inner products for these states extend over the unit cell.

The matrix equation for Equations 7.260d and 7.261 can be found by substituting Equation 7.262 into Equation 7.260d and then operating with $\langle u_{\alpha 0}|$ as follows.

$$\sum_{\beta} a_{\beta\vec{k}}^{(n)} \langle u_{\alpha 0} | \hat{\mathcal{H}} | u_{\beta 0} \rangle = \mathcal{E}_{\alpha\vec{k}} a_{\alpha\vec{k}}^{(n)} \rightarrow \sum_{\beta} \hat{\mathcal{H}}_{\alpha\beta} a_{\beta\vec{k}}^{(n)} = \mathcal{E}_{\alpha\vec{k}} a_{\alpha\vec{k}}^{(n)} \rightarrow [\mathcal{H}][a] = \mathcal{E}_{\alpha\vec{k}}[a] \quad (7.263)$$

We need to find the matrix of the Hamiltonian in order to proceed. Therefore, we must specify the starting basis set $\{u_{n0}\}$.

7.12.4 INITIAL BASIS SET

Recall from Section 7.1, the definitions of the s and p atomic orbitals. The $|s\rangle$ state has spherical symmetry and can be related to the basis set for angular momentum according to

$$|s\rangle = |\ell=0, \ell_z=0\rangle = Y_{\ell=0, \ell_z=0}(\theta, \phi) = \frac{1}{\sqrt{4\pi}} \quad (7.264)$$

In the s state, the wave function does not have any angular variation. It obviously has even parity $\hat{P}|s\rangle = +1|s\rangle$.

The p-orbitals correspond to the lowest nonzero orbital angular momentum states.

$$|\phi_x\rangle = |X\rangle = \frac{1}{\sqrt{2}} \{|\ell=1, \ell_z=1\rangle - |\ell=1, \ell_z=-1\rangle\} \sim \frac{1}{\sqrt{2}} \{Y_{1,1} + Y_{1,-1}\} = \sqrt{\frac{3}{4\pi}} \frac{x}{r} \quad (7.265a)$$

$$|\phi_y\rangle = |Y\rangle = \frac{i}{\sqrt{2}} \{|\ell=1, \ell_z=1\rangle + |\ell=1, \ell_z=-1\rangle\} \sim \frac{i}{\sqrt{2}} \{Y_{1,1} - Y_{1,-1}\} = \sqrt{\frac{3}{4\pi}} \frac{y}{r} \quad (7.265b)$$

$$|\phi_z\rangle = |Z\rangle = |\ell=1, \ell_z=0\rangle \sim Y_{10}(\theta, \phi) = \sqrt{\frac{3}{4\pi}} \cos\theta = \sqrt{\frac{3}{4\pi}} \frac{z}{r} \quad (7.265c)$$

where $Y_{\ell m}$ represents a spherical harmonic. We will normally use the uppercase letters to avoid confusing the orbitals with the linear momentum. It is easy to see that the states in Equation 7.265a through c are orthonormal. The symbols X, Y, Z indicate two properties. First, a symbol represents

the direction of odd parity with the other directions having even parity. Second, it gives the “direction” in which the p orbital “points” as described in Section 7.1. These states are convenient for their parity properties.

The basis states for the degenerate band theory consist of linear combinations of those in Equation 7.265a through c. Near the valence band edges (i.e., the maximum) the states u_{n0} most closely match the p orbitals. Near the conduction band edge, the states resemble the s orbitals. We must include spin up and spin down (as represented by upward and downward pointing arrows respectively). We use the states that are linear combinations of those in Equation 7.265a through c and reduce to the spherical harmonics.

State Number	Basis State
1	$ iS\downarrow\rangle = i S\downarrow\rangle$
2	$\left \frac{x-iy}{\sqrt{2}}\uparrow\right\rangle = \frac{1}{\sqrt{2}}(X\uparrow\rangle - i Y\uparrow\rangle)$
3	$ Z\downarrow\rangle$
4	$\left -\frac{x+iy}{\sqrt{2}}\uparrow\right\rangle = \frac{1}{\sqrt{2}}(X\uparrow\rangle + i Y\uparrow\rangle)$

Four other states come from reversing the spin in those listed above.

The parity of the states allows us to quickly and efficiently reduce matrix elements as seen in the next topic.

$$\langle S|\hat{p}_i|\phi_j\rangle = 0 \quad \text{when } i \neq j$$

where $\phi_x = X$ (and so on). For example, consider $i = x, j = y$. Then insert the parity operator for x using $\hat{1} = \hat{p}_x^+ \hat{p}_x$

$$\langle S|\hat{p}_x|\phi_y\rangle = \langle S|\hat{1}\hat{p}_x\hat{1}|\phi_y\rangle = (\langle S|\hat{p}_x^+)\hat{p}_x\hat{p}_x^+(\hat{p}_x|\phi_y\rangle) = -\langle S|\hat{p}_x|\phi_y\rangle$$

since both S and $\phi_y = Y$ are symmetric in x , and

$$\hat{p}_x\hat{p}_x\hat{p}_x^+ = \hat{p}_x\frac{\hbar}{i}\frac{\partial}{\partial x}\hat{p}_x^+ = \frac{\hbar}{i}\frac{\partial}{\partial(-x)} = -\hat{p}_x$$

7.12.5 MATRIX OF HAMILTONIAN

We now demonstrate the matrix of the Hamiltonian using the basis states

$$\mathcal{H} = \begin{bmatrix} E_s & 0 & kP & 0 \\ 0 & E_p - \Delta/3 & \sqrt{2}\Delta/3 & 0 \\ kP & \sqrt{2}\Delta/3 & E_p & 0 \\ 0 & 0 & 0 & E_p + \Delta/3 \end{bmatrix} \quad (7.266a)$$

where the Kane parameter P and the SO energy Δ have the form

$$P = \frac{-i\hbar}{m_0} \langle S|p_z|Z\rangle \quad \Delta = \frac{3\hbar i}{4m_0^2 c^2} \langle X|\frac{\partial V}{\partial x}p_y - \frac{\partial V}{\partial y}p_x|Y\rangle \quad (7.266b)$$

We first consider the term $\mathcal{H}_{11} = \langle iS \downarrow | \hat{\mathcal{H}} | iS \downarrow \rangle$ using Equation 7.261.

$$\hat{\mathcal{H}} = \underbrace{\hat{H}_0}_1 + \underbrace{k\hat{p}_z}_2 + \underbrace{\frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} \cdot \hat{\sigma}}_3 \quad (7.267)$$

where $\hat{H}_0$ represents the energy of the atomic orbitals. The term #1 provides

$$\langle iS \downarrow | \#1 | iS \downarrow \rangle = \langle iS \downarrow | \hat{H}_0 | iS \downarrow \rangle = (-i)(i) \langle S | \hat{H}_0 | S \rangle \langle \downarrow | \downarrow \rangle = +1 \langle S | \hat{H}_0 | S \rangle = E_s \quad (7.268a)$$

Term #2 gives

$$\langle iS \downarrow | \text{Term #2} | iS \downarrow \rangle = \langle iS \downarrow | k\hat{p}_z | iS \downarrow \rangle = k \langle S | \hat{p}_z | S \rangle \langle \downarrow | \downarrow \rangle = 0 \quad (7.268b)$$

Since the state $\hat{p}_z | S \rangle$ has odd parity in z while the state $\langle S |$ has even parity and therefore the integral produces zero. The inner products over the orbital states extend over the unit cell. Term #3 in Equation 7.267 has the spin operator.

$$\langle iS \downarrow | \text{Term #3} | iS \downarrow \rangle = \langle iS \downarrow | \frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} \cdot \hat{\sigma} | iS \downarrow \rangle = \langle S | \frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} | S \rangle \cdot \langle \downarrow | \hat{\vec{\sigma}} | \downarrow \rangle$$

where $\hat{\vec{\sigma}} = \hat{\sigma} = \hat{\sigma}_x \tilde{x} + \hat{\sigma}_y \tilde{y} + \hat{\sigma}_z \tilde{z}$. The last term has produces

$$\langle \downarrow | \hat{\vec{\sigma}} | \downarrow \rangle = \tilde{x} \langle \downarrow | \hat{\sigma}_x | \downarrow \rangle + \tilde{y} \langle \downarrow | \hat{\sigma}_y | \downarrow \rangle + \tilde{z} \langle \downarrow | \hat{\sigma}_z | \downarrow \rangle = \tilde{x} \langle \downarrow | \uparrow \rangle - i\tilde{y} \langle \downarrow | \uparrow \rangle - \tilde{z} \langle \downarrow | \downarrow \rangle = -\tilde{z}$$

Therefore the matrix element of the third term produces

$$\langle iS \downarrow | \text{Term #3} | iS \downarrow \rangle = -\langle S | \frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} \cdot \tilde{z} | S \rangle = -\frac{\hbar}{4m_0^2c^2} \langle S | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | S \rangle$$

For a crystal symmetric in the interchange of the x - and y -coordinates, this last term must be zero since interchanging x and y produces an extra minus sign. The full 11 matrix element comes from combining the three terms.

Next consider the matrix element $\mathcal{H}_{22} = \left\langle \frac{X-iY}{\sqrt{2}} \uparrow \left| \hat{\mathcal{H}} \right| \frac{X-iY}{\sqrt{2}} \uparrow \right\rangle$ using the Hamiltonian

$$\hat{\mathcal{H}} = \underbrace{\hat{H}_0}_1 + \underbrace{k\hat{p}_z}_2 + \underbrace{\frac{\hbar}{4m_0^2c^2} \nabla V \times \hat{p} \cdot \hat{\sigma}}_3 \quad (7.269)$$

Consider Term #1. Assuming the orbitals X and Y have the same energy (no external magnetic field), we find

$$\left\langle \frac{X}{\sqrt{2}} \left| H_0 \right| \frac{X}{\sqrt{2}} \right\rangle + \left\langle \frac{iY}{\sqrt{2}} \left| H_0 \right| \frac{iY}{\sqrt{2}} \right\rangle = \frac{1}{2} \{ \langle X | H_0 | X \rangle + \langle Y | H_0 | Y \rangle \} = \frac{1}{2} \{ E_p + E_p \} = E_p$$

The other parts of Term #1 must be zero since they are eigenfunctions and produce

$$\langle X | \hat{H}_0 | Y \rangle = E_p \langle X | Y \rangle = 0$$

Term #2 does not contribute since $\hat{p}_z$ has negative parity in z whereas X and Y have positive parity.

Finally, Term #3 is a little more complicated.

$$\left\langle \frac{X - iY}{\sqrt{2}} \uparrow \left| \text{Term \#3} \right| \frac{X - iY}{\sqrt{2}} \uparrow \right\rangle = \frac{\hbar}{4m_0^2 c^2} \left\langle \frac{X - iY}{\sqrt{2}} \uparrow \left| \nabla V \times \hat{p} \cdot \hat{\sigma} \right| \frac{X - iY}{\sqrt{2}} \uparrow \right\rangle$$

Only the z -component of the spin operator contributes since the others flip the spin. The cross product becomes

$$\tilde{z}\hat{\sigma}_z \cdot \nabla V \times \hat{p} = \begin{vmatrix} 0 & 0 & \hat{\sigma}_z \\ \partial_x V & \partial_y V & \partial_z V \\ \hat{p}_x & \hat{p}_y & \hat{p}_z \end{vmatrix} = (\partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x) \hat{\sigma}_z$$

where we were careful to maintain the order. The decimal point in $\partial_x V \cdot \hat{p}_y$ indicates that the derivative applies only to the potential V (we do not think of ∂_x as the momentum operator either). Now using the spin operator, we find

$$\left\langle \frac{X - iY}{\sqrt{2}} \uparrow \left| \text{Term \#3} \right| \frac{X - iY}{\sqrt{2}} \uparrow \right\rangle = \frac{\hbar}{4m_0^2 c^2} \left\langle \frac{X - iY}{\sqrt{2}} \uparrow \left| \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x \right| \frac{X - iY}{\sqrt{2}} \uparrow \right\rangle$$

The matrix elements $\langle X | \text{Term \#3} | X \rangle + \langle Y | \text{Term \#3} | Y \rangle = 0$ by the symmetry of the p orbitals. Let $\hat{I}_{x \leftrightarrow y}$ be an operator that switches x and y . Then $\hat{I}_{x \leftrightarrow y} X(x, y, z) = Y$ and $\hat{I}_{x \leftrightarrow y} Y(x, y, z) = X$ and $\hat{I}_{x \leftrightarrow y} (\partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x) = -(\partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x)$ assuming symmetric V . By symmetry of the p orbital, we expect the switch $x \leftrightarrow y$ to leave the inner product invariant so then

$$\begin{aligned} & \frac{\hbar}{8m_0^2 c^2} \{ \langle X | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | X \rangle + \langle Y | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | Y \rangle \} \\ &= \frac{\hbar}{8m_0^2 c^2} \{ \hat{I}_{x \leftrightarrow y} \langle X | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | X \rangle + \langle Y | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | Y \rangle \} \\ &= \frac{\hbar}{8m_0^2 c^2} \{ -\langle Y | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | Y \rangle + \langle Y | \partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x | Y \rangle \} = 0 \end{aligned}$$

since interchanging x and y produces a negative sign.

Consider the mixed parts of the third term

$$\frac{-i\hbar}{8m_0^2 c^2} \left\{ \langle X | \underbrace{\partial_x V \cdot \hat{p}_y}_{3a} - \partial_y V \cdot \hat{p}_x | Y \rangle - \langle Y | \underbrace{\partial_x V \cdot \hat{p}_y}_{3b} - \partial_y V \cdot \hat{p}_x | X \rangle \right\} = \frac{-i\hbar}{4m_0^2 c^2} \langle X | \underbrace{\partial_x V \cdot \hat{p}_y}_{3c} - \partial_y V \cdot \hat{p}_x | Y \rangle$$

which holds because the function X and Y are real and assuming boundary terms disappear since the functions must reproduce from one primitive cell to the next. Consider term $3b$.

$$\begin{aligned} \langle Y | \partial_x V \cdot \hat{p}_y | X \rangle &= \int Y \partial_x V \cdot \hat{p}_y X = - \int \hat{p}_y Y \cdot \partial_x V \cdot X = - \int Y \frac{\hbar}{i} \partial_x \partial_y V \cdot X \\ &= - \int X \partial_x V \cdot \hat{p}_y Y = - \int Y \frac{\hbar}{i} \partial_x \partial_y V \cdot X = - \langle X | \partial_x V \cdot \hat{p}_y | Y \rangle = - \int Y \frac{\hbar}{i} \partial_x \partial_y V \cdot X \end{aligned}$$

Therefore

$$(3a) + (3b) = 2(3c) + \int \dots \partial_x \partial_y \dots$$

The other terms behave similarly and produce an integral to cancel $\int \dots \partial_x \partial_y \dots$.

Finally combine all of the terms to find the result

$$\underline{\mathcal{H}}_{22} = \left\langle \frac{X - iY}{\sqrt{2}} \uparrow \left| \hat{\mathcal{H}} \right| \frac{X - iY}{\sqrt{2}} \uparrow \right\rangle = E_p + \frac{-i\hbar}{4m_0^2 c^2} \langle X | \underbrace{\partial_x V \cdot \hat{p}_y - \partial_y V \cdot \hat{p}_x}_{=0} | Y \rangle = E_p - \frac{\Delta}{3}$$

The remaining matrix elements are similarly handled. Refer to the problems. We only need to calculate half of the elements since the other half are determined by the Hermiticity of the Hamiltonian.

7.12.6 EIGENVALUES

We now look for the eigenvalues of the system $\underline{\mathcal{H}}\underline{A} = \mathcal{E}\underline{A}$

$$\begin{bmatrix} E_s & 0 & kP & 0 \\ 0 & E_p - \Delta/3 & \sqrt{2}\Delta/3 & 0 \\ kP & \sqrt{2}\Delta/3 & E_p & 0 \\ 0 & 0 & 0 & E_p + \Delta/3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \mathcal{E} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \quad (7.270)$$

We look for the possible eigenvalues by calculating $\det[\underline{\mathcal{H}} - \mathcal{E}\underline{1}]$ as usual. Evaluating the determinant along the bottom row produces

$$-(E_p + \Delta/3 - \mathcal{E}) \begin{vmatrix} E_s - \mathcal{E} & 0 & kP \\ 0 & E_p - \Delta/3 - \mathcal{E} & \sqrt{2}\Delta/3 \\ kP & \sqrt{2}\Delta/3 & E_p - \mathcal{E} \end{vmatrix} = 0 \quad (7.271a)$$

where the front factor comes from $\mathcal{H}_{44}$ in Equation 7.270. The left-hand factor produce the eigenvalue for the basis state #4, $\mathcal{E} = E_p + \Delta/3$, for the HH band. It is customary to set the zero of energy for the top of the band to zero, which provides $0 = \mathcal{E} = E_p + \Delta/3 \rightarrow E_p = -\Delta/3$. The value E_s is defined to be the bandgap energy $E_s = E_g$. The remaining determinant becomes

$$\begin{vmatrix} E_g - \mathcal{E} & 0 & kP \\ 0 & -\frac{2\Delta}{3} - \mathcal{E} & \frac{\sqrt{2}\Delta}{3} \\ kP & \frac{\sqrt{2}\Delta}{3} & -\frac{\Delta}{3} - \mathcal{E} \end{vmatrix} = 0 \quad (7.271b)$$

which provides three more eigenvalues through the equation

$$\mathcal{E}(\mathcal{E} - E_g)(\mathcal{E} + \Delta) - k^2 P^2 \left(\mathcal{E} + \frac{2\Delta}{3} \right) = 0 \quad (7.271c)$$

For $k=0$, the last equation produces the band-edge energies of $\mathcal{E}_{\text{LH}}=0$, $\mathcal{E}_{\text{CB}}=E_g$, $\mathcal{E}_{\text{SO}}=-\Delta$. For small but nonzero k , the energy must differ from the band edge energy by only a small amount ε that depends on k . We assume the condition $0 \sim \varepsilon \ll \Delta$, E_g . First consider the conduction band. Substitute $\mathcal{E} = E_g + \varepsilon$ into Equation 7.271c and retain only linear terms in ε to find

$$\mathcal{E}_{\text{CB}} \cong E_g + \varepsilon = E_g + \frac{(kP)^2 (E_g + \frac{2\Delta}{3})}{E_g(E_g + \Delta)} \quad (7.272a)$$

The other band energies can be found similarly

$$\mathcal{E}_{\text{LH}} = 0 + \varepsilon = -\frac{2(kP)^2}{3E_g} \quad (7.272b)$$

$$\mathcal{E}_{\text{SO}} = -\Delta - \frac{(kP)^2}{3(E_g + \Delta)} \quad (7.272c)$$

and we previously found

$$\mathcal{E}_{\text{HH}} = E_p + \Delta/3 = 0 \quad (7.272d)$$

where Equation 7.260b relates these to the band structure through

$$\mathcal{E}_{n\vec{k}} = E_{n\vec{k}} - \frac{\hbar^2 k^2}{2m_o} \quad (7.273)$$

The designation of “CB, LH, SO, HH” in Equation 7.272 come from the magnitude of the 0th order energy except for the “LH, HH” bands since they are degenerate at $k=0$. The “LH, HH” bands can be distinguished by their effective masses.

7.12.7 EFFECTIVE MASS

The effective mass can be found from Equation 7.272 by noting the correction terms all have k^2 . Define the correction terms by $F_n k^2 = F_n(P, \Delta, E_g) k^2$ and let $\mathcal{E}_n(0)$ denote the band edge energy. Using Equation 7.273 provides

$$E_{n\vec{k}} = \mathcal{E}_{n\vec{k}}(0) + \frac{\hbar^2 k^2}{2m_o} + F_n k^2 = \mathcal{E}_{n\vec{k}}(0) + \frac{\hbar^2 k^2}{2m_{\text{eff}}}$$

which defines the effective mass for band n . We find

$$\frac{1}{m_{\text{eff}}} = \frac{1}{m_o} + \frac{2F_n}{\hbar^2}$$

The four bands have the following effective masses

Band	Effective Mass
CB	$\frac{1}{m_{\text{eff}}} = \frac{1}{m_o} + \frac{2F_n}{\hbar^2} = \frac{1}{m_o} + \frac{2P^2(E_g + \frac{2\Delta}{3})}{\hbar^2 E_g(E_g + \Delta)}$
LH	$\frac{1}{m_{\text{eff}}} = \frac{1}{m_o} + \frac{2F_n}{\hbar^2} = \frac{1}{m_o} - \frac{4P^2}{3\hbar^2 E_g}$
SO	$\frac{1}{m_{\text{eff}}} = \frac{1}{m_o} + \frac{2F_n}{\hbar^2} = \frac{1}{m_o} - \frac{2P^2}{3\hbar^2(E_g + \Delta)}$
HH	$\frac{1}{m_{\text{eff}}} = \frac{1}{m_o}$

The HH band has the wrong mass since too few bands were included in the calculation. One can refer to the references for the Luttinger–Kane model for corrections.

7.12.8 WAVE FUNCTIONS

Having found the eigenvalues $\mathcal{E}$, we can find the expansion coefficients a that lead to the band edge functions through Equation 7.262

$$|u_{n\vec{k}}\rangle = \sum_{\beta} a_{\beta\vec{k}}^{(n)} |u_{\beta 0}\rangle \quad (7.274)$$

Working with Equation 7.270 with $E_p = -\Delta/3$ and $E_s = E_g$.

$$\begin{bmatrix} E_g - \mathcal{E} & 0 & kP & 0 \\ 0 & -\frac{2\Delta}{3} - \mathcal{E} & \frac{\sqrt{2}\Delta}{3} & 0 \\ kP & \frac{\sqrt{2}\Delta}{3} & -\frac{\Delta}{3} - \mathcal{E} & 0 \\ 0 & 0 & 0 & -\mathcal{E} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = 0 \quad (7.275)$$

There must be four distinct sets of coefficients $\{a_1, a_2, a_3, a_4\}$ to define the four bands. These four sets come from substituting the four eigenvalues in Equation 7.272. In the usual method of finding the eigencolumn vectors, three of the coefficients must be related to the fourth. However, the a_4 component does not link with any of the others. This means that we can only solve for a_1 and a_2 in terms of a_3 . The a_4 component, corresponding to the HH band is completely separate and the wave function $u_{HH,k}$ does not enter the mix. We need to normalize each wave function using $\sqrt{a_1^2 + a_2^2 + a_3^2}$.

We demonstrate the following lowest order solutions

Band	n	a_1	a_2	a_3	a_4	Band Function $u_{n\vec{k}}$
CB	1	1	0	0	0	$u_{CB,\vec{k}} = iS\downarrow\rangle$
LH	2	0	$\frac{1}{\sqrt{3}}$	$\frac{\sqrt{2}}{\sqrt{3}}$	0	$u_{LH,\vec{k}} = \frac{1}{\sqrt{6}} X - iY\uparrow\rangle + \sqrt{\frac{2}{3}} Z\downarrow\rangle$
SO	3	0	$\frac{\sqrt{2}}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	0	$u_{SO,\vec{k}} = \frac{1}{\sqrt{3}} X - iY\uparrow\rangle + \sqrt{\frac{1}{3}} Z\downarrow\rangle$
HH	4	0	0	0	1	$u_{HH,\vec{k}} = \frac{-1}{\sqrt{2}} X + iY\uparrow\rangle$

Working with the upper left 3×3 block in Equation 7.275 and solving for a_1 and a_2 in terms of a_3 , we find

$$a_1 = -\frac{kP}{E_g - \mathcal{E}} a_3 \quad a_2 = \frac{\frac{\Delta}{3} + \mathcal{E} + \frac{(kP)^2}{E_g - \mathcal{E}}}{\sqrt{2}\frac{\Delta}{3}} a_3 \quad (7.276)$$

First, consider the CB. We require a_1 to be nonzero since the CB must be predominantly S-like. Equation 7.276 require

$$a_3 = -\frac{E_g - \mathcal{E}}{kP} a_1 = \frac{kP(E_g + \frac{2\Delta}{3})}{E_g(E_g + \Delta)} a_1 \rightarrow 0 \quad \text{for } k \text{ near } 0.$$

Therefore we have $a_2 \sim 0$ for the conduction band. This gives the first line in the table for the bands.

Next consider the light-hole (LH) band. Substituting Equation 7.272b into Equation 7.276 provides for $k \sim 0$

$$a_1 = -\frac{kp}{E_g + \frac{2k^2p^2}{3E_g}}a_3 \sim 0$$

$$a_2 = \frac{1}{\sqrt{2}\frac{\Delta}{3}} \left[\frac{\Delta}{3} - \frac{2(kP)^2}{3E_g} + \frac{(kP)^2}{E_g + \frac{2(kP)^2}{3E_g}} \right] a_3 \sim \frac{a_3}{\sqrt{2}}$$

The normalization factor becomes

$$\sqrt{a_1^2 + a_2^2 + a_3^2} = \sqrt{\frac{3}{2}}a_3$$

and therefore the LH periodic function becomes

$$u_{\text{LH},\vec{k}} = \frac{a_1u_{1\vec{0}} + a_2u_{2\vec{0}} + a_3u_{3\vec{0}}}{\sqrt{a_1^2 + a_2^2 + a_3^2}} = \frac{1}{\sqrt{3}}u_{2\vec{0}} + \sqrt{\frac{2}{3}}u_{3\vec{0}} = \frac{1}{\sqrt{6}}|X - iY\rangle + \sqrt{\frac{2}{3}}|Z\rangle$$

which agrees with the table.

The other wave functions are similarly found and can be worked as exercises.

7.13 INTRODUCTION TO DENSITY OF STATES

Many physical phenomena depend on the number of states within an energy range (energy density of states [EDOS]). When a semiconductor absorbs light, for example, electrons can be promoted from occupied valence states to empty conduction states. The energy of the photons must match the energy difference between the occupied and the empty states. Without the empty states, the transitions cannot occur. More occupied valence states and more unoccupied conduction states mean the possibility of greater transition rates and therefore higher levels of absorption. The same reasoning applies to thermal transitions.

The discussion distinguishes between localized and extended states only in their role in semiconductor processes. The localized states provide a convenient starting picture for developing the EDOS. Bands in a semiconductor can be viewed as extended states. This section discusses density of states for electrons in bulk and heterostructure with special focus on reduced dimensional structures. For bulk semiconductors, we indicate how satellite valleys in the bands affect the density of states. The discussion includes the reduced density of states necessary for light emitters and absorbers.

7.13.1 INTRODUCTION TO LOCALIZED AND EXTENDED STATES

Often “localized states” refer to traps and recombination centers. Electrons (or holes) moving in a semiconductor collide with the traps and become immobilized. The localized electron or hole has a wave function with finite size and remains in a given small region. For example, Figure 7.56 shows an electron caught in a trap (the electron is shown with a Gaussian-like wave function). The trap might be thought of as finitely deep square well potential at least for the lowest energy states. The localized states occur in the bulk or at the surface of a semiconductor. The surface states trap either electrons or holes or act as surface recombination centers.

The position of the state within the bandgap determines whether it behaves as a trap or as a recombination center. Shallowly trapped electrons require little energy to become free. A semiconductor at room temperature supplies sufficient numbers of phonons to release the electron. States near the

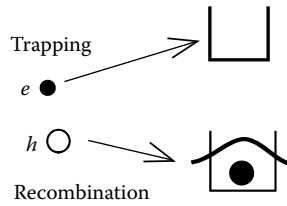


FIGURE 7.56 Trapping states with localized wave functions. Top: an electron enters a trap (shown as a quantum well). Bottom: a recombination event can occur when a hole encounters the trapped electron (other types of recombination events are possible).

center tend to be recombination centers since few phonons have enough energy to release the electron. Eventually a free hole encounters the trapped electron and recombines with it. Therefore the depth of the trap controls the rate of release and determines its function for recombination and optical processes.

The localized surface and bulk states affect the efficiency of electronic and optoelectronic components. As just mentioned, the localized states can function as recombination centers that lower the efficiency of the device. For example, consider a light emitting diode operating under forward bias. The recombination centers provide recombination current that does not contribute to the optical emission. Therefore, the efficiency of the bias current for producing light must be reduced as compared with the case for a high-quality semiconductor without recombination or trapping centers.

The extended states refer to plane waves with infinite extent and to electrons (or holes) with unrestricted motion within the semiconductor. In particular, they describe electrons and holes within the valence or conduction band. The Bloch wave function has the plane wave as the envelope function. A finite system can support only certain plane waves which give rise to the quantized energy and discrete states. The boundary conditions produce the allowed wave vectors.

7.13.2 DEFINITION OF DENSITY OF STATES

This section discusses the counting procedure for the electronic density of states (EDOS) starting with the localized states for simplicity. This section repeats earlier discussion for convenience and continuity.

The EDOS function measures the number of energy states in each unit interval of energy and in each unit volume of the crystal

$$g(E) = \frac{\text{\#States}}{\text{Energy} * \text{XalVol}} \quad (7.277)$$

We need to explore the reasons for dividing by the energy and the crystal volume.

First, consider the reason for the “per unit energy.” Suppose we have a system with the energy levels shown at the left of Figure 7.57. Assume for now the number of states occurs in a unit volume of material (say 1 cm³). Maybe the system consists of a few quantum wells with slightly different widths distributed throughout the material. The figure shows the energy levels from all of the wells in the unit volume. The figure shows four energy states in the energy interval between 3 and 4 eV. The density of states at $E = 3.5$ must be

$$g(3.5) = \frac{\text{\#States}}{\text{Energy} \times \text{Vol}} = \frac{4}{1 \text{ eV} \times 1 \text{ cm}^3} = 4$$

Similarly, between 4 and 5 eV, we find two states and the density of states function has the value $g(4.5) = 2$ and so on. Essentially, just add up the number of states at each energy. The graph shows the

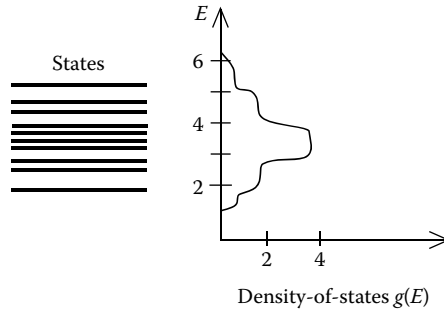


FIGURE 7.57 The density of states for the discrete levels shown on the left-hand side. The plot assumes the system has unit volume (1 cm^3) and the levels have energy measured in eV.

number of states versus energy; for illustration, the graph has been flipped over on its side. Generally we use finer energy scales and the material has larger numbers of states (10^{17}) so that the graph generally appears much smoother than the one in Figure 7.57 since the energy levels essentially form a continuum. The “per unit energy” characterizes the type of state and the type of material.

The definition of density of states uses “per unit crystal volume” in order to remove geometrical considerations from the measure of the type of state. Obviously, if each unit volume has N_v traps given by

$$N_v = \int_0^{\infty} dE g(E) = \int d(\text{energy}) \frac{\# \text{States}}{\text{Energy} * \text{Vol}} = \frac{\# \text{States}}{\text{Vol}} \quad (7.278)$$

then the volume V must have $N = N_v V$ traps. Changing the volume changes the total number. To obtain a measure of the “type of state,” we need to remove the trivial dependence on crystal volume.

What are the states? The states can be those in an atom. The states can also be traps that an electron momentarily occupies until being released back into the conduction band. The states might be recombination centers that electrons enter where they recombine with holes. Traps and recombination centers can be produced by defects in the crystal. Surface states occur on the surface of semiconductors as an inevitable consequence of the interrupted crystal structure. The density of defects can be low within the interior of the semiconductor and high near the surface; as a result, the density of states can depend on position.

Let us consider several examples. First, suppose a crystal has two discrete states (i.e., single states) in each unit volume of crystal. Figure 7.58 shows the two states on the left side of the graph. The density-of-state function consists of two Dirac delta functions of the form

$$g(E) = \delta(E - E_1) + \delta(E - E_2)$$

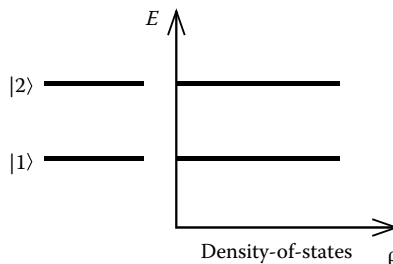


FIGURE 7.58 The density of states for two discrete states shown on the left side.

Integrating over energy gives the number of states in each unit volume

$$N_v = \int_0^\infty dE g(E) = \int_0^\infty dE [\delta(E - E_1) + \delta(E - E_2)] = 2$$

If the crystal has the size $1 \times 4 \text{ cm}^3$ then the total number of states in the entire crystal must given by

$$N = \int_0^4 dV N_v = 8$$

as illustrated in Figure 7.59. Although this example shows a uniform distribution of states within the volume V , the number of states per unit volume N_v can depend on the position within the crystal. For example, the growth conditions of the crystal can vary or perhaps the surface becomes damaged after growth.

As a second example, consider localized states near the conduction band of a semiconductor as might occur for amorphous silicon. Figure 7.60 shows a sequence of graphs. The first graph shows the distribution of states versus the position x within the semiconductor. Notice that the states come closer together (in energy) near the conduction band edge. As a note, amorphous materials have mobility edges rather than band edges. The second graph shows the density of states function versus energy where a sharp Gaussian spike represents the number of states at each energy. At 7 eV, the material has six states (traps) per unit length in the semiconductor as shown in the first graph. The second graph shows a spike at 7 eV. Actual amorphous silicon has very large numbers of traps near the upper mobility edge and they form a continuum as represented in the third graph. This example shows how the density of states depends on position and how closely space discrete levels form a continuum.

As a final example for localized states, let us consider the role of localized states for nanoscale devices. Suppose a small cube of length L represents a small electronic device. Suppose electrons and holes are created in the bulk and on the surface either by electrical or optical pumping. Suppose the device should function when carriers recombine in the bulk of the material (for example, the device might be a small LED). However, some of the carriers will recombine at surface states,

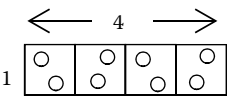


FIGURE 7.59 Each unit volume has two states and the full volume has eight.

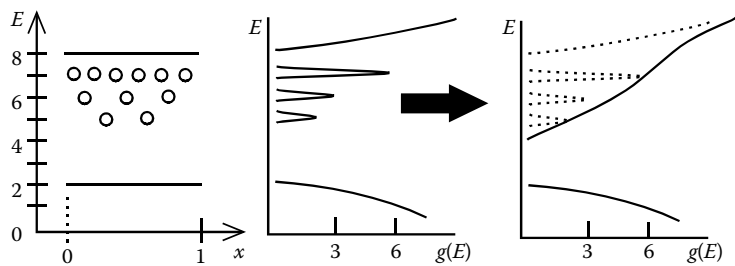


FIGURE 7.60 Transition from discrete localized states to the continuum.

which does not contribute to the device operation. We can reasonably assume the bulk and surface recombination rates depend on the total number of states at the surface and in the bulk. The bulk surface recombination rates must be

$$R_{\text{bulk}} = C_v N_v V = C_v N_v L^3 \quad R_{\text{surf}} = C_A N_A A = C_A N_A L^2 \quad (7.279)$$

where

C_v, C_A are constants

N_v, N_A represent the total number of states per volume and area, respectively

The ratio of surface to bulk recombination rates must be

$$\frac{R_{\text{surf}}}{R_{\text{bulk}}} = \frac{C_A N_A L^2}{C_v N_v L^3} \sim \frac{1}{L} \quad (7.280)$$

We therefore see that surface recombination can dominate the bulk recombination at sufficiently small sizes. For a device intended to operate using traditional bulk processes, the surface states pose significant problems and render the device nonoperative. For recombination involving phonons, the surface becomes the prime heating agent.

7.13.3 RELATION BETWEEN DENSITY OF EXTENDED STATES AND BOUNDARY CONDITIONS

So far we have discussed the density of states for the localized states. We can add up the number of extended states using similar techniques. However, the extended states correspond to plane waves characterized by a wave vector $\vec{k}$ and angular frequency ω_k . For electron and hole wave functions, the band diagrams interrelate the wave vector and angular frequency. Therefore allowed values of energy $E = \hbar\omega$ must be related to allowed values of k .

The electron can be either confined to a finite region of space or not confined at all. Confining an electron to a finite region of space places conditions on the allowed electron wavelength and hence also on the allowed wave vectors. Finite regions of space produce discrete allowed wave vectors and therefore discrete energy values. Boundary conditions mathematically model the effect of the finite regions. Either fixed-endpoint conditions or periodic boundary conditions can be applied to the wave function for the confined electron. The fixed-endpoint boundary conditions produce sine and cosine standing waves for the energy eigenfunctions. The wave vectors $\vec{k}$ have only positive components as given by the Fourier summations in Chapter 2. The periodic boundary conditions over a finite distance L usually applies to plane waves even though the wave must be restricted to length L . In this case, the wave vectors $\vec{k}$ have both positive and negative components. We apply these periodic conditions to macroscopic size L . For those electrons not confined to finite regions, we apply the periodic boundary conditions over the length L . Here the length L appears artificial in order to provide normalization for an infinitely sized wave. Nevertheless, the finite size of L leads to discrete allowed wavelengths, wave vectors, and therefore also energy. For infinitely sized regions, we can let the repetition length L increase unbounded. The allowed wavelengths, wave vectors and energy become infinitesimally close together and essentially form a continuum. The transition from the Fourier series to the Fourier transform appears very similar to this procedure for letting L increase without bound. In real crystals with finite sizes, the length of the crystal must be identical with the repetition size. In such a case, the size of the crystal sets a minimum spacing for allowed k . We find that each atom contributes one state to each band. The number of states in each band must be the same as the number of atoms N .

Once we know the allowed energies for a finite system, we can count the number of allowed states. Figure 7.61 shows the discrete states for the conduction band. We can count the number of states in the energy range ΔE to find $g(E)$. However, the figure makes it clear that

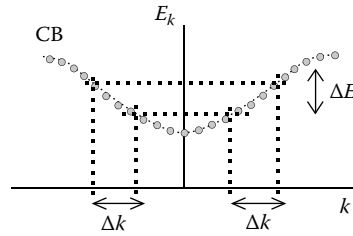


FIGURE 7.61 The density of energy states must be related to the density of k -states.

the number of states along the energy axis must be related to the number along the k -axis. In fact the total number of states in the range ΔE comes from the two regions marked Δk . For 2-D systems the Δk region corresponds to an annular region between two circles.

7.13.4 FIXED-ENDPOINT BOUNDARY CONDITIONS

The fixed-endpoint boundary conditions require a wave to be zero at the edges of the bounding region. Our main interest in the fixed-endpoint conditions is to find the spacing between allowed wave vectors so as to be able to compare with the more ubiquitous results for the periodic boundary conditions. We do not apply the results to a crystal in this section and do not consider the number of modes for the FBZ.

The fundamental modes in the range 0 to L have the form of sine and cosine waves as shown in Figure 7.62. The wavelengths can be no larger than

$$\lambda_1 = 2L$$

In fact, the wave must exactly fit into the distance L according to the relation

$$\lambda = \frac{2L}{1}, \frac{2L}{2}, \frac{2L}{3}, \dots, \frac{2L}{n}, \dots$$

Therefore, the allowed wave vectors must be

$$k_n = \frac{2\pi}{(2L/n)} = \frac{n\pi}{L} \quad n = 1, 2, 3, \dots \quad (7.281a)$$

and for the interval 0 to L , the eigenfunctions have the form

$$\phi_n(x) = \sqrt{\frac{2}{L}} \sin(k_n x) \quad (7.281b)$$

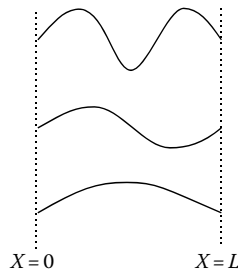


FIGURE 7.62 The endpoint boundary conditions.

The solution of the partial differential equation (Schrödinger's equation) limits n to positive numbers. Zero is not included since the boundary conditions would require ϕ_n to be zero (which indicates the particle does not exist contrary to assumption). One can see the spacing between the k values must be

$$\Delta k = k_{n+1} - k_n = \pi/L \quad (7.281c)$$

As a reminder, Fourier series expansion in sine and cosine basis set (not the sine basis in Equation 7.281b) given by

$$B = \left\{ \frac{1}{\sqrt{2L}}, \frac{1}{\sqrt{L}} \cos(k_n x), \frac{1}{\sqrt{L}} \sin(k_n x) \right\} \quad (7.282)$$

uses only positive integers and k -values. On the other hand, the integers n must be positive and negative $n = 0, \pm 1, \pm 2, \dots$ for the equivalent exponential basis set

$$B' = \left\{ \frac{e^{ik_n x}}{\sqrt{2L}} \right\} \quad (7.283)$$

Although the range of n is larger for the exponential basis set, the two sets contain the same number of basis functions.

Three-dimensional problems require 3-D wave vectors. For a cube, with sides of length L , the allowed wave vectors can be written as

$$\vec{k} = \frac{n_x \pi}{L} \tilde{x} + \frac{n_y \pi}{L} \tilde{y} + \frac{n_z \pi}{L} \tilde{z} \quad (7.284)$$

where $n_x, n_y, n_z = \pm 1, \pm 2, \dots$ for plane waves. As we will see, traveling waves most naturally use the periodic boundary conditions since then the waves do not need to be zero at the boundaries.

7.13.5 PERIODIC BOUNDARY CONDITION

Periodic boundary conditions describe macroscopically sized real crystals. The electron wave function must repeat itself every distance L , which usually matches the physical size of the crystal. For infinitely sized media, such as free space, the length L can be increased without bound. We are primarily interested in finite physical crystals. In this case, the waves can be imagined to have infinite extent by imagining copies of the physical crystal next to each other as shown in Figure 7.63.

Two allowed modes with the longest wavelengths appear in Figure 7.64. The allowed wavelengths must be given by

$$\lambda_n = \frac{L}{n}$$

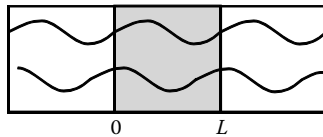


FIGURE 7.63 Repeating the physical crystal every distance L .

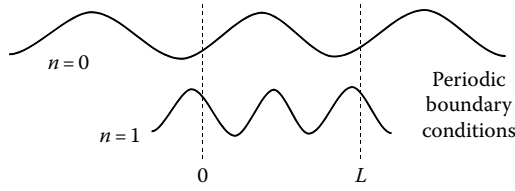


FIGURE 7.64 The first two allowed modes that satisfy the periodic boundary conditions.

and the allowed 1-D wave vectors must be

$$k_n = \frac{2\pi}{\lambda_n} = \frac{2\pi n}{L} \quad n = 0, \pm 1, \dots \quad (7.285a)$$

These are traveling waves so the basis set can be found by solving the Schrödinger wave equation with periodic boundary conditions to be (for 1-D)

$$\phi_k = e^{ikx} / \sqrt{L} \quad (7.285b)$$

Notice that the values of the wave vector can be zero, positive, or negative because of the periodic boundary conditions.

Now one has an interest in the number of k -values in each Brillouin zone. To find the number, one only needs to find how many k -states can be found within the k -range $(-\pi/a, \pi/a]$ where a represents the atomic spacing. Recall one expects strong Bragg reflections at wavelength $\lambda = 2a$ for which the group velocity will be zero. This value of wavelength corresponds to the edge of the FBZ at $k_{\text{FBZ}} = G/2 = \pi/a$ where G represents the smallest reciprocal lattice vector. To find the number of k -values in the FBZ, one merely divides the FBZ width,

$$G = 2\pi/a \quad (7.286)$$

by the minimum spacing of the k -values

$$\Delta k = k_{n+1} - k_n = 2\pi/L \quad (7.287)$$

The number of k -values is

$$N = \frac{G}{\Delta k} = \frac{L}{a} \quad (7.288)$$

Now an important point, the number of states in any single Brillouin zone equals the number of atoms in the length L . One can easily see this from the last equation since two atoms are spaced by a and so $N = L/a$ must be the number of atoms. The allowed k -states in the FBZ (1-D) have the values

$$k = \frac{\pi}{a}, \frac{\pi}{a} - \frac{2\pi}{L}, \dots, \frac{2\pi}{L}, 0, -\frac{2\pi}{L}, \dots, -\frac{\pi}{a} + \frac{2\pi}{L} \quad (7.289a)$$

Notice that $-\pi/a$ is not included since if π/a belongs to the FBZ then $-\pi/a$ can be omitted for periodicity in k . People often write this last sequence as

$$k_n = 0, \pm \frac{2\pi}{L}, \dots, \pm n \frac{2\pi}{L}, \dots \quad (7.289b)$$

without regard to the maximum value for n . However, if n is interpreted as the number of atoms, the sequence can be written as

$$k_n = \frac{\pi}{a} - (n-1) \frac{2\pi}{L} \quad n = 1, 2, \dots, N \quad (7.289c)$$

where N is the number of atoms in length L .

In free space, the wavelength can be as small as desired. However, as previously discussed with the Kronig–Penney model, for example, electron transport through a crystal has an associated E – k dispersion curve that repeats across Brillouin zones. To confine attention to the FBZ, the electron wavelength should be no smaller than $\lambda = 2a$. As soon as one imposes this condition, the number of states in the FBZ becomes fixed at N . However, note that while “ $2a$ ” appears as a lower limit to the propagation of phonons, it does not have exactly the same role for electrons. For example, high energy electrons (and ions) can be injected into a material and they can propagate into the bulk. However, collisions with atoms will limit the distance.

The periodic boundary conditions similarly apply to 3-D cubic systems to give an allowed set of wave vectors

$$\vec{k} = \frac{2\pi n_x}{L_x} \hat{x} + \frac{2\pi n_y}{L_y} \hat{y} + \frac{2\pi n_z}{L_z} \hat{z} \quad n_x, n_y, n_z = 0, \pm 1, \pm 2, \dots \quad (7.290)$$

where each component k_x, k_y, k_z has the same range as in Equation 7.289a and c. For the case of Equation 7.290, each component uses a different length L_i which matches the normalization length for the particular axis. In principle, all three terms in Equation 7.290 can have the same denominator and this would not change the method of calculating the density of states. If there are N_i atoms along axis $\#i$ with atomic spacing a_i , then $L_i = N_i a_i$ and the total number of atoms will be $N_x N_y N_z$. Regardless of the approach, one can see the 3-D case is a simple extension of that for 1-D.

Now for an important point regarding the state-counting procedure. Compare the 3-D case for fixed-endpoint boundary conditions with that for the “periodic boundary” conditions.

Type of BCs	Spacing of k -Values	Range of n
Fixed-endpoint	$\Delta k_x = \Delta k_y = \Delta k_z = \frac{\pi}{L}$	Positive
Periodic	$\Delta k_x = \Delta k_y = \Delta k_z = \frac{2\pi}{L}$	Positive or negative

One can see that the spacing between the k -values for the periodic boundary conditions is double that for the fixed endpoint ones. This means there are fewer points in each unit length of k -space for the periodic ones than for the fixed-endpoint conditions. However, the “higher density” points for the fixed-endpoint conditions are confined the region of k -space (3-D in this case) to where k_x, k_y, k_z are all positive, whereas for the periodic conditions, the k_x, k_y, k_z can be positive or negative. So when adding up the points in a given volume of k -space centered on the k -origin, one will always find the same number of states enclosed by the volume. This fact allows one to calculate the number of states using either the fixed or periodic boundary conditions.

7.13.6 DENSITY OF k -STATES

The density of k -states measures the number of possible modes in a given region of k -space. Figure 7.65 shows a 2-D region of k -space for the vectors

$$\vec{k} = \frac{2\pi m}{L} \hat{x} + \frac{2\pi n}{L} \hat{y} \quad m, n = 0, \pm 1, \dots$$

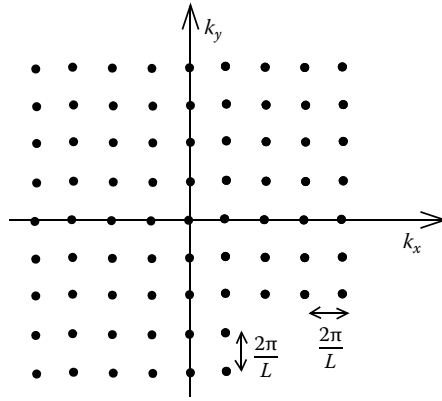


FIGURE 7.65 The allowed values of $\vec{k}$ as determined by periodic boundary conditions.

that assumes periodic boundary conditions over the length L , which is the same for both the x - and y -direction. Consider just the horizontal direction for a moment. The distance between adjacent points can be calculated as

$$\frac{2\pi(m+1)}{L} - \frac{2\pi m}{L} = \frac{2\pi}{L}$$

Therefore, each elemental area of k -space

$$\frac{2\pi}{L} \cdot \frac{2\pi}{L} = \left(\frac{2\pi}{L}\right)^2$$

corresponds to precisely one mode. The number of modes per unit area of $\vec{k}$ -space must then be given by

$$g_{\vec{k}}^{(2-D)} = \frac{1}{(2\pi/L)^2} = \frac{L^2}{4\pi^2} = \frac{A_{\text{xal}}}{4\pi^2} \quad (7.291)$$

where A_{xal} represents the area of the crystal. Note the use of the vector k as opposed to the scalar k as a subscript on g .

The same type of calculation provides the density of states for a 3-D crystal. In this case, we find one mode in each elemental volume of k -space

$$g_{\vec{k}}^{(3-D)} = \frac{1}{(2\pi/L)^3} = \frac{L^3}{8\pi^3} = \frac{V_{\text{xal}}}{8\pi^3} \quad (7.292)$$

where V_{xal} is the total volume of the crystal (in direct space). Many times the “density of k -states” has units of “#modes per unit crystal volume per unit k -space volume” thereby requiring us to divide the last equation by V_{xal} . The density of k -states becomes

$$g_{\vec{k}}^{(3-D)} = \frac{1}{8\pi^3} \quad (7.293)$$

We can likewise surmise the density of states for the 1-D crystal

$$g_k^{(1-D)} = \frac{1}{(2\pi/L)} = \frac{L}{2\pi} \quad (7.294)$$

The previous equations show that the density of states for n -dimensions can be written as

$$g_k^{(n-D)} = \frac{1}{(2\pi/L)^n} = \left(\frac{L}{2\pi}\right)^n \quad (7.295)$$

7.13.7 ELECTRON DENSITY OF ENERGY STATES FOR TWO-DIMENSIONAL CRYSTAL

In this section and the next section, we discuss the density of energy states for 2-D and 3-D arrays of atoms. We need to clearly distinguish these cases from those encountered with reduce dimensional systems such as quantum wells, wires, and dots. These latter systems still have 3-D arrangements of atoms. However, the 3-D pattern of atoms (heterostructure) produces potentials that tend to confine electrons to wells. In this topic, we discuss 2-D and 3-D arrays of atoms without regard to confining the electron to smaller wells. For simplicity, we apply the procedure to portions of the band having a parabolic shape. The density of states for the entire band requires the full dispersion curve $E = E(k)$ and not just the portion at the top or bottom of the band.

For simplicity of drawing figures, first consider the 2-D case for the electronic density of energy states. We need the energy versus wave vector k . Keep in mind that the k -vector refers to the envelope of the Bloch wave function and therefore the calculations will use the effective mass. We can write the energy dispersion relation for the electron near the bottom of the conduction band as

$$E = \frac{\hbar^2 k^2}{2m_e} \quad (7.296)$$

where we have shifted the energy scale for convenience to set the bottom of the conduction band at $E_c = 0$. This last equation does not represent the full band $E = E(\vec{k})$ but only a portion of it. The previous section calculated the number of k -states per unit wave number without restriction to the shape of the dispersion curve.

Now we determine the number of energy states in each unit interval of energy E . The number of energy states must be related to the number of allowed k -states. Equation 7.296 relates the magnitude of the wave vector to the energy of the wave. In order to know how many states fall within each unit interval of energy, one needs to know how many k -states fall within each unit length of $k = |\vec{k}|$. That is, one first must find the number of states per unit k -length. Figure 7.66 shows the total number of states within the k -space area of a circle of radius κ is given by

$$N_T = \text{Total number} = \sum \frac{\text{Number}}{k\text{-area}} \Delta(k\text{-area}) \quad (7.297)$$

which can be rewritten as

$$N_T = \int g_k^{(2-D)} (dk |k| d\varphi) \quad (7.298)$$

Substituting for the density of k -states and using a dummy variable κ provides

$$N_T = \int_0^k \kappa d\kappa \int_0^{2\pi} d\varphi g_k^{(2-D)} = \int_0^k \kappa d\kappa \int_0^{2\pi} d\varphi \frac{A_{\text{xal}}}{4\pi^2} = \int_0^k \kappa d\kappa \frac{A_{\text{xal}}}{4\pi^2} 2\pi \quad (7.299)$$

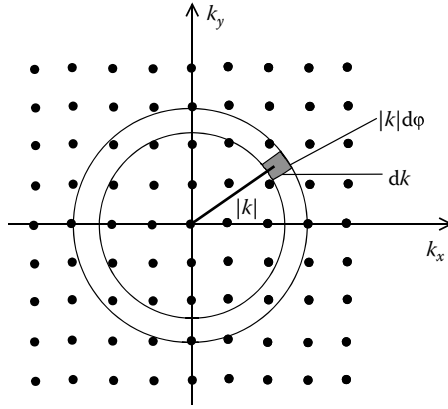


FIGURE 7.66 The number of modes in length dk (over the angular range of 2π) depends on the radius k .

Integrating over k provides

$$N_T = \frac{A_{\text{xal}}}{4\pi^2} \pi k^2 = g_k^{(2-D)} A_{k-\text{sp}} \quad (7.300)$$

where $A_{k-\text{sp}} = \pi k^2$ gives the area of the circle. We could have written Equation 7.300 right from the start since $g_k^{(2-D)}$ is independent of k . Although not needed at present, the density of states per unit “magnitude k ” can be found if desired from the last equation by differentiating

$$g_k^{(2-D)} = \frac{\partial N_T}{\partial k} = \frac{A_{\text{xal}} k}{2\pi}$$

Notice that this last equation differs from Equation 7.291 because this one gives the number of states per unit k -length whereas Equation 7.291 gives the number of states per unit k -area.

We can find the density of energy states by solving for the magnitude of the wave vector in the dispersion relation $E = \hbar^2 k^2 / (2m_e)$ and then substituting into Equation 7.300.

$$N_T = g_k^{(2-D)} A_{k-\text{sp}} = \frac{A_{\text{xal}}}{4\pi^2} \pi k^2 = \frac{A_{\text{xal}}}{2\pi} \frac{m_e}{\hbar^2} E$$

Therefore, the number of states per unit energy must be given by

$$g_E^{(2-D)} = \frac{\partial N_T}{\partial E} = \frac{A_{\text{xal}}}{2\pi} \frac{m_e}{\hbar^2} \quad (7.301a)$$

where A_{xal} represents the area of the 2-D crystal. Usually, the physical size of the crystal is removed to write the density of states as a number per energy per area

$$g_E^{(2-D)} = \frac{1}{2\pi} \frac{m_e}{\hbar^2} \quad (7.301b)$$

Notice that the 2-D density of energy states does not depend on the energy. In the 3-D case, the volume will be divided out rather than the area.

7.13.8 ELECTRON DENSITY OF ENERGY STATES FOR THREE-DIMENSIONAL CRYSTAL

The 3-D case proceeds in a similar fashion to the 2-D case. We know that the density of states in k -space is

$$g_k^{(3-D)} = \frac{V_{\text{xal}}}{8\pi^3}$$

The total number of states to a radius of κ is given by

$$N_T = \int_0^\kappa dk \int_0^{2\pi} d\phi \int_0^\pi d\theta k \sin \theta g_k^{(3-D)}$$

where the integral is given in spherical coordinates with a differential volume element of

$$(dk)(k d\phi)(d\theta k \sin \theta)$$

The two angular integrals can be evaluated since the density of states does not depend on the angles. Using a dummy variable κ , we find

$$N_T = 4\pi \int_0^\kappa d\kappa \kappa^2 g_k^{(3-D)} = 4\pi \int_0^\kappa d\kappa \kappa^2 \frac{V_{\text{xal}}}{(2\pi)^3} = \frac{V_{\text{xal}}}{(2\pi)^3} 4\pi \int_0^\kappa d\kappa \kappa^2 \quad (7.302)$$

This last equation gives the total number of states in a k -space sphere of radius k

$$N_T = g_k^{(3-D)} V_{k\text{-sph}} = \frac{V_{\text{xal}}}{(2\pi)^3} \frac{4\pi}{3} k^3 \quad (7.303)$$

Although not needed at present, the density of states in magnitude k -space can be written if desired by differentiating either Equation 7.302 or 7.303 to find

$$g_k^{(3-D)} = \frac{\partial N_T}{\partial k} = \frac{V_{\text{xal}} k^2}{2\pi^2} \quad (7.304)$$

The density of states for E -space comes from differentiating Equation 7.303 and using the dispersion relation $E = \hbar^2 k^2 / (2m_e)$

$$g_E^{(3-D)} = \frac{dN_T}{dE} = \frac{dk}{dE} \frac{dN_T}{dk} = \left(\frac{dE}{dk} \right)^{-1} \frac{d}{dk} \left(\frac{V_{\text{xal}}}{(2\pi)^3} \frac{4\pi}{3} k^3 \right) = \frac{m_e}{\hbar^2 k} \frac{4\pi V_{\text{xal}} k^2}{(2\pi)^3} = \frac{m_e V_{\text{xal}}}{2\pi^2 \hbar^2} k \quad (7.305)$$

The density of energy states must be written in terms of energy. The dispersion relation then provides

$$g_E^{(3-D)} = \frac{m_e V_{\text{xal}}}{2\pi^2 \hbar^2} \sqrt{\frac{2m_e E}{\hbar^2}} = \frac{m_e^{3/2} V_{\text{xal}}}{\sqrt{2} \pi^2 \hbar^3} \sqrt{E} \quad (7.306a)$$

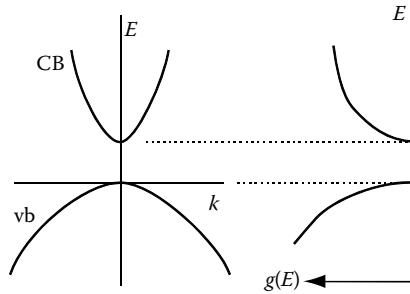


FIGURE 7.67 The conduction and valence band both have a density of states function.

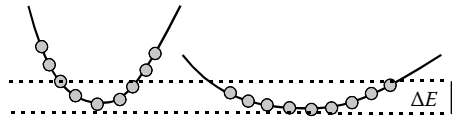


FIGURE 7.68 Different curvatures place different numbers of states in a fixed energy interval.

Usually we divide out the crystal volume as appropriate for the definition of density of energy states to get

$$g_E^{(3-D)} = \frac{m_e^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E} \quad (7.306b)$$

As an important note, the electron can have either spin up or spin down. For the present case, the spin degeneracy can be included in the density of states by multiplying by 2.

The 3-D density of energy states can be plotted next to the band diagram as illustrated in Figure 7.67. Both the conduction band and heavy-hole valence band produces a density of states. The two bands have different density of states although they both increase as $\sqrt{E}$. Notice the conduction band has been shifted back to its proper location and the density of states for the conduction band actually increases as $\sqrt{E - E_c}$ where E_c represents the bottom of the conduction band. The effective mass controls the shape of the density of states.

We can see the reason that the effective mass enters into the formula 7.306 for the density of states from Figure 7.68. The two depicted bands have different curvatures. The boundary conditions produce equally spaced states along the horizontal k -axis. Let ΔE represent a small fixed energy interval. The curvature of the bands produces two different numbers of states within the energy interval. The band with the larger curvature and therefore smaller effective mass has fewer states within the energy interval. The flatter band with the larger effective mass has more states within the interval.

7.13.9 GENERAL RELATION BETWEEN k AND E MODE DENSITY

The previous section shows how to find the k -space and E -space density of states. More generally, we can trace through the development of the previous two sections to find a general formula relating the density of states for the magnitude of $\vec{k}$ (c.f., Equation 7.304) and the density of energy states. Integrating over $|\vec{k}|$ up to some value κ must give the same number of states as integrating the energy E up to some value Ω . For example in 2-D, the radius of the circle in Figure 7.66 can be written in terms of either $k = |\vec{k}|$ or $E = \hbar^2 k^2 / (2m_e)$. Therefore, the dispersion relation relates

the limits of the integral to give the same number of states within the circle $\omega = \hbar^2 \kappa^2 / (2m_e)$. The total number of states can be written in two ways

$$\int_0^{\kappa} dk g_k = N_T = \int_0^{\Omega} dE g_E \quad (7.307)$$

Similar considerations can be applied to a variety of density of states including those for phonons and EM waves traveling in free space. Therefore, we expect

$$g_k dk = g_E dE \quad (7.308)$$

since κ , Ω are assumed to describe the same “region of mode space” as discussed below.

To see that relation 7.308 holds, consider the electron dispersion relation near the bottom of the conduction band $E = \frac{\hbar^2 k^2}{2m_e}$. Equation 7.307 becomes

$$\int_0^E dE' g_E(E') = N_T = \int_0^k dk' g_k(k')$$

Differentiate both sides with respect to E .

$$\begin{aligned} \frac{d}{dE} \int_0^E dE' g_E(E') &= \frac{d}{dE} \int_0^k dk' g_k(k') \\ g_E(E) &= \frac{dk}{dE} \frac{d}{dk} \int_0^k dk' g_k(k') = \frac{dk}{dE} g_k(k) \end{aligned}$$

Therefore,

$$g_E(E) dE = g_k(k) dk$$

7.13.10 TENSOR EFFECTIVE MASS AND DENSITY OF STATES

So far we have assumed symmetric bands in k_x , k_y , k_z . Now we repeat the derivation using the tensor effective mass

$$(m^{-1})_{ij} = \frac{1}{\hbar^2} \frac{\partial}{\partial k_i} \frac{\partial}{\partial k_j} E \quad (7.309)$$

We must use ellipsoid-shaped constant energy surfaces unlike the spherical constant energy surfaces for the symmetrical bands. We can see this as follows. The energy as a function of the components of the wave vector can be written as (see Section 7.72)

$$E = \frac{\hbar^2}{2} \sum_{ij} (m^{-1})_{ij} k_i k_j \quad (7.310a)$$

For a diagonal mass matrix

$$\underline{m} = \begin{pmatrix} m_x & 0 & 0 \\ 0 & m_y & 0 \\ 0 & 0 & m_z \end{pmatrix} \quad (7.310b)$$

we find the energy relation

$$E = \frac{\hbar^2}{2} \left[\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z} \right] \quad (7.310c)$$

We put this last dispersion relation in the standard form for an ellipse

$$1 = \frac{k_x^2}{\left(\sqrt{\frac{2m_x E}{\hbar^2}}\right)^2} + \frac{k_y^2}{\left(\sqrt{\frac{2m_y E}{\hbar^2}}\right)^2} + \frac{k_z^2}{\left(\sqrt{\frac{2m_z E}{\hbar^2}}\right)^2} \quad (7.311a)$$

which can be written as

$$1 = \frac{k_x^2}{a^2} + \frac{k_y^2}{b^2} + \frac{k_z^2}{c^2} \quad \text{with} \quad a = \sqrt{\frac{2m_x E}{\hbar^2}} \quad b = \sqrt{\frac{2m_y E}{\hbar^2}} \quad c = \sqrt{\frac{2m_z E}{\hbar^2}} \quad (7.311b)$$

We now determine the density of states by finding the number of states within the constant energy surface as illustrated in Figure 7.69. As before, the density of states in $\vec{k}$ space is

$$g_{\vec{k}}^{(3-D)} = \frac{V_{\text{xal}}}{8\pi^3} \quad (7.312)$$

The volume of the ellipsoid can be written as

$$V = \frac{4\pi}{3} abc \quad (7.313)$$

The number of states within the constant energy surface must be

$$N(E) = g_{\vec{k}}^{(3-D)} V = \frac{V_{\text{xal}}}{8\pi^3} \frac{4\pi}{3} abc = \frac{V_{\text{xal}}}{6\pi^2} \left(\frac{2}{\hbar^2}\right)^{3/2} E^{3/2} \sqrt{m_x m_y m_z} \quad (7.314)$$

The density of energy states can be written as

$$g(E) = \frac{dN}{dE} = \frac{V_{\text{xal}} m_e^{3/2}}{\sqrt{2} \pi^2 \hbar^3} \sqrt{E} \quad (7.315a)$$

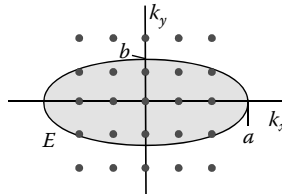


FIGURE 7.69 An ellipse in k -space as a constant energy surface.

where the effective mass must be

$$m_e = (m_x m_y m_z)^{1/3} \quad (7.315b)$$

Taking into account the two possible directions for electron spin and dividing out the crystal volume, we find

$$g(E) = \frac{dN}{dE} = \frac{\sqrt{2} m_e^{3/2}}{\pi^2 \hbar^3} \sqrt{E} \quad (7.315c)$$

7.13.11 OVERLAPPING BANDS

Gallium arsenide has overlapping heavy-hole (HH) and light-hole (LH) valence bands as shown in Figure 7.70. We will find overlapping subbands for the reduced dimensional structures such as quantum wells. Each band must have states corresponding to the allowed discrete wave vectors k . Therefore the number of states within the energy range ΔE must include states from both the HH and LH bands. The present section shows a simple calculation in preparation for the more mathematical version in a subsequent section. In fact, it is the calculation that one most often encounters and therefore should be carefully examined.

We now discuss the method for calculating the density of states for overlapping subbands. For simplicity, consider two overlapping bands with positive curvature as shown in Figure 7.70; the situation for bulk semiconductors such as in Figure 7.70 has similar development. We can easily demonstrate that the density of states must be given by

$$g_E^{(3-D)}(E) = \frac{m_1^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E - E_1} \Theta(E - E_1) + \frac{m_2^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E - E_2} \Theta(E - E_2) \quad (7.316)$$

where the step function has the definition

$$\Theta(E - E_1) = \begin{cases} 0 & E < E_1 \\ +1 & E \geq E_1 \end{cases}$$

We can intuitively see that Equation 7.316 holds (refer to Figure 7.71). At $E = 0$ in Figure 7.71, there is not any band structure and therefore there can not be any states. As E increases, we eventually encounter band 1 starting at energy E_1 where the states start. The density of states (3-D crystal) must therefore increase as $\sqrt{E - E_1}$ according to Equation 7.306a (or 7.306b). At energy E_2 , the number of states in band 2 must be included. The density of states in band 2 increases as $\sqrt{E - E_2}$ again according to Equation 7.306. To find the total number of states for energy larger than E_2 , we must add the states from bands 1 and 2. Therefore, we find Equation 7.316 (Figure 7.71).

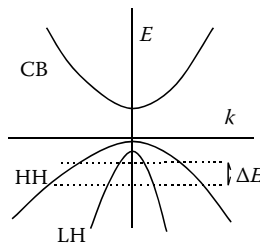


FIGURE 7.70 Light and HH valence bands.

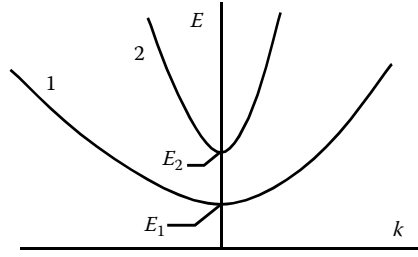


FIGURE 7.71 Two overlapping 3-D bands (inverted for convenience).

The density of states can also be demonstrated using relation 7.308, specifically $g_E(E)dE = g_k(k)dk$. Looking at the band #1, the dispersion relation can be written as

$$E = E_1 + \frac{\hbar^2 k^2}{2m_1} \quad E > E_1 \quad (7.317a)$$

where, unlike in Section 7.13.8, the bottom of the band remains shifted from $E=0$ and where m_1 represents the effective mass for band #1. The $|\vec{k}|$ density of states relation in Equation 7.304 remains unchanged

$$g_k^{(3-D)} = \frac{\partial N_T}{\partial k} = \frac{V_{\text{sal}} k^2}{2\pi^2} \quad (7.317b)$$

Therefore, Equation 7.308 provides

$$g_E^{(1)}(E) = g_k(k) \left(\frac{dE}{dk} \right)^{-1} = \frac{V_{\text{sal}} k^2}{2\pi^2} \left(\frac{\hbar^2 k}{m_1} \right)^{-1} = \frac{m_1}{\hbar^2 2\pi^2} k \quad (7.318)$$

However, solving for k in Equation 7.317a, we find

$$k = \sqrt{\frac{2m_1}{\hbar^2} (E - E_1)} \Theta(E - E_1)$$

where the step function ensures k does not become imaginary. Therefore, we find

$$g_E^{(1)}(E) = \frac{m_1^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E - E_1} \Theta(E - E_1) \quad (7.319a)$$

Similar reasoning applied to band 2 provides

$$g_E^{(2)}(E) = \frac{m_2^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E - E_2} \Theta(E - E_2) \quad (7.319b)$$

Therefore, the total density of states can be found just by adding Equation 7.319a and b together

$$g_E(E) = g_E^{(1)}(E) + g_E^{(2)}(E) = \frac{m_1^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E - E_1} \Theta(E - E_1) + \frac{m_2^{3/2}}{\sqrt{2} \pi^2 \hbar^2} \sqrt{E - E_2} \Theta(E - E_2)$$

as required.

7.13.12 DENSITY OF STATES FROM PERIODIC AND FIXED-ENDPOINT BOUNDARY CONDITIONS

This section finds the density of states using the periodic boundary conditions. The length L in Figures 7.63 and 7.64 appears to be rather arbitrary. For the fixed-endpoint boundary conditions, the length L matches the physical length of the crystal. We make the same requirement for the length L in the periodic boundary conditions as illustrated in Figure 7.63. However, the fixed-endpoint conditions might seem to give the more accurate density of states since electrons must surely be confined to the crystal and cannot therefore be a standing wave that repeats every length L . Let us examine how the choice of the type of boundary conditions affects the density of states. We will find that both types give precisely the same density of state function.

The following table compares the wavelength, wave vectors, and minimum wave vector spacing using periodic and fixed-endpoint boundary conditions for a 2-D crystal (for example).

Periodic BCs	Fixed-Endpoint BCs
$\lambda_x = L/m \quad \lambda_y = L/n$	$\lambda_x = 2L/m \quad \lambda_y = 2L/n$
$k_x = 2\pi m/L \quad k_y = 2\pi n/L$	$k_x = \pi m/L \quad k_y = \pi n/L$
$\Delta k_x = 2\pi/L \quad \Delta k_y = 2\pi/L$	$\Delta k_x = \pi/L \quad \Delta k_y = \pi/L$
Traveling waves	Standing waves
m, n can be positive and negative	m, n must be nonnegative

The spacing between allowed k -values is twice the size for the periodic boundary conditions compared with the fixed-endpoint ones. As shown in Figure 7.72, the density of k -states from the periodic boundary conditions (PBC) must be 25% that for the fixed-endpoint boundary conditions (FEBC)

$$g_{\vec{k}(\text{PBC})}^{(2\text{-D})} = \frac{g_{\vec{k}(\text{FEBC})}^{(2\text{-D})}}{4}$$

(7.320a)

Next, we see that the portion of the area of the constant energy circle covering the allowed states for the periodic boundary conditions is four times that for the fixed-endpoint point conditions.

$$A_{\text{PBC}} = 4A_{\text{FEBC}}$$

(7.320b)

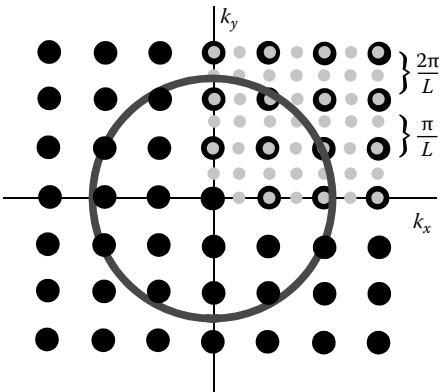


FIGURE 7.72 Full black circles represent allowed k for periodic BC while the light circles represent the fixed-endpoint BCs.

The density of energy states can then be calculated from the product of Equation 7.320a and b. We find the same result for either set of boundary conditions.

$$g_{\vec{k}(\text{PBC})}^{(2\text{-D})} A_{(\text{PBC})} = \frac{g_{\vec{k}(\text{FEBC})}^{(2\text{-D})}}{4} 4A_{(\text{FEBC})} = g_{\vec{k}(\text{FEBC})}^{(2\text{-D})} A_{(\text{FEBC})} \quad (7.320c)$$

So one finds the same $g(E)$ with either type of boundary condition.

7.13.13 CHANGING SUMMATIONS TO INTEGRALS

We often use the density-of-states (i.e., density-of-modes) to find the total number of carriers when we know the number per state (Fermi–Dirac distribution). However, the same reasoning applies to other quantities besides the number of carriers. Let us call the amount of some quantity per state as amount/state. We can write

$$\text{Total amount} = \sum \frac{\text{Amount}}{\text{State}} \cdot \frac{\#\text{States}}{k\text{-space}} \cdot \Delta(k\text{-space})$$

Let, $A(\vec{k})$ be the “amount” per state at wave vector $\vec{k}$ and let $g_{\vec{k}}$ be the $\vec{k}$ -space density-of-modes. The “total amount” can be written by

$$\text{Total amount} = \int_{k\text{-vol}} A(\vec{k}) g_{\vec{k}} d^3\vec{k}$$

The differential $d^3\vec{k}$ represents a small element of volume in $\vec{k}$ -space such as, for example, the volume element in the previous section of the form

$$d^3\vec{k} = k^2 \sin \theta dk d\phi d\theta$$

The density-of-states and density-of-modes can be used to convert summations to integrals. Suppose we start with a summation of coefficients $C_{\vec{k}}$ of the form

$$S = \sum_{\vec{k}} C_{\vec{k}}$$

The index $\vec{k}$ on the summation means to sum over allowed values of k_x, k_y, k_z ; that is, think of the 2-D plot in the previous sections and imagine that $C_{\vec{k}}$ has a different value at each point on the plot. For one dimension, a plot of “ C_k versus k ” might appear as in the Figure 7.73. Suppose the allowed values of k are close to one another. Let ΔK_i be a small interval along the k -axis; this interval is small but assume that it contains many of the k points. Let K_i be the center of each of these intervals. The figure shows that

$$S = (C_{1.00} + C_{1.01} + C_{1.02} + C_{1.03}) + (C_{1.04} + C_{1.05} + C_{1.06} + C_{1.07}) + \cdots$$

The sum can be recast into

$$S = 4C_{1.00} + 4C_{1.04} + \cdots = \sum [g(k)\Delta k] C_k \equiv \sum C(k)g(k)\Delta k = \int dk C(k)g(k)$$

where, for the figure, $\Delta k = 0.04$ and $g(k) = 4/0.04 = 100$.

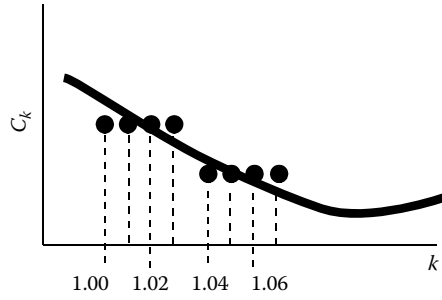


FIGURE 7.73 Example of closely spaced modes.

Now let prove the above conjecture in general—it works for any slowly varying function $f(x)$. Suppose f is defined at the points in the set $\{x_1, x_2, \dots\}$ where the points x_i are equally spaced and separated by the common distance Δx . The summation can be rewritten as

$$\sum_i f(x_i) = \sum_i \frac{1}{\Delta x_i} f(x_i) \Delta x_i$$

We recognize the quantity $1/\Delta x$ as the density of states; that is, $g = 1/\Delta x$. Recognizing the second summation as an integral for sufficiently small Δx , the summation can be written as

$$\sum_i f(x_i) \cong \int dx g(x) f(x) \quad (7.321)$$

The last expression generalizes to a 3-D case most commonly applied to the wave vectors discussed in the preceding topics.

$$\sum_{\vec{k}} f(\vec{k}) \rightarrow \int d^3k g(\vec{k}) f(\vec{k}) = \frac{V}{(2\pi)^3} \int d^3k f(\vec{k}) \quad (7.322)$$

where V represents the normalization volume coming from periodic boundary conditions. We essentially use this last integral when we find the total number of discrete states within a sphere or circle.

7.13.14 COMMENT ON PROBABILITY

The previous section discusses the use of the density of states for computing summations. This section points out the difference among the average, probability and the density of states function.

Suppose that repeated measurement of a random variable X produces a discrete set $x_1, x_2, x_3, \dots$. The average value of that set is given by

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i$$

Suppose we plot the value of X versus the measurement number as shown in Figure 7.74. Suppose, for example, that x_1, x_5 have the same value as x_1 , that x_2, x_3, x_6 have the same value as x_2 , that x_4, x_7 have the same value as x_4 , and $N=7$. The summation can be written as

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{7} (2x_1 + 3x_2 + 2x_4)$$

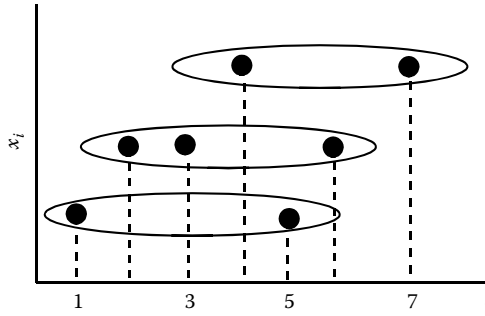


FIGURE 7.74 Regrouping points for calculations involving probability.

The probability of x_1 occurring is $P(x_1) = 2/7$. Similarly, the probability of x_2, x_4 is given by $P(x_2) = 3/7$ and $P(x_4) = 2/7$. Now the average value can be written as

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i = \sum_{x_i} x_i P(x_i)$$

where it is crucially important to note the second summation extends over the possible values rather than the index “ i ” since P accounts for the multiple values.

At this point, it should be clear that the indices are unnecessary. The average value can be written as

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i = \sum_x x P(x)$$

The point is this: the summation over the N observations can be rearranged into a summation over the observed values. The figure shows that this is a horizontal grouping and does not involve the number of states i per unit i -space. Instead, the average is more related to the number of states per unit x -space. This is more apparent for the integral version.

From calculus

$$\langle f(x) \rangle = \frac{1}{L} \int_0^L dx f(x) = \sum_i f(x_i) \frac{1}{L} \Delta x_i$$

By regrouping the possible values of $y_i = f(x_i)$ into like values, the summation can be rewritten as before

$$\langle f(x) \rangle = \frac{1}{L} \int_0^L dx f(x) = \int y_i \rho(y_i) dy_i = \int y \rho(y) dy$$

where ρ is the probability density. The advantage of the formula using the probability density is that we do not need to know the functional form of $f(x)$.

7.14 INFINITELY DEEP QUANTUM WELL IN A SEMICONDUCTOR

Leading edge research focuses on the theory, fabrication, and experiments on reduced dimensional structures. These structures can have fewer than a hundred atoms. Such small sizes induce quantum confinement effects in the systems that radically affect the band structure and most of the optoelectronic properties.

Many devices use epitaxially grown heterostructure where the composition of the material changes along the z -axis as shown, for example, in Figure 7.75. As an example for using the effective mass equation, we discuss separation of variables and the resulting Sturm–Liouville equation for the case of an electron confined along the z -direction. In this section, we approximate the finitely deep well with the infinitely deep one. For the finitely deep well, one would need to use the results in Section 5.3 for a finite well with an effective mass.

We want to model the electron and hole dynamics in crystals incorporating spatially varying potentials that confine these electrons and holes. The crystals might be 1-D, 2-D, or 3-D. The dimension of the embedded structure describes the number of unconfined directions. Bulk material does not confine the carriers and it can be considered a 3-D microstructure. The quantum well confines along one spatial dimension and therefore represents a 2-D structure. The quantum wire confines along two directions and can therefore be classified as a 1-D nanostructure. The quantum dot confines in all directions and is often given the designation as a 0-D structure.

As an example, Figure 7.75 shows a heterostructure with varying aluminum concentration along the growth axis z . The crystal atoms produce a periodic potential V_L (L for lattice) and the interfaces produce the confining potential V . The 3-D character of the structure leads to a 2-D equation for the x – y directions and a 1-D equation for the z -direction. All three directions must use a form of the Bloch wave functions. Figure 7.76 shows the Bloch wave function for the z -direction. The finitely

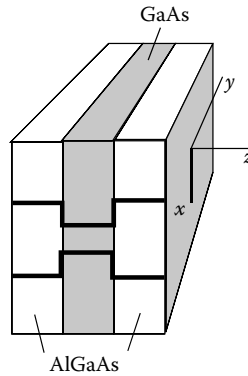


FIGURE 7.75 The band offset produces quantum wells in a heterostructure.

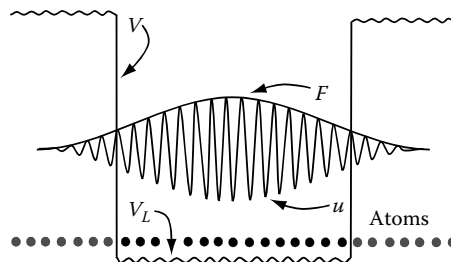


FIGURE 7.76 Cartoon representation of the wave function for a finite well. Note the waves in the lines for the barrier tops and well bottom are due to the periodic potential of the atoms.

deep well requires boundary conditions at the interfaces. Notice how the Bloch function u is periodic in the atomic spacing and the envelope F changes the amplitude of the wave function.

7.14.1 ENVELOPE FUNCTION APPROXIMATION FOR INFINITELY DEEP WELL

The Schrödinger wave equation for the heterostructure can be written as

$$-\frac{\hbar^2}{2m}\nabla^2\Psi + (V + V_L)\Psi = i\hbar\frac{\partial}{\partial t}\Psi \quad (7.323)$$

where

m denotes the free mass of the electron

V is the confining heterostructure potential

V_L is the potential with the periodicity of the lattice

We consider only the conduction band to avoid the difficulties introduced by the degenerate valence bands.

There are some differences between the bulk crystal and the heterostructure. In either case, a general wave function in the Hilbert space has the form

$$|\Psi(t)\rangle = \sum_{\vec{k}} \beta_{n\vec{k}}(t) |n, \vec{k}\rangle = \sum_{\vec{k}} \beta_{n\vec{k}}(0) |n, \vec{k}\rangle e^{-iE_{n\vec{k}}t/\hbar} \quad (7.324)$$

The basis consists of the exact energy eigenfunctions $|n, \vec{k}\rangle$. The coefficient β represents the probability of finding the electron in the extended state $|n, \vec{k}\rangle$. For the infinite crystal, the basis set has the form

$$|n, \vec{k}\rangle \sim \psi(\vec{r}) = \phi_{\vec{k}}(\vec{r}) u_{n,\vec{k}}(\vec{r}) = \frac{1}{\sqrt{\Omega}} e^{i\vec{k}\cdot\vec{r}} u_{n,\vec{k}}(\vec{r}) \quad (7.325)$$

and the normalization volume Ω comes from the periodic boundary conditions. The envelope and periodic parts of this wave function satisfy the usual orthonormality relations

$$\langle \phi_{\vec{k}} | \phi_{\vec{k}} \rangle = \delta_{\vec{k}\vec{k}} \quad \langle u_{n\vec{k}} | u_{m\vec{k}} \rangle_{uc} = \int_{uc} dV u_{n\vec{k}}^* u_{m\vec{k}} = V_{uc} \delta_{mn} \quad (7.326)$$

where uc restricts the integration over any unit cell with volume V_{uc} and we represent the conduction band by $n=2$. The envelope approximation uses the fact that $u_{n,\vec{k}}(\vec{r}) \sim u_{n,0}(\vec{r}) = u_n(\vec{r})$ so that an arbitrary vector in the Hilbert space becomes

$$\Psi(\vec{r}, t) = \sum_{\vec{k}} \beta_{\vec{k}}(t) \phi_{\vec{k}} u_{n,\vec{k}}(\vec{r}) \cong \left[\sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}}(\vec{r}) \right] u_{n,\vec{k}}(\vec{r}) = F(\vec{r}, t) u_{n,\vec{k}}(\vec{r}) \quad (7.327)$$

The envelope wave function F carries the system dynamics.

The use of a heterostructure rather than the infinite crystal alters the basis set and requires different boundary conditions from those used with free space. The form of Bloch energy basis in Equation 7.325 requires the system to be invariant with respect to translations through lattice vectors. However, the heterostructure interrupts the periodicity of the lattice thereby invalidating the assumption on invariance. We assume that the Bloch wave functions still approximately hold.

Although somewhat not physical for most materials, the infinitely deep well uses particularly simple boundary conditions that require the wave function to be zero at the internal interfaces and elsewhere outside of the well along the z -direction. However, along the x - and y -directions shown in Figure 7.75, the electron can propagate in a 2-D crystal with the translational symmetry required for the Bloch states. We will find that the basis set for the infinitely deep quantum well must have the form

$$\psi_n(\vec{r}) = \sqrt{\frac{2}{L}} \sin(k_z z) \frac{e^{i\vec{k}_\perp \cdot \vec{r}_\perp}}{\sqrt{A}} u_n(\vec{r}) \quad (7.328)$$

for the conduction band $n = 2$. The confinement along z requires us to single out the z -component so that $\vec{k} = k_x \hat{x} + k_y \hat{y} + k_z \hat{z} \equiv \vec{k}_\perp + k_z \hat{z}$ and $\vec{k}_\perp$ gives the component of the wave vector perpendicular to the confinement direction (i.e., $k_\perp$ gives the envelope wave vector for the Bloch state along the plane of the quantum well). The position vector is treated similarly. The general wave function then has the form

$$\Psi_n(\vec{r}, t) = \left\{ \sum_{\vec{k}} \beta_{\vec{k}}(t) \phi_{\vec{k}}(\vec{r}) \right\} u_n(\vec{r}) = \left\{ \sum_{\vec{k}} \beta_{\vec{k}}(t) \left[\sqrt{\frac{2}{L}} \sin(k_z z) \frac{e^{i\vec{k}_\perp \cdot \vec{r}_\perp}}{\sqrt{A}} \right] \right\} u_n(\vec{r}) \quad (7.329)$$

The envelop wave function consists of a plane wave (complex exponential) moving in the plane of the quantum well and it consists of a sinusoidal for the confinement direction along z . We assume the electron remains free of the interfaces due to the barriers.

The finitely deep well must have boundary conditions that allow the electron to penetrate into the barrier. In free space, we require the wave function and the first derivative to be continuous across the interface. Although these provide reasonable results for the quantum well, we will use a better approximation. In addition to requiring the wave function to be continuous across the boundary, we will require the product of the effective mass and the first derivative of the wave function to be continuous.

These cases all show that the periodic part of the wave function can be removed from consideration and we only need to work with the envelop basis function $\phi_{\vec{k}}$ or a summation over the basis functions. The energy basis functions satisfy the eigenvector equation

$$\hat{H} \phi_{\vec{k}} = E_{\vec{k}} \phi_{\vec{k}} \quad (7.330)$$

where $E_{\vec{k}} = E_{2, \vec{k}}$ and $\hat{H} = -\frac{\hbar^2}{2m_e} \nabla^2 \Psi + V \Psi$ and V represents the potential (notice the “ $\hbar$ ” subscript on V used in the previous section has been dropped).

7.14.2 SOLUTIONS FOR INFINITELY DEEP QUANTUM WELL IN 3-D CRYSTAL

We now find the energy states associated with the quantum well. The results must have aspects of an electron free to move in a 2-D crystal in the plane of the quantum well x - y and also of a confined electron along the confinement direction z . Once having found the states, one can determine the density-of-states (DOS) for the quantum well. The procedure will be carried out in the next few sections.

To start, separate variables in the time-independent Schrödinger wave Equation 7.330.

$$-\frac{\hbar^2}{2m_e} \nabla^2 \psi(x, y, z) + V(z) \psi(x, y, z) = E \psi(x, y, z) \quad (7.331)$$

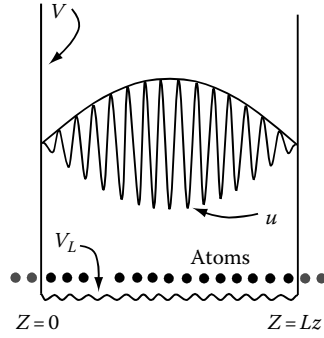


FIGURE 7.77 Lowest energy eigenfunction for the semiconductor infinitely deep well.

where E represents the total energy of the electron. The total kinetic energy comes from motion perpendicular and parallel to the interfaces in the heterostructure. Inside the quantum well, a zero potential $V(z) = 0$ for $0 \leq z \leq L$ (effective mass theory) requires the total energy to be the same as the kinetic energy. A standing wave describes the electron motion along the confinement direction z as shown in Figure 7.77. As usual, we separate the equation by substituting

$$\psi = X(x)Y(y)Z(z) \quad (7.332)$$

and then divide by ψ to find

$$\underbrace{-\frac{\hbar^2}{2m_e} \frac{1}{X} \frac{\partial^2}{\partial x^2} X(x)}_{E_x} - \underbrace{\frac{\hbar^2}{2m_e} \frac{1}{Y} \frac{\partial^2}{\partial y^2} Y(y)}_{E_y} - \underbrace{\frac{\hbar^2}{2m_e} \frac{1}{Z} \frac{\partial^2}{\partial z^2} Z(z) + V(z)}_{E_z} = E \quad (7.333)$$

The total energy consists of the sum of the energies for motion in the x -, y -, z -directions

$$E = E_x + E_y + E_z \quad (7.334)$$

We already know the eigenfunctions and eigenvalues for motion in the x - and y -directions.

$$X_{k_x} = e^{ik_x x} \quad Y_{k_y} = e^{ik_y y} \quad (7.335)$$

where

$$E_x = \frac{\hbar^2 k_x^2}{2m_e} \quad E_y = \frac{\hbar^2 k_y^2}{2m_e} \quad (7.336)$$

These last two equations represent dispersion curves for the x - and y -directions; the electron acts as a free electron so long as the effective mass replaces the free mass. Equation 7.336 assumes spherical bands but the effective masses can be replaced with m_x and m_y as necessary. The allowed values of k_x and k_y come from macroscopic boundary conditions as usual. The equation for the z -direction takes the form of

$$-\frac{\hbar^2}{2m_e} \frac{\partial^2}{\partial z^2} Z + V(z)Z = E_z Z \quad (7.337)$$

We need to find the eigenfunctions and eigenvalues for this last equation.

For the infinitely deep well, we assume that the envelope wave function must be zero outside the well as given by the fixed-endpoint boundary conditions $z = 0$ and $z = L_z$. We have solved this type of equation in Chapter 5 and found

$$Z(z) = \sqrt{\frac{2}{L_z}} \sin(k_z z) \quad \text{where } k_z = n\pi/L_z \quad n = +1, +2, \dots \quad (7.338a)$$

and the energy eigenvalues

$$E_{(z)n} = \frac{\hbar^2 k_z^2}{2m_e} = \frac{n^2 \pi^2 \hbar^2}{2m_e L_z^2} \quad (7.338b)$$

These last equations for the wave function and energy corresponding to the z -direction required the fixed-endpoint boundary conditions.

Now that we have the allowed energies and the eigenfunctions for the z -direction, the general solution to the original Schrödinger time-dependent equation can be determined. The total energy consists of the quantum well energy E_z plus the energy due to motion parallel to the interfaces.

$$E = E_z + E_{xy} = E_z + \frac{\hbar^2}{2m_e} (k_x^2 + k_y^2) \quad (7.339)$$

and the general wave function must have the form

$$\begin{aligned} \Psi &= \sum_{k_x, k_y, k_z} C_{k_x, k_y, k_z} X_{k_x} Y_{k_y} Z_{k_z} e^{-itE/\hbar} \\ &= \sum_{k_x, k_y, k_z} C_{k_x, k_y, k_z} X_{k_x} Y_{k_y} Z_{k_z} e^{-itE_z/\hbar} e^{-itE_{xy}/\hbar} \end{aligned} \quad (7.340)$$

where the energy E_{xy} for motion in the x - y -plane is

$$E_{xy} = \frac{\hbar^2}{2m_e} (k_x^2 + k_y^2) \quad (7.341)$$

The dispersion relation 7.341 applies to directions parallel to the plane of the quantum well. It appears to describe the motion of a free particle with mass m_e . The effective mass m_e must depend on the wave vector since the band must curve and produce a gap. We can make the effective mass a constant so long as we only apply Equation 7.341 to the bottom of the conduction band.

The total energy in Equation 7.339 represents a sequence of paraboloids as shown Figure 7.78. The vertex of each one increases in energy according to the energy E_z in Equation 7.338b. Electron motion in the x - y -planes of Figure 7.75 is similar to the motion of free electrons because of the parabolic dispersion relations. Each paraboloid in Figure 7.78 corresponds to the portion of the electron motion parallel to the layers. However the paraboloids must be displaced from the origin along the energy axis because of the additional discrete energy levels due to the quantum well. Normally the dispersion curves in Equation 7.339 have a 2-D appearance for convenience as shown in Figure 7.79.

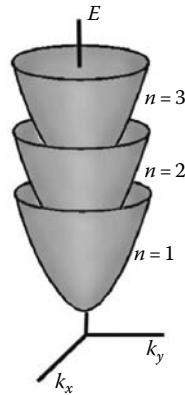


FIGURE 7.78 The energy subbands from Equation 7.339.

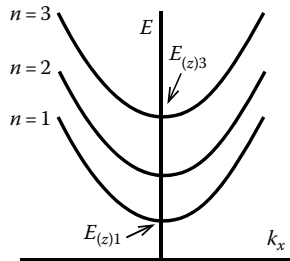


FIGURE 7.79 Subbands for the quantum well in the 3-D crystal as viewed for the single component k_x .

7.14.3 INTRODUCTION TO THE DENSITY OF STATES

The dispersion surface associated with the quantum well (for example) can be viewed as consisting of “subbands” as in Figures 7.78 and 7.79, each of which represent motion of the electron along the “free” direction parallel to x - y -plane. One subband has the same shape as the every other. However, they are displaced from each other according to the possible electron energies for motion along z . The effects of confinement include the change from a single dispersion surface into multiple subbands separated by the confinement energy and the fact that each subband represents a 2-D density of states (for a quantum well) rather than the 3-D version for bulk crystal. One should note that the density-of-states depends on position within the semiconductor. For the quantum well region, the dispersion surfaces have the form indicated in Figure 7.79 for example. However, outside the quantum well, the dispersion surface reverts to the type discussed in Section 7.13 for the 3-D case.

To find the density of states, one only needs to calculate the density of states for one subband and then include the effects of the displacement along the energy axis. For example, suppose one is interested in the density of states at energy E (c.f., Figure 7.80). If $E < E_1$ then there are not any states and the density of states must be zero. If $E_1 \leq E < E_2$, then the density of states comes only from the $n = 1$ dispersion curve. For larger energy E , one must include the density of states of the other subbands. For example, the energy range ΔE for the energy E shown in Figure 7.80 includes two subbands and so the states from these two bands contribute to the total density of states for the quantum well structure. One just imagines moving the energy E to larger values and adding states from bands as they are encountered. A complete calculation can be found in the next section but it does not add significantly to the concept.

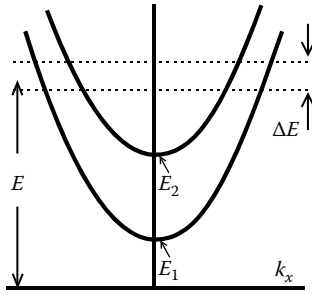


FIGURE 7.80 Density of states depends on the number of subbands at energy E . The value E_1 refers to the energy at the vertex of the lowest sub-band while E_2 refers to the vertex energy of the upper sub-band.

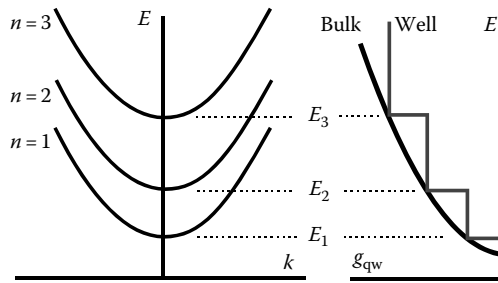


FIGURE 7.81 The density of energy states for the quantum well and its relation to the subband diagram.

One can now calculate the density of states. The electron in the quantum well can propagate freely within the plane of the quantum well. In such a case, the 2-D density of states described in Section 7.13 applies to each subband

$$g_E^{(2-D)} = \frac{1}{2\pi} \frac{m_e}{\hbar^2} \quad (7.344)$$

Then for example, $E_2 \leq E < E_3$ gives the density-of-states for the quantum well as

$$g_{QW} = \frac{1}{2\pi} \frac{m_e}{\hbar^2} \Big|_{n=1} + \frac{1}{2\pi} \frac{m_e}{\hbar^2} \Big|_{n=2} \quad (7.345a)$$

where each term must be evaluated at energy E . In the present case, we assume that m_e is independent of energy so that each term is a constant. The density of states becomes

$$g_{QW} = \frac{1}{\pi} \frac{m_e}{\hbar^2} \quad (7.345b)$$

Figure 7.81 shows how the density of states for the quantum well increase with energy in a step-like fashion.

7.15 DENSITY OF STATES FOR REDUCED DIMENSIONAL STRUCTURES

One of the most exciting areas of modern research focuses on the possibility of fabricating reduced dimensional structures. These structures incorporate potentials with dimensions on the order of tens to hundreds of atoms. Such small sizes induce quantum confinement effects in the system which radically affects the band structure and the optoelectronic properties.

In this section, we develop the density of states for these reduce dimensional structures after briefly reviewing the solution to the Schrödinger wave equation in the effective mass approximation.

7.15.1 ENVELOPE FUNCTION APPROXIMATION

We want to model 3-D crystals with potentials that reduce the Schrödinger wave equation to simpler 1-D or 2-D problems. To fix our thoughts, Figure 7.82 shows a 3-D heterostructure with varying aluminum concentration along the growth axis z . The GaAs region forms a 2-D reduced dimensional structure, namely a quantum well. The crystal atoms produce a periodic potential V_L (L for lattice) and the interfaces produce the confining potential V . The 3-D character of the structure leads to a 2-D equation for the x - y -directions and a 1-D equation for the z -direction. All three directions must use the Bloch wave functions. Figure 7.83 shows the Bloch wave function for the z -direction.

The Schrödinger wave equation for the heterostructure can be written as

$$-\frac{\hbar^2}{2m} \nabla^2 \Psi + (V + V_L) \Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (7.346)$$

where m denotes the free mass of the electron. The wave function has the form

$$|\Psi(t)\rangle = \sum_{\vec{k}} \beta_{n\vec{k}}(t) |n, \vec{k}\rangle = \sum_{\vec{k}} \beta_{n\vec{k}}(0) |n, \vec{k}\rangle e^{-iE_{n\vec{k}}t/\hbar} \quad (7.347)$$

where the eigenfunctions have the form

$$|n, \vec{k}\rangle \sim \psi(\vec{r}) = \frac{1}{\sqrt{\Omega}} e^{i\vec{k} \cdot \vec{r}} u_{n\vec{k}}(\vec{r}) \quad (7.348)$$

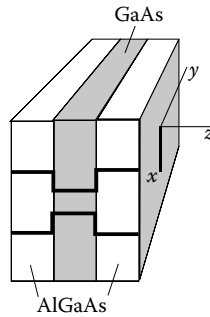


FIGURE 7.82 The band offset produces quantum wells in a heterostructure.

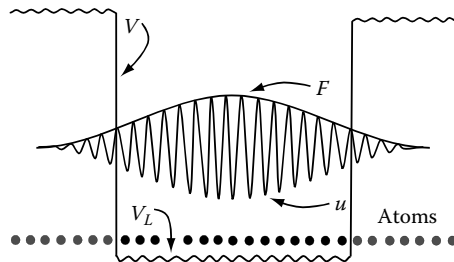


FIGURE 7.83 Cartoon representation of the wave function for the well.

where u is periodic in the lattice and Ω represents the normalization volume. We confine our attention to the conduction band. A similar expression can be used for the valence bands so long as the light and HH bands have sufficient separation in energy (nondegenerate bands). The basis functions for the Hilbert space of envelope functions

$$\phi_{\vec{k}}(\vec{r}) = \frac{1}{\sqrt{\Omega}} e^{i\vec{k} \cdot \vec{r}} \quad (7.349a)$$

satisfy the orthonormality relation

$$\langle \phi_{\vec{k}} | \phi_{\vec{k}} \rangle = \delta_{\vec{k}\vec{k}} \quad (7.349b)$$

The Bloch functions $u_{n,\vec{k}}$ are periodic on the crystal so that the values of $u_{n,\vec{k}}$ repeat from one unit cell to the next. The Bloch function $u_{n,\vec{k}}$ satisfy an inner product over the unit cell of the form.

$$\langle u_{n\vec{k}} | u_{m\vec{k}} \rangle_{uc} = \int_{uc} dV u_{n\vec{k}}^* u_{m\vec{k}} = V_{cell} \delta_{mn} \quad (7.350a)$$

Consider only the conduction band ($n=2$) and define $u_{2,\vec{k}} = u_{\vec{k}}$. So that

$$\langle u_{2\vec{k}} | u_{2\vec{k}} \rangle_{uc} \equiv \langle u_{\vec{k}} | u_{\vec{k}} \rangle_{uc} = \int_{uc} dV u_{\vec{k}}^* u_{\vec{k}} = V_{uc} \quad (7.350b)$$

where uc restricts the integration over any unit cell and represent the conduction band by $n=2$.

The general vector in the space spanned by the basis set

$$|n, \vec{k}\rangle \sim \psi_{n,\vec{k}}(\vec{r}) = \frac{1}{\sqrt{\Omega}} e^{i\vec{k} \cdot \vec{r}} u_{n,\vec{k}}(\vec{r}) \quad (7.351a)$$

has the form

$$\Psi(\vec{r}, 0) = \sum_{\vec{k}} \beta_{\vec{k}} \psi_{n,\vec{k}}(\vec{r}) = \sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}} u_{n,\vec{k}}(\vec{r}) \quad (7.351b)$$

The envelope approximation uses the fact that $u_{n,\vec{k}}(\vec{r})$ must be relatively independent of the wave vector $\vec{k}$ since it corresponds to a wavelength having the size of many unit cells whereas $u_{n,\vec{k}}(\vec{r})$ has distinct values only within the unit cell. Therefore, using $u_{n,\vec{k}}(\vec{r}) \sim u_{n,0}(\vec{r}) \equiv u_n(\vec{r})$, one can write

$$\Psi(\vec{r}, 0) = \sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}} u_{n,\vec{k}}(\vec{r}) \cong \left[\sum_{\vec{k}} \beta_{\vec{k}} \phi_{\vec{k}}(\vec{r}) \right] u_n(\vec{r}) = F(\vec{r}) u_n(\vec{r}) \quad (7.351c)$$

The envelope function $F(\vec{r})$ resides in the Hilbert space spanned by the envelope basis set $\{\phi_{\vec{k}}(\vec{r})\}$. The solution to the Schrödinger wave equation has the form of that for a modulated carrier.

For each state $\phi_{\vec{k}}(\vec{r})$, there exists a basis function $|2, \vec{k}\rangle$ and a Bloch function $u_{2,\vec{k}} = u_{\vec{k}}$. Therefore, we can count the allowed values of k to find the number of allowed states. This most importantly means we can use the effective mass approximation to find the number of states (density of states).

The effective mass equation drops the periodic potential V_L but replaces the free mass with the more complicated effective mass m_e .

$$-\frac{\hbar^2}{2m_e} \nabla^2 \Psi + V\Psi = i\hbar \frac{\partial}{\partial t} \Psi \quad (7.352a)$$

The solution has the form

$$|\Psi(t)\rangle = \sum_{\vec{k}} \beta_{2\vec{k}}(t) |2\vec{k}\rangle = \sum_{\vec{k}} \beta_{\vec{k}}(0) \phi_{\vec{k}} e^{-iE_{\vec{k}}t/\hbar} \quad (7.352b)$$

As usual, the functions $\phi_{\vec{k}}$ satisfy the eigenvector equation

$$\hat{H}\phi_{\vec{k}} = E_{\vec{k}}\phi_{\vec{k}} \quad (7.353)$$

where $E_{\vec{k}} = E_{2,\vec{k}}$ and $\hat{H} = -\frac{\hbar^2}{2m_e} \nabla^2 \Psi + V\Psi$.

7.15.2 DENSITY OF ENERGY STATES FOR QUANTUM WELL

We can calculate the density of energy states by either of two methods. The first, more intuitive method appears at the end of Section 7.14. The present section substantiates the intuitive approach starting with the density of $\vec{k}$ states, which involves a Dirac delta function for the discrete values of k_z . Next we integrate over the density of states using a spherical surface. The integral reduces to the summation of 2-D density of $\vec{k}$ states. Finally, we take a derivative with respect to energy to find the density of energy states.

First find the density of energy states for the quantum well structure in the 3-D crystal. The first step consists of plotting the allowed wave vectors $\vec{k}$ assuming periodic boundary conditions in the x - and y -directions (but not for the confinement direction z). Assume the crystal has size $L_x = L_y = L \gg L_z$. For each value of k_z given by

$$k_z = \frac{n_z \pi}{L_z} \quad \text{for } n_z = 1, 2, \dots \quad (7.354a)$$

there exists a range of closely spaced values of k_x and k_y given by

$$k_x = \frac{2\pi n_x}{L} \quad \text{and} \quad k_y = \frac{2\pi n_y}{L} \quad \text{for } n_x, n_y = \pm 1, \pm 2, \dots \quad (7.354b)$$

The density of k -states for k_x and k_y (based on Equation 7.354b and periodic boundary conditions for those directions) does not depend on the 2-direction and its fixed endpoint boundary conditions. It might appear that the geometry of the well (infinite or finite) makes very little difference to the calculation of the density of states given that Figure 7.78 essentially plots the dispersion curves for E vs. k_x and k_y . Not true though because the bottom of each parabola depends on the confined energy values, which in turn depend on the allowed k_z . We will find $g_{\text{well}}^{(3-D)}(E) \sim \sum_{E_z} \dots \theta(E - E_z)$ where the step function $\theta(E - E_z) = +1$ for $E > E_z$ and zero otherwise. The well size L_z being much smaller than crystal size L produces k_z spacing much larger than either k_x or k_y . Figure 7.84 shows the large separation between the k_z values making the figure appear as multiple parallel planes spaced along the k_z -axis.

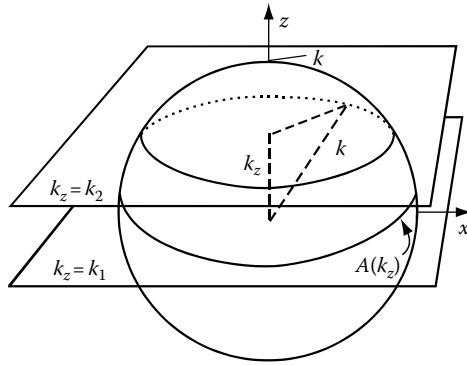


FIGURE 7.84 The allowed k_z points for the quantum well form planes. A sphere encloses points on the various planes. The sphere of radius k intersects the plane to form a circle.

To find the density of $|\vec{k}|$ -states, enclose the $\vec{k}$ points within a sphere and then k -volume integrate the density of $\vec{k}$ -states as usual. One can use a sphere because as shown in the following equation, m_e is assumed independent of direction and the energy is then symmetric in k_x , k_y , and k_z .

$$\begin{aligned} E &= E_{(z)n} + \frac{\hbar^2}{2m_e}(k_x^2 + k_y^2) \\ &= \frac{\hbar^2}{2m_e}(k_x^2 + k_y^2 + k_z^2) = \frac{\hbar^2 k^2}{2m_e} \end{aligned} \quad (7.355)$$

We will find that the geometry of the $\vec{k}$ states in Figure 7.84 reduces the integral to one involving cylindrical coordinates. The density of $\vec{k}$ -states involves closely spaced points forming parallel planes that come from macroscopic boundary conditions $L \sim 1$ cm while the spacing between planes comes from smaller microscopic boundary conditions $L_z \sim 50$ Å. We view the allowed values of k_z as discrete points rather than the more continuous ones in the planes. We can represent the allowed k_z using Dirac delta functions.

We now show the Dirac delta function nature of the k_z points. Figure 7.85 shows the allowed k_z points. A single point has a density function (1-D) represented by a delta function according to

$$g(k_z) = \delta(k_z - k_1)$$

so that the total number of states must be

$$N = \int dk_z \delta(k_z - k_1) = 1$$

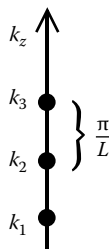


FIGURE 7.85 A line of discrete points with the same values as k_z for the quantum well.

For the case of the quantum well, we have discrete values of k_z , denoted by k_n , and we can write the density of states as

$$g(k_z) = \sum_{k_n} \delta(k_z - k_n) \quad (7.356a)$$

where we assume $k_n > 0$ to match the allowed values of k_z for the quantum well. Integrating to a fixed value of k along the z -axis provides

$$N = \int_0^k dk_z \sum_{k_n} \delta(k_z - k_n) = \sum_{k_n} \int_0^k dk_z \delta(k_z - k_n) = \sum_{k_n < k} \Theta(k - k_n) \quad (7.356b)$$

where the step function $\Theta(k_z - k_n)$ has the value of +1 for $k_z \geq k_n$ and zero otherwise. If there exists n values of k_z smaller than k then the number of states must be $N = n$. We find for n values of k_n smaller than k

$$N = \sum_{k_n < k} \Theta(k - k_n) = n \cong \frac{k_z}{(\pi/L_z)} = \frac{k_z L_z}{\pi} \rightarrow k_z = \frac{n\pi}{L_z} \quad (7.356c)$$

where k_z has been substituted for k . Actually we can obtain the same result by defining the average number of states per unit k_z as

$$\langle g \rangle \equiv \frac{1}{k_z} \int_0^{k_z} dk_z g(k_z) = \frac{1}{k_z} \int_0^{k_z} dk_z \sum_{k_n} \delta(k_z - k_n) = \frac{n}{k_z} = \frac{1}{\pi/L_z}$$

We therefore see that closely spaced points actually produce an average density of states as if the points were smeared out.

Now we can write the density of $\vec{k}$ -states assuming closely spaced points in a plane

$$g_{\text{well}}^{(3-D)}(\vec{k}) = \left(\frac{L}{2\pi}\right)^2 \sum_{k_n} \delta(k_z - k_n) \quad (7.357)$$

which has units of # k -states per k -volume (dividing by the crystal-area L^2 would include a factor of “per crystal area”). One can see this last equation is correct by supposing a single k -plane such as for k_1 in Figure 7.84, and integrating over k_z to find $L^2/(2\pi)^2$ which gives the density of k -states for a 2-D sheet as expected. The number of $\vec{k}$ -states within a sphere of fixed radius $k = |\vec{k}|$ must be

$$N(k) = \int_0^k d^3k g_{\text{well}}^{(3-D)}(\vec{k}) \quad (7.358a)$$

We see that $g_{\text{well}}^{(3-D)}(\vec{k})$ depends only on k_z , which suggests factoring the integral into one over an area and into another over k_z (Figure 7.84).

$$N(k) = \int_0^k dk_z g_z^{(1-D)}(k_z) \int_{A(k_z)} d^2k g_{\text{area}}^{(2-D)}(k_x, k_y) \quad (7.358b)$$

As shown in Figure 7.84, the integration along k_z extends to the radius of the circle, namely k . Substituting Equation 7.357 into Equation 7.358a shows that Equation 7.358b has the form

$$N(k) = \int_0^k d^3k g_{\text{well}}^{(3-D)}(\vec{k}) = \int_0^k dk_z \int_{A(k_z)} d^2k \left(\frac{L}{2\pi}\right)^2 \sum_{k_n} \delta(k_z - k_n) = \int_0^k dk_z \sum_{k_n} \delta(k_z - k_n) \int_{A(k_z)} d^2k \left(\frac{L}{2\pi}\right)^2$$

Note in particular the limits of the integral. Figure 7.84 shows that the area A (a circle) becomes smaller for larger values of k_n . That is, the area $A(k_z)$ depends on the discrete values of k_z , namely k_n but this will become apparent after integrating over the delta function. The area integral produces

$$N(k) = \int_0^k dk_z \sum_{k_n} \delta(k_z - k_n) \left(\frac{L}{2\pi}\right)^2 A(k_n) = \sum_{k_n < k} \left(\frac{L}{2\pi}\right)^2 A(k_n) = \sum_{k_n < k} g_k^{(2-D)} A(k_n) \quad (7.358c)$$

which shows the number of states must be computed from the number of states in the parallel planes. Recall that the 2-D density of $\vec{k}$ -states has the definition

$$g_k^{(2-D)} = \left(\frac{L}{2\pi}\right)^2 \quad (7.359)$$

The area of each plane enclosed by the sphere must be

$$A(k_n) = \pi(k^2 - k_n^2) \quad (7.360)$$

as shown in the figure. Therefore, the number of states in the sphere must be

$$N(k) = \sum_{k_n < k} g_k^{(2-D)} \pi(k^2 - k_n^2) = \sum_{k_n} g_k^{(2-D)} \pi(k^2 - k_n^2) \Theta(k^2 - k_n^2) \quad (7.361)$$

where the step function has been inserted in order to remove the restriction on the summation. When we differentiate this last equation, we must use a product rule on the last two terms. The derivative of the step function produces a term that we assume goes to zero

$$\pi(k^2 - k_n^2) \delta(k^2 - k_n^2) = \pi(k_n^2 - k_n^2) \delta(k^2 - k_n^2) \rightarrow 0$$

As an alternative, we could leave off the step function, take a derivative and then reinsert the step function to find the same result. We use the step function for notational convenience rather than for computation.

Now we can compute the electron density-of-states (EDOS) according to

$$g_{\text{well}}(E) = \frac{dN(k)}{dE} = \frac{dk}{dE} \frac{dN}{dk} \quad (7.362)$$

The derivative of $N(k)$ becomes

$$\frac{d}{dk} N(k) = \frac{d}{dk} \sum_{k_n} g_k^{(2-D)} \pi(k^2 - k_n^2) \Theta(k^2 - k_n^2) = g_k^{(2-D)} 2\pi k \sum_{k_n} \Theta(k^2 - k_n^2) \quad (7.363a)$$

where we have ignored the derivative of the step function. We need to differentiate the energy in Equation 7.355

$$\frac{dE}{dk} = \frac{d}{dk} \frac{\hbar^2 k^2}{2m_e} = \frac{\hbar^2}{m_e} k \quad (7.363b)$$

Combining Equation 7.363a and b into Equation 7.362 produces

$$\begin{aligned} g_{\text{well}}(E) &= \frac{dN(k)}{dE} = \frac{dk}{dE} \frac{dN}{dk} = \frac{m_e}{\hbar^2 k} g_k^{(2-D)} 2\pi k \sum_{k_n} \Theta(k^2 - k_n^2) \\ &= \frac{2\pi m_e}{\hbar^2} \left(\frac{L}{2\pi} \right)^2 \sum_{k_n} \Theta(k^2 - k_n^2) \end{aligned} \quad (7.364)$$

Finally rewrite the step function in terms of the energy

$$g_{\text{well}}(E) = \frac{2\pi m_e}{\hbar^2} \left(\frac{L}{2\pi} \right)^2 \sum_{E_n} \Theta(E - E_n) \quad (7.365)$$

Usually we define the density of states as either the (1) number of states per crystal area per energy or (2) as the number of states per crystal volume per energy. Therefore two versions of the density of states for the quantum well in a 3-D crystal can be encountered in the literature. In the first case, we divide by L^2 and in the second by $L^2 L_z$. The density of state can then be written in either of the following forms.

$$\begin{aligned} g_{\text{well}}^A(E) &= \frac{m_e}{2\pi\hbar^2} \sum_{E_n} \Theta(E - E_n) \quad \text{per xal area} \\ g_{\text{well}}^V(E) &= \frac{m_e}{2\pi\hbar^2 L_z} \sum_{E_n} \Theta(E - E_n) \quad \text{per xal vol} \end{aligned} \quad (7.366)$$

These equations do not include spin degeneracy.

The subbands produce the step functions in Equation 7.366. For energy E smaller than E_1 (Figure 7.86), the range ΔE does not include any of the subbands and therefore $g_{\text{well}} = 0$. As E increases to the range $E_1 < E < E_2$, the energy states in the first subband must be counted. The density of states for the first subband is independent of energy. For larger E , multiple subbands must be included in the energy range ΔE . We must therefore add multiple constants representing the 2-D density of states $\frac{m_e}{2\pi\hbar^2}$.

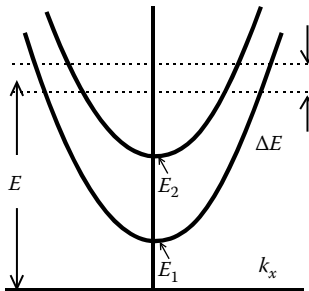


FIGURE 7.86 Density of states depends on the number of subbands at energy E .

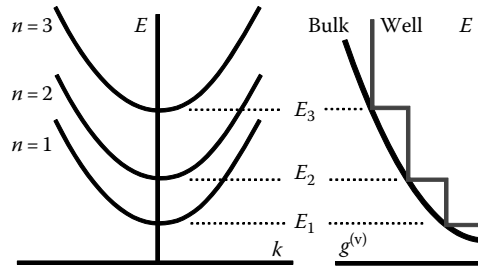


FIGURE 7.87 The density of energy states for the quantum well and its relation to the subband diagram.

Typically, books show the density of energy states for the quantum well next to a plot of the subbands as in Figure 7.87. Each 2-D plane of k vectors leads to a constant density of energy states (independent of energy). Larger energy E requires more subbands be included in the energy state counting. A step occurs in the well density of states at the start of each subband as shown. By the way, the step-like form for the well density of states means that thermal electrons occupy narrower range of energy than for the bulk material. This also means that the population inversion required for lasing also occupies a narrower range.

7.15.3 DENSITY OF ENERGY STATES FOR QUANTUM WIRE

The quantum wire confines the electron in two directions, say the x - and y -directions. For example, Figure 7.88 shows a GaAs “wire” embedded within AlGaAs. We assume the wire length along z has macroscopic size compared with the microscopic size along either x or y

$$L_x, L_y \ll L_z$$

The lengths along x or y are approximately 50–100 Å.

We can solve Schrödinger’s equation in the effective mass approximation for the infinitely deep well. The solutions along x and y must have a sinusoidal form while those for the z -direction can be taken as traveling waves. The allowed wave vectors can be written as

$$k_m^{(x)} = \frac{m\pi}{L_x} \quad k_n^{(y)} = \frac{n\pi}{L_y} \quad k_q^{(z)} = \frac{2q\pi}{L_z} \quad m, n = +1, +2, \dots \quad q = \pm 1, \pm 2, \dots \quad (7.367)$$

with the spacing between states

$$\Delta k_x = \frac{\pi}{L_x} \quad \Delta k_y = \frac{\pi}{L_y} \quad \Delta k_z = \frac{2\pi}{L_z} = \text{small} \quad (7.368)$$

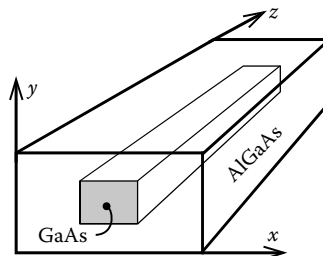


FIGURE 7.88 The quantum wire confines electrons along the x - and y -direction.

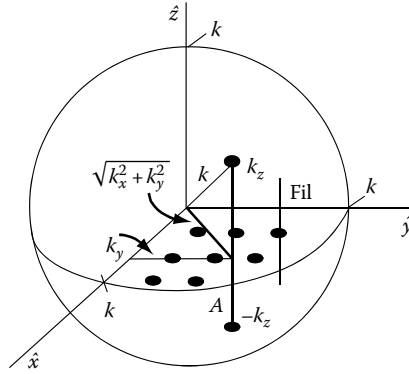


FIGURE 7.89 Each allowed k in the $k_x - k_y$ plane has a “filament” denoted 1-D representing the continuous range of k_z .

and the energy

$$E = E_x + E_y + E_z = \frac{\hbar^2 k_x^2}{2m_e} + \frac{\hbar^2 k_y^2}{2m_e} + \frac{\hbar^2 k_z^2}{2m_e} = \frac{\hbar^2 k^2}{2m_e} \quad (7.369)$$

where $k^2 = k_x^2 + k_y^2 + k_z^2$. The z -component of the wave vector k_z is essentially continuous.

As usual, the first step consists of plotting the allowed values of $\vec{k}$. The close spacing of the k_z points and wide spacing of k_x, k_y produces lines of allowed $\vec{k}$. Figure 7.89 shows one of the filaments for a point just in front of the k_y -axis. These filaments consist of the continuous range of k_z -values. The points shown in the $k_x - k_y$ plane are widely separated compared with those along the filament. As shown, the filament runs along the k_z -axis with both positive and negative values as required by the positive and negative integers q in Equation 7.367. Each point in the $k_x - k_y$ plane has a filament (denoted by “Fil” in the figure) running along the z -direction. Also notice that we only need to consider $k_x > 0$ and $k_y > 0$ because of the range of the integers m, n . Next, write the density of $\vec{k}$ states using the Dirac delta function similar to the previous topic. The $\vec{k}$ density of states can be written as

$$g_{\text{wire}}(\vec{k}) = g_{1\text{-D}} \sum_{k_m^{(x)}} \delta(k_x - k_m^{(x)}) \sum_{k_n^{(y)}} \delta(k_y - k_n^{(y)}) \quad (7.370)$$

where

$$g_{1\text{-D}} = L_z / \pi \quad (7.371)$$

The equation for energy has spherical symmetry (Equation 7.369), regardless of the discrete or nondiscrete nature of the components of the wave vectors. We therefore draw a sphere with fixed radius k . The number of states enclosed by the sphere can be written as

$$N(k) = \int dk'_x \int dk'_y \sum_{k_m^{(x)}} \delta(k'_x - k_m^{(x)}) \sum_{k_n^{(y)}} \delta(k'_y - k_n^{(y)}) \int_{-k_z}^{+k_z} dk'_z g_{1\text{-D}} \quad (7.372a)$$

The primes indicate the integration variables and Equation 7.372a integrates over the filament passing through the point (k_x, k_y) . The values of k'_z extend from the bottom of the sphere to the top. The values of the end points depend on the position of the filament so that $k_z = k_z(k_x, k_y)$. As shown in Figure 7.89, the coordinates (k_x, k_y, k_z) specify a point on the surface of the sphere. Carrying out the k_z integral gives

$$N(k) = \int dk'_x \int dk'_y \sum_{k_m^{(x)}} \delta(k'_x - k_m^{(x)}) \sum_{k_n^{(y)}} \delta(k'_y - k_n^{(y)}) 2g_{1-D} k_z(k'_x, k'_y) \quad (7.372b)$$

We need to be cognizant of the fact that the variables k_x, k_y, k_z are constrained to lie on the surface of the sphere with fixed radius k .

We can evaluate the two integrals in the previous equation. The integral over k_y provides

$$N(k) = 2g_{1-D} \int dk'_x \sum_{k_m^{(x)}} \delta(k'_x - k_m^{(x)}) \sum_{k_n^{(y)2} < k^2 - k_x^2} k_z(k'_x, k_n^{(y)}) \quad (7.373a)$$

We need to comment on the index $k_n^{(y)2} < k^2 - k_x^2$ for the second summation. The integration variable k'_y can range from 0 to $\sqrt{k^2 - k_x^2}$ since at the sphere surface in the $k_x - k_y$ plane, the distance from the origin must be the radius k of the sphere. The Dirac delta function requires $k_n^{(y)} < \sqrt{k^2 - k_x^2}$. Then in preparation for the energy variable, we can square both sides to find $k_n^{(y)2} < k^2 - k_x^2$. Similar to the previous topic, the “<” sign can be removed from the index on the second summation by including the step function Θ

$$N(k) = 2g_{1-D} \int_0^k dk'_x \sum_{k_m^{(x)}} \delta(k'_x - k_m^{(x)}) \sum_{k_n^{(y)2}} k_z(k'_x, k_n^{(y)}) \Theta(k_x^2 + k_n^{(y)2} < k^2) \quad (7.373b)$$

where the step function has been rearranged according to the prescription

$$\Theta(k_n^{(y)2} < k^2 - k_x^2) = \Theta(k_x^2 + k_n^{(y)2} < k^2)$$

The final integral provides

$$N(k) = 2g_{1-D} \sum_{k_m^{(x)2}} \sum_{k_n^{(y)2}} k_z(k_m^{(x)}, k_n^{(y)}) \Theta(k_m^{(x)2} + k_n^{(y)2} < k^2) \quad (7.374a)$$

Technically, the first summation should be summed in accordance with the constraint $k_m^{(x)2} < k^2$ but the step function already includes it. The value of k_z can be written as

$$k_z(k_m^{(x)}, k_n^{(y)}) = \sqrt{k^2 - k_m^{(x)2} - k_n^{(y)2}} \quad (7.374b)$$

since it must reside on the surface of the sphere.

The results in Equation 7.374a and b

$$N(k) = 2g_{1-D} \sum_{k_m^{(x)2}} \sum_{k_n^{(y)2}} \sqrt{k^2 - k_m^{(x)2} - k_n^{(y)2}} \Theta(k_m^{(x)2} + k_n^{(y)2} < k^2) \quad (7.375)$$

must be converted to the energy variable E given by Equation 7.369 as reprinted here

$$E = E_x + E_y + E_z = \frac{\hbar^2 k_x^2}{2m_e} + \frac{\hbar^2 k_y^2}{2m_e} + \frac{\hbar^2 k_z^2}{2m_e} = \frac{\hbar^2 k^2}{2m_e}$$

Defining E_{mn} as

$$E_{mn} = E_x + E_y = \frac{\hbar^2 k_m^{(x)2}}{2m_e} + \frac{\hbar^2 k_n^{(y)2}}{2m_e} \quad (7.376)$$

changes Equation 7.376 into

$$N(E) = 2g_{1-D} \sqrt{\frac{2m_e}{\hbar^2}} \sum_{E_{mn}} \sqrt{E - E_{mn}} \Theta(E_{mn} < E) \quad (7.377)$$

Finally, we find the density of energy states for quantum wire to be

$$g_{\text{wire}}(E) = \frac{dN(E)}{dE} = \frac{L}{2\pi} \sqrt{\frac{2m_e}{\hbar^2}} \sum_{E_{mn}} \frac{1}{\sqrt{E - E_{mn}}} \Theta(E_{mn} < E) \quad (7.378)$$

where as before, we ignore the derivative of the step function. As usual, we divide out the crystal dimensions. Sometimes the literature reports the number of states per crystal length L (and per energy) and at other times, the number of states per crystal volume (much more common). The results of the calculation can be written as

$$\begin{aligned} g_{\text{wire}}^{\text{Len}}(E) &= \frac{1}{2\pi} \sqrt{\frac{2m_e}{\hbar^2}} \sum_{E_{mn}} \frac{1}{\sqrt{E - E_{mn}}} \Theta(E_{mn} < E) \\ g_{\text{wire}}^{\text{Vol}}(E) &= \frac{1}{2\pi L_x L_y} \sqrt{\frac{2m_e}{\hbar^2}} \sum_{E_{mn}} \frac{1}{\sqrt{E - E_{mn}}} \Theta(E_{mn} < E) \end{aligned} \quad (7.379)$$

These equations do not include the factor of 2 for electron spin degeneracy.

The density of energy states for the quantum wire appears in Figure 7.90. The density of states rapidly falls off after each new confined energy. Combining this density of states with the Fermi-Dirac distribution produces very sharp electron (vs. energy) distributions.

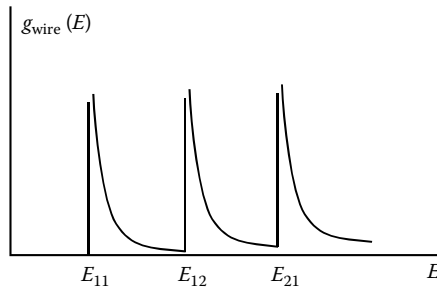


FIGURE 7.90 The density of energy states for the quantum wire.

As an important note, we have assumed an infinitely deep well. The energy levels for the finitely deep well have different values than those for the infinitely deep ones. Therefore we expect a density of states function that appears similar to Figure 7.90 except that the steps must occur at different values of E . Further, the finitely deep well only binds the electron for a fixed number of states; the remaining states correspond to plane waves. Therefore only a finite number of steps appear in the density of states plot.

7.16 REVIEW EXERCISES

These problems do not necessarily state all necessary assumptions needed to work the problem. It will be up to the reader to state any missing assumptions.

- 7.1** Show the relation $r^2 + t_{12}t_{21} = 1$ for the simple interface as discussed in Section 7.2.
- 7.2** Using the definitions for the transmittance T and R from Section 7.2, show that $R = 1$ when $T = 0$ for real k_1 and complex k_2 (i.e., $E < V_2$).
- 7.3** Work out the reflectance and transmittance for the step potential in Section 7.2 when the effective masses m_1 and m_2 for $z < 0$ and $z > 0$, respectively, are different.
- 7.4** Discuss how current conservation $I_{\text{inc}} = I_{\text{ref}} + I_{\text{trans}}$ holds or does not hold for the case in Section 7.2 when the masses m_1 and m_2 for $z < 0$ and $z > 0$ are different.
- 7.5** Consider a barrier in free space (Figure P7.5) with an electron incident from either side. Suppose the phase α of beam $a_2 = a_{20}e^{-ik_2x+\alpha}$. Show the effect on the two output beams.

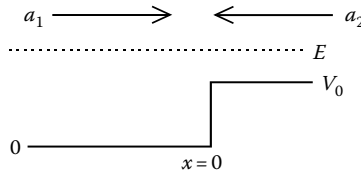


FIGURE P7.5 Electrons incident from the right and left.

7.6 Starting with (Section 7.3)

$$|S_{11}|^2 = R \frac{1 + \exp(-4\phi_i) - 2 \exp(-2\phi_i) \cos(2\phi_r)}{1 + R^2 \exp(-4\phi_i) - 2R \exp(-2\phi_i) \cos(2\phi_r + 2\alpha)}$$

show

$$|S_{11}|^2 = R \frac{1 + \frac{\mathcal{R}^2}{R^2} - 2 \frac{\mathcal{R}}{R} \cos(2\phi_r)}{1 + \mathcal{R}^2 - 2\mathcal{R} \cos(2\phi_r + 2\alpha)}$$

where $\mathcal{R} = R \exp(-2\phi_i)$, $R = |r|^2$, $r = |r|e^{i\alpha}$, and $\phi = \phi_r + i\phi_i$.

- 7.7** Consider Figure P7.7 with $V = 0$ for regions 1 and 3, and $V = V_0$ for region 2, and $E > V_0$. Starting with three transfer matrices and real wave vectors (not the complex ones used in Section 7.3), show all steps leading to

$$J_{\text{ref}} = r^2 \frac{2R - 2R \cos(2k_2L)}{1 + R^2 - 2R \cos(2k_2L)} J_{\text{inc}} \quad \text{and} \quad J_{\text{trans}} = \frac{(1 - R)^2}{1 + R^2 - 2R \cos(2k_2L)} J_{\text{inc}}$$

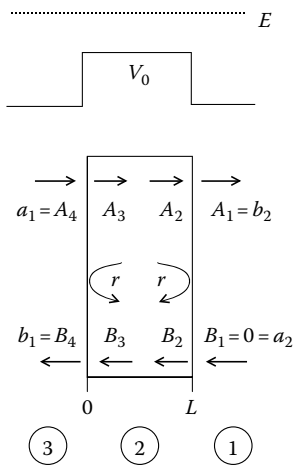


FIGURE P7.7 Potential barrier.

7.8 Repeat Problem 7.7 for $0 > E > V_0$ explicitly using $k_2 = i\kappa_2$ (and not the complex version found in Section 7.3) to show

$$\frac{J_{\text{trans}}}{J_{\text{inc}}} = \frac{|t|^4}{4 \sin^2(\kappa_2 L) + |t|^4}$$

7.9 Solve Schrödinger’s wave equation for the case of a free-space plane wave propagating from the right as shown in Figure P7.9. Assume the energy of the particle is larger than the top of the potential $E > V_0$. Using the solution, write the scattering matrix.

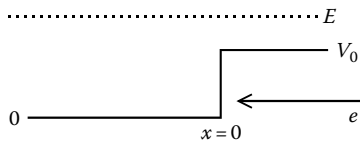


FIGURE P7.9 Particle incident from the right.

7.10 An electron traveling through free space from the left encounters a step potential particle at $x=0$ as shown in Figure P7.10. Assume the energy of the particle is smaller than the top of the potential $E < V_0$. By solving Schrödinger’s wave equation, write the reflected and transmitted amplitudes in terms of the incident amplitude.

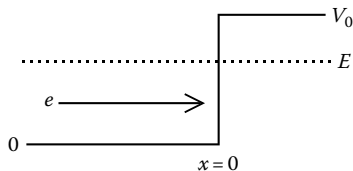


FIGURE P7.10 Particle incident from the left.

7.11 Discuss whether or not the reflectance $R = 1$ and transmittance $T = 0$ for an electron incident on a barrier when $0 < E < V_0$ as shown in Figure P7.11 when either (or both) the barrier height

and width becomes very large. Region 1 corresponds to $x < 0$, Region 2 corresponds to $0 < x < L$, Region 3 corresponds to $x > L$.

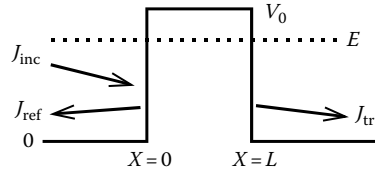


FIGURE P7.11 Reflected and transmitted current for a barrier.

- 7.12** Work out all of the details for the reflected current J_{ref} and the transmitted current J_{tr} in Figure P7.11 using the transfer matrices developed in the chapter. Assume $k_2 = i\kappa_2$ with $\kappa_2 = \text{real}$ and Region 2 refers to $0 < X < L$.
- 7.13** Using the results for quantum tunneling from the chapter, write the solution for the case when $L \rightarrow 0$ and $V_0 \rightarrow \infty$ such that $V_0 L = \text{constant}$.
- 7.14** Consider an infinitely large barrier at $x=0$ as shown in Figure P7.14. The infinite barrier requires $V_0 \rightarrow \infty$.
- Show that the penetration depth into region 2 must be zero. Use the expression for κ_2 in $k_2 = i\kappa_2$ take $V_0 \rightarrow \infty$. The penetration depth is defined to be the distance L such that $e^{-\kappa_2 L} = e^{-1}$.
 - Find the reflected amplitude b_1 in terms of the incident amplitude a_1 . You should find $r = -1$.

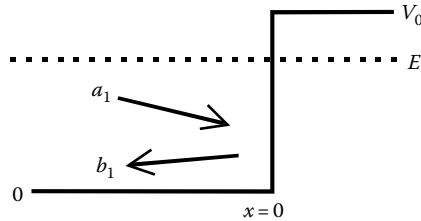


FIGURE P7.14 The barrier becomes infinite when $V_0 \rightarrow \infty$.

- 7.15** Is it possible to use transfer and scattering matrices to determine the standing waves in an infinitely deep well.
- 7.16** Use transfer and scattering matrices to find the amplitude A_3 at $X=L$ in Figure P7.16 in terms of A_1 .

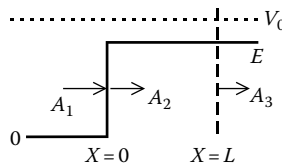


FIGURE P7.16 Interface and guide.

- 7.17** The $E - k$ relationship around minimum of the GaAs conduction band is slightly nonparabolic and has the form $E - E_c = ak^2 - bk^4$. Find the effective mass as a function of k .
- 7.18** Electrons in GaAs can transfer from the conduction band Γ minimum to the L minimum. If electrons transfer from Γ to L , does their effective mass increase or decrease? If the transfer occurs when the charge is moving, then what happens to the current density?

7.19 Section 7.4 finds the energy gap by examining the expectation value of the Hamiltonian for standing wave states. The section defines

$$\Delta E = \langle \psi_- | \hat{H} | \psi_- \rangle - \langle \psi_+ | \hat{H} | \psi_+ \rangle = E_- - E_+$$

where $\hat{H} = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x)$ and $\hat{H}\psi_{\pm} = E_{\pm}\psi_{\pm}$, and

$$\psi_+ \sim e^{ikx} + e^{-ikx} \sim C_+ \cos(kx) = C_+ \cos\left(\frac{\pi x}{a}\right)$$

$$\psi_- \sim e^{ikx} - e^{-ikx} \sim C_- \sin(kx) = C_- \sin\left(\frac{\pi x}{a}\right)$$

and $V(x) = \sum_G V_G \frac{e^{iGx}}{\sqrt{a}} = \sum_n V_n \frac{e^{i\frac{2\pi n x}{a}}}{\sqrt{a}}$. Show $|\Delta E| = 2|V_1|$.

- 7.20** Using the information in the previous problem, show $E_+ = E_0 + V_1$ where E_0 is the energy of the free particle with wave vector k .
- 7.21** Find the effective mass matrix for $E - E_c = 3(k_x - 1)^2 + 3(k_y - 2)^2$. Be sure to discuss the effective mass m_{zz} .
- 7.22** If $\vec{a} = a\hat{x}$ and the effective mass can be represented as a tensor, show the force can be written as $\vec{F} = \hat{x}m_{xx}a + \hat{y}m_{yx}a + \hat{z}m_{zx}a$
- 7.23** Find the density of electron states for a 1-D crystal assuming periodic boundary conditions over the length L .
- 7.24** The Kronig–Penney model predicts energy bands and gaps. Figure 7.48 in the chapter plots Equation 7.174 for two different values of U . Find the bandwidth and energy gap for the region near $(0, \pi)$ for the two different values of U .
- 7.25** This problem gives an alternative derivation to the density of states for unsymmetrical bands in Section 7.13.10. Define new wave vectors $\kappa_i = \frac{k_i}{m_i}$ where $i = x, y, z$ and m_i is the effective mass in direction i .
- Show the spacing between new κ_i values must be $\Delta\kappa_i = \frac{2\pi}{L\sqrt{m_i}}$
 - Show $E = \hbar^2\kappa^2/2$
 - Find the number of states N in the spherical surface of part b.
 - Differentiate to find the density of states.
- 7.26** Draw the electron and hole distribution as a function of energy for the quantum well and quantum wire. Assume the Fermi–Dirac distribution holds.
- 7.27** Suppose that a crystal is invariant with respect to operations $\hat{O}$ that constitute a group. Suppose the vector $|n\rangle$ is an eigenvector of the Hamiltonian $\hat{\mathcal{H}}|n\rangle = E_n|n\rangle$. Show the vector $|\psi\rangle = |n\rangle + \hat{O}|n\rangle + \hat{O}^2|n\rangle$ must also be an eigenvector producing the eigenvalue E_n . *Hint:* consider commutation relation $[\hat{\mathcal{H}}, \hat{O}]$.
- 7.28** Repeat the derivation for the CB quantum well in Section 7.14.2. Use an effective mass diagonal in x, y, z . Assume all entries of the effective mass tensor are different and given by m_x, m_y, m_z .
- 7.29** Show the relation

$$\underbrace{\cos(\kappa_{[w\xi]}\xi)\cos(\kappa_{[b\eta]}\eta) - \frac{\kappa_{[b\eta]}^2 - \kappa_{[w\xi]}^2}{2\kappa_{[w\xi]}\kappa_{[b\eta]}} \sin(\kappa_{[w\xi]}\xi)\sin(\kappa_{[b\eta]}\eta)}_{F(E)} = \cos(ka)$$

Starting with

$$\left[\sin(\kappa_{[w\xi]}\xi) + \frac{\kappa_{[w\xi]}}{\kappa_{[b\eta]}} e^{ika} \sin(\kappa_{[b\eta]}\eta) \right] [-\kappa_{[w\xi]} \sin(\kappa_{[w\xi]}\xi) - \kappa_{[b\eta]} e^{ika} \sin(\kappa_{[b\eta]}\eta)] + \\ - [\cos(\kappa_{[w\xi]}\xi) - e^{ika} \cos(\kappa_{[b\eta]}\eta)] [\kappa_{[w\xi]} \cos(\kappa_{[w\xi]}\xi) - \kappa_{[w\xi]} e^{ika} \cos(\kappa_{[b\eta]}\eta)] = 0$$

given in Section 7.8.

7.30 Show the following states are orthonormal

$$|X\rangle = \frac{1}{\sqrt{2}} \{ |\ell = 1, \ell_z = 1\rangle - |\ell = 1, \ell_z = -1\rangle \} \sim \frac{1}{\sqrt{2}} \{ Y_{1,1} + Y_{1,-1} \} = \sqrt{\frac{3}{4\pi}} \frac{x}{r}$$

$$|Y\rangle = \frac{i}{\sqrt{2}} \{ |\ell = 1, \ell_z = 1\rangle + |\ell = 1, \ell_z = -1\rangle \} \sim \frac{i}{\sqrt{2}} \{ Y_{1,1} + Y_{1,-1} \} = \sqrt{\frac{3}{4\pi}} \frac{y}{r}$$

$$|Z\rangle = |\ell = 1, \ell_z = 0\rangle \sim Y_{10}(\theta, \varphi) = \sqrt{\frac{3}{4\pi}} \cos \theta = \sqrt{\frac{3}{4\pi}} \frac{z}{r}$$

7.31 Derive all the terms in the matrix

$$\underline{\mathcal{H}} = \begin{bmatrix} E_s & 0 & kP & 0 \\ 0 & E_p - \Delta/3 & \sqrt{2}\Delta/3 & 0 \\ kP & \sqrt{2}\Delta/3 & E_p & 0 \\ 0 & 0 & 0 & E_p + \Delta/3 \end{bmatrix}$$

7.32 Show the approximations for the degenerate k-p theory

$$\mathcal{E}_{\text{LH}} = 0 + \varepsilon = -\frac{2(kP)^2}{3E_g}$$

$$\mathcal{E}_{\text{SO}} = -\Delta - \frac{(kP)^2}{3(E_g + \Delta)}$$

7.33 Show the approximate functions from the degenerate k-p theory

$$u_{\text{SO},\vec{k}} = \frac{1}{\sqrt{3}} |X - iY \uparrow\rangle + \sqrt{\frac{1}{3}} |Z \downarrow\rangle \\ u_{\text{SO},\vec{k}} = \frac{-1}{\sqrt{2}} |X + iY \uparrow\rangle$$

7.34 Draw the electron and hole distribution as a function of energy for the quantum well and quantum wire. Assume the Fermi-Dirac distribution holds.

7.35 Find the density of energy states for the finitely deep quantum well in a 3-D crystal.

7.36 Find the solutions to the 1-D Schrödinger equation for periodic boundary conditions (BCs) over $0 - L$ by separating variables and applying the BCs.

7.37 Suppose a crystal extends from $x = -W$ to $x = +W$ with a quantum well centered between $x = -L$ to $x = +L$. Find the density of states as a function of x .

7.38 Repeat the previous problem assuming a small voltage V_x is applied to the crystal. Assume for convenience that the voltage does not affect the shape of the well. Assume negligible current flow and that $V_x = 0$ at the center of the well.

- 7.39 Discuss the density of states for a finitely deep well. How does it differ from the infinitely deep well?
- 7.40 Assume five infinitely deep wells of width L , separated by width b . Write a piecewise relation for the density of states as a function of x .
- 7.41 How does the density of states in the previous problem change for five finitely deep wells?
- 7.42 Do a library search to find the meaning and use for “antibonding” states.
- 7.43 Repeat the derivations for the k - p theory. Fill in the details.

REFERENCES AND FURTHER READINGS

General References

1. Blakemore J.S., *Solid State Physics*, 2nd ed., W.B. Saunders Company, Philadelphia, PA (1974).
2. Ashcroft N.W. and Mermin N.D., *Solid State Physics*, Holt, Rinehart & Winston, New York 1976.
3. Kittel C., *Introduction to Solid State Physics*, 5th ed., John Wiley & Sons, New York (1976).
4. Bhattacharya P., *Semiconductor Optoelectronic Devices*, 2nd ed., Prentice Hall, Upper Saddle River, NJ (1997). Good general reference on most aspects of solid state including fabrication, electronic processes, bands, junctions, contacts and optoelectronic devices.
5. Sze S.M., *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, New York (1981).
6. Brennan K.F., *The Physics of Semiconductors with Applications to Optoelectronic Devices*, Cambridge University Press, Cambridge, U.K. (1999).
7. Pierret R.F., *Advanced Semiconductor Fundamentals, Volume VI in the Modular Series on Solid State Devices*, edited by R.F. Pierret and G.W. Neudeck, Addison Wesley Publishing, Reading, MA (1989). Very thin and readable text.

Effective Mass, Bands and k - p Theory

8. Datta S., *Quantum Phenomena, Volume VIII in the Modular Series on Solid State Devices* edited by R.F. Pierret and G.W. Neudeck, Addison Wesley Publishing Company, Reading, MA (1989).
9. Yu P.Y. and Cardona M., *Fundamentals of Semiconductors: Physics and Materials Properties*, 2nd ed., Springer, Berlin (1999).
10. Chuang S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, Inc., New York (1995).

Quantum Structures

11. Davies J.H., *The Physics of Low Dimensional Semiconductors*, Cambridge University Press, Cambridge, U.K. (1998).
12. Jaros M., *Physics and Applications of Semiconductor Microstructure*, Clarendon Press, Oxford (1989).
13. Davies J.H. and Long A.R., Eds., *Physics of nanostructures, Proceedings of the Thirty-Eighth Scottish Universities Summer School in Physics*, St. Andrews, July–August 1991, Published by SUSSP Publications (Edinburgh) and IOP Publishing Ltd. (London), 1992. ISBN:0-7503-0169-4 (pbk), 0-7503-0170-8 (hbk), published 1992.

Carrier Transport

14. Ferry D.K., *Semiconductor Transport*, Taylor & Francis, New York (2000). A good readable book.
15. Datta S., *Quantum Transport: Atom to Transistor*, Cambridge University Press, New York (2005).
16. Rammer J., *Quantum Transport Theory*, Perseus Books, Reading, MA (1998).
17. Lundstrom M., *Fundamentals of Carrier Transport*, 2nd ed., Cambridge University Press, Cambridge, U.K. (2000).

8 Statistical Mechanics

Modeling the behavior of real devices requires information on the types and numbers of carriers in each band as well as the mechanisms affecting mobility. The number and energy distribution of carriers is determined by whether or not the particles are distinguishable and whether or not they are Fermions. Electrons, as Fermions, follow the Fermi-Dirac distribution that can be found from the basic definition of entropy by incorporating the Pauli exclusion principle and indistinguishability. Knowing the Fermi-Dirac distribution and the density of states allows one to calculate a wide variety of properties for devices.

The chapter first reviews topics in thermodynamics and statistical mechanics that form the basis for equilibrium carrier statistics. These sections develop the idea of ensembles, systems, and reservoirs. The development includes the derivation of the Boltzmann distribution using two methods. The first method uses the definition of thermal equilibrium. The second method appeals to the ensemble and maximizes the entropy by maximizing the number of states available to a system. Semiconductors require the Fermi-Dirac distribution, which can be derived using the same ensemble methods. However, the Boltzmann distribution sometimes serves as a useful approximation to the Fermi-Dirac distribution. The chapter applies the Fermi-Dirac distribution to the pn junction and derives the diode current-voltage characteristics.

Any description of electronic and optoelectronic devices must necessarily focus on equilibrium and nonequilibrium processes in semiconductors. Equilibrium statistics for carrier occupation numbers describes the number of carriers in the conduction band without the application of light or voltage. In other words, the equilibrium statistics describe a type of quiescence. Applying light or voltage necessarily upsets the equilibrium conditions and changes the carrier occupation numbers. Therefore, the probability that an electron occupies a given state must change and the new distribution must be described by nonequilibrium statistics. The chapter presents the equilibrium statistics and focuses on the Fermi function, carrier density, carrier recombination, and generation. Electrical conduction and photoconduction can be expected to involve nonequilibrium statistics and will be left for books on the physics of optoelectronics.

8.1 INTRODUCTION TO RESERVOIRS

The reservoir has played a very important role since before the inception of thermodynamics. We most commonly recognize it as the “large bath” that maintains the temperature of the object. The reservoir also provides a conceptual basis for deriving many thermodynamical properties and for the study of the thermal systems on the microscopic level—statistical mechanics.

The reservoir also becomes important for the study of optoelectronics. The semiclassical gain and rate equations for light emitters and detectors can be derived starting with the Hamiltonian

$$\hat{H} = \hat{H}_0 + \hat{H}_{\text{other}} = \hat{H}_a + \hat{V} + \hat{H}_{\text{other}}$$

where

$\hat{H}_a$ denotes the free-atom Hamiltonian (such as for a single-atom detector or emitter)

$\hat{V}$ represents the semiclassical matter-light interaction

The term $\hat{H}_{\text{other}}$ refers to other influences on the smaller atomic system (consisting of an atom in this case). The “other influences” include the pump (power source) and collisions between the atom in question and other atoms or phonons (etc.). The density operator must then satisfy the Liouville

equation for the density operator $\hat{\rho}$ (as shown in the companion volume on the physics of optoelectronics).

$$\frac{\partial \hat{\rho}}{\partial t} = \frac{1}{i\hbar} [\hat{H}_o, \hat{\rho}] + \left(\frac{\partial \hat{\rho}}{\partial t} \right)_{\text{other}} = \frac{1}{i\hbar} [\hat{H}_o, \hat{\rho}] + \left(\frac{\partial \hat{\rho}}{\partial t} \right)_{\text{pump}} + \left(\frac{\partial \hat{\rho}}{\partial t} \right)_{\text{coll}} + \left(\frac{\partial \hat{\rho}}{\partial t} \right)_{\text{spont}}$$

The commutator involving the Hamiltonian $\hat{H}_o = \hat{H}_a + \hat{V}$ provides the evolution of the density operator due to the perturbing potential $\hat{V}$ and due to the natural motion of the electron within the atom. The density operator describes the statistical average and the microscopic quantum mechanical average as an extension to the wave function. The derivative terms on the right-hand side represent the effects of the pump, collisions, and spontaneous recombination. They come from relaxation effects and they therefore usually lead to decaying exponentials. These “other” terms, the derivatives, can be modeled by the effect of a thermal reservoir upon the smaller atomic system. The reservoir induces rapid fluctuations (Langevin noise) as well as damping. This section introduces the notion of a reservoir and discusses the associated fluctuation-dissipation theorem.

8.1.1 DEFINITION OF RESERVOIR

We divide a complete system into a small system under study and a collection of reservoirs. The reservoirs have an extremely large number of degrees of freedom and provide equilibrium for the smaller system. For example, a reservoir of two-level atoms or harmonic oscillators necessarily contains a large number of atoms or oscillators. For a more abstract example, a reservoir of light consists of a set of modes where the number of such modes must be extremely large. Typically, a specific energy distribution must be assumed to exist in a reservoir. For example, if the reservoir consists of point particles (such as gas molecules) then one might assume a Boltzmann distribution for the energy.

Bringing the reservoir into contact with the small system allows energy to flow between the system and the reservoir (Figure 8.1). The reservoir has such a large number of degrees of freedom that any energy transferred from the small system to the reservoir has negligible affect on the reservoir energy distribution. For a concrete example, suppose the small system consists of a single gas molecule and the reservoir has a large number of molecules all at thermal equilibrium (i.e., a Boltzmann energy distribution). The temperature of the small system will eventually match the temperature of the larger system. However, temperature measures kinetic energy in this case. Therefore, to say that the temperatures are the same is to say that the average kinetic energy of the single molecule is the same the average kinetic energy of all the molecules in the reservoir. Note the use of the word “average” which indicates the possibility (and actual fact) that some molecules move faster than others.

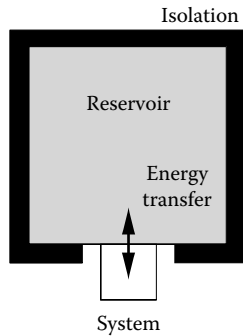


FIGURE 8.1 The reservoir can exchange energy with the system under study.

Suppose the molecule in the small system has a much larger than average kinetic energy (maybe a factor of 10). The extra energy must eventually transfer to the reservoir. This extra energy is distributed to all of the molecules in the reservoir, which makes negligible changes in the total reservoir distribution. In effect, the reservoir has “absorbed” the “extra” system energy and the motion of the single molecule must be “damped.”

The reservoir energy distribution defines average quantities for the reservoir. The contact between the two systems brings the small system into equilibrium with the reservoir, which therefore defines the average quantities for the small system. Suppose the single atom in the small system to initially be in equilibrium with the reservoir. Occasionally, a large chunk of energy will be transferred from the reservoir to the small system as a thermal fluctuation. As a result, the single atom will have more energy than its equilibrium value. Eventually this extra energy will damp out due to reservoir interactions. The correlation between fluctuations is assumed to occur on such short times scales as to be negligible. The process of transferring energy between the small system and the reservoir constitutes an example application of the fluctuation-dissipation theorem. The theorem basically states that a reservoir (or other system) both damps the small system and induces fluctuations in the small system. The two processes go together and cannot be separated. Often, on a phenomenological level, the fluctuations are included in rate equations through a Langevin function.

8.1.2 EXAMPLE OF FLUCTUATION-DISSIPATION THEOREM

Brownian motion of a small particle in a liquid consists of rapid, uncorrelated movements. The motion of the small particle is a result of the interaction between the particle and its liquid environment, which acts as a reservoir. The phenomenological equations can be derived from Newton’s second law. For one-dimensional motion

$$m\ddot{x} = -m\gamma\dot{x} + f(t) \quad (8.1)$$

where $f(t)$ represents the Langevin force and the term proportional to the velocity is the damping term. The discussion in the previous Section 8.1.1 demonstrates the intimate relation between the damping term and the Langevin force $f(t)$. Including one term in an equation necessarily requires the other one to be there. We define the Langevin force $f(t)$ to rapidly vary and to have an average value of $\langle f \rangle = 0$.

The fluctuations associated with the Langevin force are “stationary” with exceedingly small correlation times. Stationary means that the probability distribution P for the fluctuations does not depend on the origin of time. An ergodic process assumes that the average of a function $f(t)$ can be computed by either

$$\langle y(t) \rangle = \frac{1}{\tau} \int_0^\tau dt y(t) \quad (\text{for sufficiently large } \tau)$$

or by using an ensemble average

$$\langle y \rangle = \int dy y P(y)$$

where $P(y)$ is the probability density. The average of Equation 8.1 can be handled by either method. To use the time average, the time interval τ must be long compared with the correlation time but short compared with the time scale of interest. In this way, $\langle f \rangle = 0$ but $\langle x \rangle$ can still depend on time τ .

Taking the average of Equation 8.1 gives

$$m \frac{d^2}{dt^2} \langle x \rangle + m\gamma \frac{d}{dt} \langle x \rangle = \langle f(t) \rangle = 0 \quad (8.2a)$$

or equivalently

$$m \frac{d}{dt} v + m\gamma v = 0 \quad \text{where } v = \frac{d}{dt} \langle x \rangle \quad (8.2b)$$

Equation 8.2b shows how the usual form of Newton's law with only the damping term can be recovered from one with the Langevin term. In particular, it points out that many classical equations apply to the average motion of all the particles while ignoring departures from the average.

The simple differential equation in Equation 8.2b has the solution

$$v(t) = v(0) \exp(-\gamma t) \quad (8.3)$$

from which one can find the position as a function of time in the usual manner of integration. However, one can see that the average position will be zero if the initial velocity is zero.

It can be shown that the variance of the position can be nonzero even with an initial macroscopic velocity of zero. This means that the particle moves away from the starting point even though the average force is zero. The nonzero variance in particle position occurs because of the fluctuations induced in the velocity of the particle by the reservoir.

8.1.3 RESERVOIRS FOR OPTICAL EMITTER

What are the reservoirs for the optical emitter? First consider collisions. For simplicity, consider an emitter with a single atom. Assume the single atom to be embedded in a “background material” such as a crystal. The reservoir might consist of phonons that exist on the lattice or free electrons that can participate in collisions. The “background material” as part of the emitter has a specific temperature and composition. This means that the reservoir might be assumed to have a Boltzmann distribution characterized by a certain temperature T .

The reservoir for spontaneous emission consists of the collection of all optical modes in space—an extremely large number. We think of the “mode” as a place to dump photons. A mode can be characterized by a given wave vector and polarization. We typically picture them as “empty” traveling waves (the quantum vacuum) devoid of photons. When an excited single atom emits a photon, the photon is “absorbed” by the reservoir and never interacts with the atom again. For this reason, people refer to the interaction of the (spontaneous emission) reservoir and atom as an irreversible process. The spontaneous emission process can be reversed only if the emitted photon can interact with the atom again. The reabsorption of the photon by the atom can be accomplished if the atom exists within a Fabry-Perot cavity, for example.

8.1.4 COMMENT

The Liouville equation for the density operator is essentially a differential equation for the energy level occupation number (i.e., $\langle n | \hat{\rho} | n \rangle$) and the induced polarization (off-diagonal terms). A quantum mechanical reservoir gives rise to damping and fluctuation terms in the Liouville equation (a.k.a., master equation). The damping term appears as

$$\left. \frac{\partial \hat{\rho}}{\partial t} \right|_{\text{other}}$$

while, as expected, the average of the fluctuations disappears. One can use the trace over the reservoir states to calculate the average. The tracing operation produces a zero average for the fluctuations and removes the reservoir degrees of freedom from the differential equation.

The formalism can be applied to spontaneous and stimulated emission from atoms. The density operator is used so that various types of EM states (Fock, coherent, and squeezed) can be used as well as various atomic states. The density operator/matrix accounts for all possible knowledge of the system.

8.2 STATISTICAL ENSEMBLES AND INTRODUCTION
TO STATISTICAL MECHANICS

The thermal reservoir finds wide-ranging applications from holding constant the temperature of a smaller system to allowing one to derive the ubiquitous Boltzmann distribution. The Boltzmann distribution characterizes the classical thermal reservoir. This section introduces the microcanonical, canonical, and grand canonical ensembles and compares/contrasts them with the reservoir.

8.2.1 MICROCANONICAL ENSEMBLE, ENTROPY, AND STATES

The behavior of physical systems composed of a large number of particles can be predicted using probability theory. Boltzmann introduced the ensemble as a mental construct to calculate these probabilities. Contemporaries of Boltzmann considered the use of probability heretical. At the time, the molecular nature of matter was not well established. The statistical mechanics represents one of the first departures from the classical possibility of knowing all the positions and momentum of all the particles in the system. However, unlike quantum mechanics, the incomplete knowledge stems from our finite ability to measure and track the trajectory of a large number of independent particles (Avogadro’s number!).

A system consists of constituents such as molecules, atoms, electrons, or sometimes highlights one aspect such as spin. These are generally entities that carry energy. Certain “macroscopic” parameters specify the system. The ensemble (Figure 8.2) consists of a large number of systems with exactly the same set of parameters with the same values for those parameters. The micro-canonical ensemble consists of duplicate systems each specified by the number of particles N in the system, the *total* energy E of those particles, and the volume V of the system. Point particles have kinetic and potential energy whereas 3-D objects additionally have rotational and vibrational energy. Each system in the ensemble must have identical values for N, V, E . The specific values of the parameters N, V, E define the “macrostate” of the system. Usually, we do not “exactly” know the total energy of the system and therefore, the total energy E must be specified as $E \pm \Delta E$. The same can be said of number N and volume V but we mainly concentrate on the energy. All systems in the ensemble must reside in identical macrostates; however, this does not require the particles comprising one system to have the same internal coordinates as another of the systems. For example, gas molecules in system #1 might have different positions than molecules in system #2. In reality, we examine a single system and consider the duplicate systems as “make believe.”

Particles comprise each system of the ensemble. The position and momentum for each gas particle describe the “classical microstate” of a system. For example, consider a box with two molecules. To specify the state of this classical system, we must provide x, y, z coordinates and x, y, z momentum (speed) for each molecule. Each molecule has 3 degrees of freedom (x, y, z). Therefore, we need to specify the number

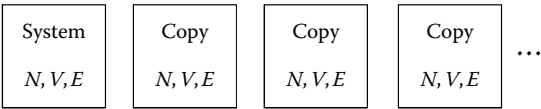


FIGURE 8.2 An ensemble of systems.

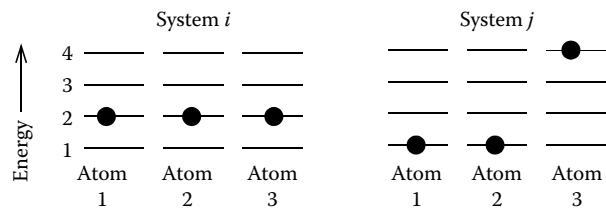


FIGURE 8.3 Two different microstates with the same energy $E=6$.

$$(\text{\# molec}) (\text{\# degrees of freedom}) \left(\begin{matrix} \text{two sets of values} \\ \text{i.e., pos. \& mom.} \end{matrix} \right) = 2 * 3 * 2 = 12$$

of phase space coordinates. A system with N molecules requires us to specify $6N$ values. Classically, if we know the position and momentum of all the particles then we also know the trajectories and total energy for all times.

We specify a “quantum mechanical microstate” by specifying the eigenvalues of all commuting operators. For N noninteracting particles in a quantum well, we must specify the energy level occupied by each particle.

It should be clear that for either the classical or quantum mechanical system, the specified macrostate can be achieved by any number of microstates. Figure 8.3 shows an example for two possible microstates that provide a total energy of $E=6$. Notice that we assume one of the systems has all atoms in energy state #2 while another system has atoms with energy 1, 1, 4. Each set of numbers specifies a different microstate. The basic postulate of statistical mechanics assumes that all microstates with the same energy must be equally probable.

The entropy S of a given system can be defined by

$$S = k \text{Ln}(\Omega) \tag{8.4}$$

where Ω represents the number of distinguishable ways the constituents of a system can occupy the available microstates, and k is the Boltzmann constant (see examples below). As we will see, the entropy describes the disorder of a system. The first example below shows systems occupying larger numbers of states must be more disordered and therefore have larger entropy. For example, suppose one were to “tidy-up” a desk by placing all of the objects (pens, paper clips, etc.) into small holders. The system (desk and objects) exhibit a high degree of order because the objects have been confined to small spaces. After a hard day of work, the objects migrate to other parts of the desk and occupy larger numbers of states (other portions of the desk). Therefore, we see the entropy must increase. Nature tends toward states of greatest entropy. Only doing work (adding energy) on a system can reverse the disorder and decrease the entropy. So for example, molecular chains can be grown by adding energy and thereby entropy.

Books on thermodynamics shows how the entropy S determine the macroscopic quantities according to

$$\left(\frac{\partial S}{\partial E} \right)_{N,V} = \frac{1}{T} \quad \left(\frac{\partial S}{\partial V} \right)_{N,E} = \frac{P}{T} \quad \left(\frac{\partial S}{\partial N} \right)_{E,V} = -\frac{\mu}{T} \tag{8.5}$$

These relations link the number $\Omega = \Omega(N, V, E)$ to the macroscopic parameters through the entropy S where E represents the total energy of the system (the sum of the energy of eigenstates in a microstate), and where the subscripts indicate the quantities to hold constant. The symbol μ represents the chemical potential.

Example 8.1

Consider two marbles in four possible bins making up a larger 2×2 square. Note the definitions whereby “bins” (as the smallest unit) makeup the larger “squares” are arranged in either an “overall rectangular pattern” or an “overall triangular pattern.” Assume both marbles can be in the same bin. The top part of Figure 8.4 shows 16 possible configurations for distinguishable marbles (maybe one is red and the other is white). Each configuration represents a microstate of the system. The entropy must be

$$S = k \ln(\Omega) = k \ln(16) = 2.8 k$$

The second part of the figure (shaped like a triangle) shows 10 different configurations for indistinguishable particles. Therefore, the entropy can be written as

$$S = k \ln(\Omega) = k \ln(10) = 2.3 k$$

In both cases the entropy S depends on the total number of states Ω available to the systems. If the large square were to have 4×4 bins but only 2×2 of them could be occupied by marbles then the entropy would remain unchanged from the values just calculated. In both cases, as the number of marbles or the number of available bins increases, so too does the entropy because the number of possible microstates must increase.

In connection with the bottom part of Figure 8.4, the reader should note the relation between the number of bins in a square and the length of the side of the triangle; this note will be important for the next example. If each bin has unit volume $V_b = 1$, then the length of each leg of the triangle ($L = 4$) must equal the volume of the square with $V_s = 4$; each square has four unit volumes and each leg has four of the large squares. As a note, all configurations (microstates) must be equally probable according to the postulate of a priori probability.

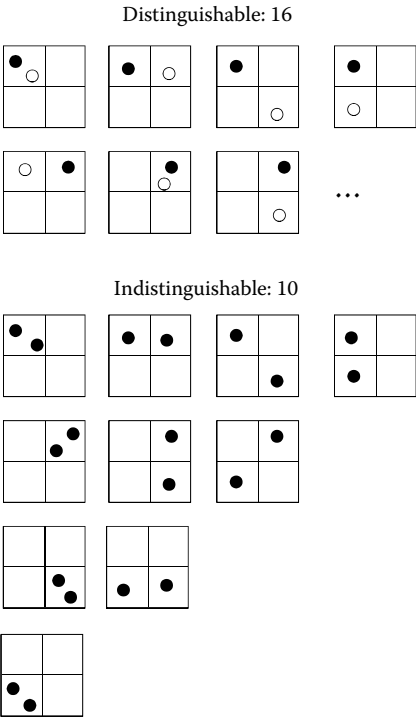


FIGURE 8.4 Counting states for distinguishable and indistinguishable particles.

Example 8.2

Calculate S as a function of the total volume V where each bin has unit volume. Assume the total volume is large. The approximate number of large squares must be the same as the area of the triangle (where each leg has length V) and each square represents a different microstate. The total number of microstates must be

$$\Omega = V^2/2$$

based on the note in the previous paragraph. Therefore

$$S = k \ln(\Omega) = 2k \ln(V) - k \ln(2)$$

If we replace the marbles with gas molecules having temperature T , we can calculate the pressure from Equation 8.5

$$\left(\frac{\partial S}{\partial V}\right)_{N,E} = \frac{P}{T}$$

The last equation provides

$$\frac{2k}{V} = \frac{P}{T} \rightarrow PV = 2kT$$

which has the correct form for the ideal gas law for two molecules.

We primarily use the ensemble to find expectation values. Suppose we want to calculate the average of a function f (such as temperature) that depends on the average speed of the molecules. We can calculate averages in two ways. First, we could watch a molecule in the system for a long time and calculate its average speed. We denote this average by $\langle v \rangle_{\text{time}}$. The other method imagines a large number of copies of the molecule (the ensemble) and calculates the average velocity observed for all of the molecules. We denote this second average by $\langle v \rangle_{\text{ensem}}$. Under certain conditions (Ergodic systems), the two averages must be the same $\langle v \rangle_{\text{time}} = \langle v \rangle_{\text{ensem}}$.

8.2.2 CANONICAL ENSEMBLE

The Canonical Ensemble consists of duplicate systems having the same number of constituent particles N , volume V , and temperature T . Consider systems of gas molecules. Unlike the micro-canonical ensemble, we do not try to make the energy remain constant. Instead, requiring the total energy to be the same among all of constituent systems, requires the translational kinetic energy of gas molecules to be the same (temperature refers to an average kinetic energy). A “thermal” reservoir can be used to control the temperature of the systems in the ensemble as indicated in Figure 8.5. A thermal reservoir is a system with a very large number of degrees of freedom and the energy states of the reservoir assume a nearly continuous range of values (i.e., finely spaced). The large number of degrees

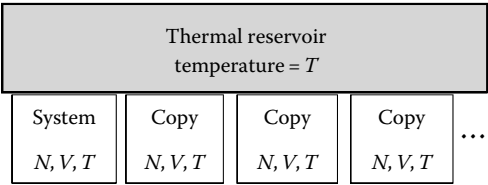


FIGURE 8.5 A thermal reservoir maintains the temperature of a system (and its copies).

of freedom means that the transfer of small amounts of energy has negligible effect on the energy distribution among states. One can say this in another way using the language of classical thermodynamics. The thermal reservoir has a very large heat capacity C so that any energy $\Delta E = C\Delta T$ transferred to a smaller system has negligible effect on the reservoir temperature.

Thermal equilibrium occurs when the temperature of the system matches the temperature of the reservoir. The temperature of the system provides a measure of the “average” kinetic energy of the gas molecules in the system. Referring to the temperature as a type of average implies that occasionally the actual total energy of the “system” (including kinetic energy) can fluctuate from its average value. We can find the probable magnitude of the fluctuation after finding the canonical probability distribution. The reader should note the use of the word “average” in connection with kinetic energy. The average can be found in two ways. The first method consists of averaging over all the systems in the ensemble. The second method views one system over a period of time that is long compared to the relaxation processes. Both methods must agree here.

The condition for thermal equilibrium (equal temperatures) can be demonstrated as follows. Suppose a reservoir makes thermal contact with a smaller system so that they can exchange energy as illustrated in Figure 8.6. Let E_r and E_s be the “total” energy in the reservoir and system, respectively. The “combined” system RS consisting of the reservoir and the little system has energy $E_r + E_s$. Thermal equilibrium occurs when the number of microstates accessible to the combined system reaches a maximum (because systems always evolve toward maximum entropy). Denote the number of microstates accessible to the small system and to the reservoir by $\Omega_s(E_s)$ and $\Omega_r(E_r)$, respectively. The total number of microstates accessible to the combined system RS when the small system has energy E_s and the reservoir has energy E_r must be given by

$$\Omega_{rs}(E_r, E_s) = \Omega_r(E_r) \Omega_s(E_s) \quad (8.6)$$

The total energy E_{rs} for the combined system must be $E_{rs} = E_r + E_s$ by energy conservation and E_{rs} must be a constant.

The entropy in Equation 8.4 for the combined system can then be written

$$S_{rs} = k \ln(\Omega_{rs}) = k \ln[\Omega_r(E_r)] + k \ln[\Omega_s(E_s)] \quad (8.7)$$

For equilibrium, one requires the full system (RS) to have maximum entropy when the energy of the small system and the reservoir have the equilibrium values $\bar{E}_s$ and $\bar{E}_r$, respectively. That is, require

$$0 = \frac{\partial S_{rs}}{\partial E_s} = k \left\{ \frac{\partial \ln[\Omega_r(E_r)]}{\partial E_r} \frac{\partial E_r}{\partial E_s} + \frac{\partial \ln[\Omega_s(E_s)]}{\partial E_s} \frac{\partial E_s}{\partial E_s} \right\}_{\substack{E_s = \bar{E}_s \\ E_r = \bar{E}_r}}$$

Using $E_r = E_{rs} - E_s$, this last expression reduces to

$$\frac{\partial \ln[\Omega_r(\bar{E}_r)]}{\partial E_r} = \frac{\partial \ln[\Omega_s(\bar{E}_s)]}{\partial E_s} \quad (8.8)$$

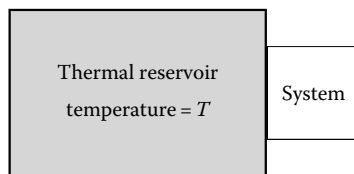


FIGURE 8.6 Reservoir in thermal contact with a system.

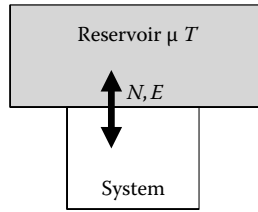


FIGURE 8.7 Reservoir exchanges particles and energy with the system for the “grand” canonical ensemble.

Recall the definition of temperature given in Equation 8.5, specifically

$$\left(\frac{\partial S}{\partial E}\right)_{N,V} = \frac{1}{T} \text{ or equivalently, } \frac{\partial \text{Ln}(\Omega)}{\partial E} = \frac{1}{kT} \equiv \beta \quad (8.9)$$

where the last equation follows from Equation 8.4 and the symbol “ k ” denotes Boltzmann’s constant. As a result, Equation 8.8 indicates that the combined system achieves equilibrium when the temperatures of the reservoir and small system agree

$$T_r = T_s$$

8.2.3 GRAND CANONICAL ENSEMBLE

The grand canonical ensemble generalizes the canonical ensemble. It consists of exact copies of a system with a specified temperature T and a specified chemical potential μ (essentially the Fermi energy). The macroscopic parameters μ, T, V specify the state of the system. The exchanges “energy” and “particles” between the system and the reservoir establishes equilibrium in this case (Figure 8.7).

8.3 THE BOLTZMANN DISTRIBUTION

The present section uses the canonical ensemble to find the Boltzmann distribution (a.k.a., the canonical distribution). The Boltzmann distribution describes how constituents (atoms or molecules for example) of the system occupy various energy levels; essentially this distribution gives the probability of a constituent occupying a given energy level. We demonstrate the Boltzmann distribution using three methods. The first method emphasizes the notion of thermal equilibrium. The second method maximizes the entropy treating the constituents as “indistinguishable” subsystems of an ensemble. The third method, most closely related to the development of the Fermi–Dirac distribution, treats the case of “distinguishable” boson-like particles. The second two methods produce identical results even though the procedure for calculating the total number of accessible states differs. We will see that only counting procedures affecting the number of particles at given energy can change the type of distribution. Subsequent sections use the third method to derive the Fermi–Dirac distribution for electrons in a semiconductor. We will see how the Fermi–Dirac distribution reduces to the Boltzmann distribution for energy differences large compared with kT .

8.3.1 PRELIMINARY DISCUSSION OF STATES AND PROBABILITY

The Boltzmann distribution describes a system in thermal equilibrium. This means that the constituents of the system occupy energy levels in accordance with the Boltzmann probability distribution. It is primarily the properties of the system (such as the type of constituent) that determines the applicable distribution. However, for equilibrium, the mathematical expressions

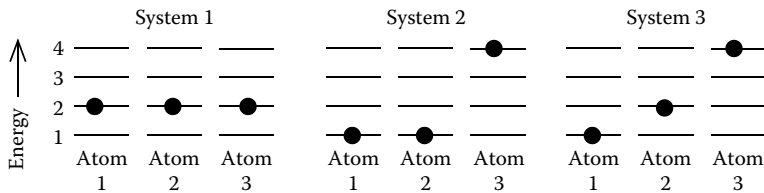


FIGURE 8.8 Systems 1, 2, 3 all consist of three atoms. The first two systems are in different microstates but have the same total energy E_s . System #3 is in a different microstate with a different energy.

obtain by bringing the system into contact with a thermal reservoir. Previous discussions show that one can find the condition for thermal equilibrium by working with the energy E_s of a system in an ensemble. The energy E_s can be viewed as the energy that a particular system has at a particular time and represents the summation of energy over all degrees of freedom. That is, an individual system has multiple constituents such as the atoms shown in Figure 8.8. Then, for E_s to be the energy of a system, the sum of the energy of all the constituents must be E_s . Alternatively said, E_s must be the total energy for a particular microstate. For example, systems 1 and 2 in the figure have energy $E_s = 6$ while system 3 has energy $E_s = 7$. The figure shows the three systems are all in different microstates but the first two systems have the same total energy. To say that the atoms have different energy can mean a variety of things. This could mean that the electrons in the atoms (or quantum wells for another example) occupy different energy levels. However, it can also mean that the atoms themselves occupy various regions of space having different potential energy (such as the discrete steps of a ladder).

Statistical mechanics makes the basic a priori postulate that all microstates with the same total energy E_s have equal probability of occurring. One can argue that microstates with greater total energy E_s must have lower probability of occurring. Therefore, systems 1 and 2 in the figure must be equally probable to occur (the microstates are equally probable). System 3 occupies a less probable microstate (since it requires more total energy). The reader should realize that a given system at temperature T can make transitions between microstates as time progresses. Also notice that if a “given” system has a “fixed” energy E (as for the microcanonical ensemble), then all “accessible” states must be equally probable (since they should all have the same energy). Usually, the microcanonical ensemble requires the energy of every system to be identically E . However, we might only know the energy to within the range $E \pm \Delta E$ (for small ΔE). In such a case, all microstates with total energy in $E \pm \Delta E$ have equal probability.

The Boltzmann distribution (also called the canonical distribution) gives the probability of a “specific” microstate with total energy E_s . We first demonstrate the Boltzmann distribution by placing a small system in contact with a large thermal reservoir. If the small system occupies a microstate with energy E_s then the reservoir can occupy any of its own compatible microstates so long as the total energy adds up to the total $E_{rs} = E_s + E_r$. The probability of finding the system in a microstate with energy E_s must then be proportional to the number of microstates available to the reservoir.

$$P(E_s) \sim \Omega_r(E_r) \tag{8.10}$$

For example, Figure 8.9 shows a very small reservoir with 3 degrees of freedom for indistinguishable particles with only one particle allowed per state. The system occupies a particular microstate with energy $E_s = 3$ and the reservoir has energy $E_r = 4$ that can occur in three ways. The following table therefore holds

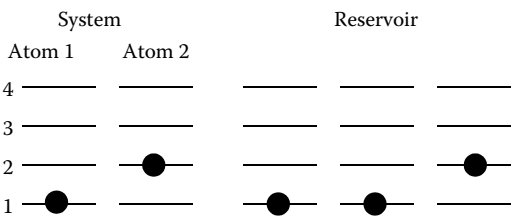


FIGURE 8.9 A (very small) reservoir has three atoms with four possible energies available to each atom. A system in thermal contact with the reservoir has two atoms, which each have four accessible levels.

System Energy (E_s)	# Reservoir Microstates	$P(E_s)$
2	6	0.6
3	3	0.3
4	1	0.1

We calculate the probability by dividing the number of reservoir microstates by the total number of accessible reservoir microstates. We have glossed over the fact that the small system can occupy two possible microstates with energy $E_s = 3$. Therefore $P(E_s = 3) = 0.3$ really means

$$P\left(\begin{matrix} \text{atom \#1} = 1 & \text{and} & \text{atom \#2} = 2 \\ \text{or} \\ \text{atom \#1} = 2 & \text{and} & \text{atom \#2} = 1 \end{matrix}\right) = 0.3$$

Figure 8.9 can be rearranged to show the states of the system (i.e., each energy E_s); each individual level (i.e., small line) represents a microstate. For example, there are two microstates in Figure 8.9 that have energy $E_s = 3$ namely {atom #1 = 1 and atom #2 = 2} and {atom #1 = 2 and atom #2 = 1}. Now we represent the entire system (consisting of two atoms) by a single dot as shown in Figure 8.10, which gives more meaning to the phrase “the system occupies a microstate.” For example, atoms #1 and #2 in states #1 and #2, respectively produce the single dot at $E_s = 3$ in Figure 8.10. Each small horizontal line represents one of the system microstates. Each microstate comes from a combination of atomic eigenstates (from Figure 8.9). For Figure 8.10, there exist four “degenerate” microstates

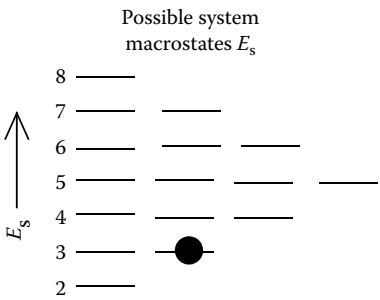


FIGURE 8.10 Available states for the system given the energy E_s .

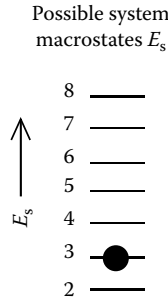


FIGURE 8.11 A nondegenerate system has only one state for each energy.

with energy $E_s = 5$ (sometimes the word “micro” is dropped). As we will see, we can include this energy degeneracy in our probability scheme by using a density of states function $g(E)$. We handle the degeneracy as a special procedure so that the Boltzmann distribution only accounts for temperature effects. For now, assume the system has “nondegenerate” microstates whereby the system has only one state with energy E_s as represented in Figure 8.11. A “macrostate” of the system only tracks the total energy and not which microstates have that energy.

8.3.2 DERIVATION OF BOLTZMANN DISTRIBUTION USING A THERMAL RESERVOIR

One can find an expression for the probability distribution of indistinguishable classical particles. Let E_s , E_r , and E_{rs} be the energy of the system, reservoir, and reservoir plus system ($E_{rs} = E_r + E_s$). As in Equation 8.4, define the entropy S by

$$S = k \text{Ln}(\Omega) \quad (8.11)$$

where Ω represents the number of distinguishable ways the constituents of a system can occupy the available microstates, and k denotes the Boltzmann constant (see examples below). The number of microstates in Equation 8.11 can be rewritten using a Taylor expansion in E_s (since $E_s \ll E_{rs}$)

$$\text{Ln}[\Omega_r(E_r)] = \text{Ln}[\Omega_r(E_{rs} - E_s)] \cong \text{Ln}[\Omega_r(E_{rs})] + \left. \frac{\partial \text{Ln} \Omega_r(E)}{\partial E} \right|_{E=E_{rs}} (-E_s) \quad (8.12)$$

Recalling the definition for temperature from Equation 8.5, specifically

$$\beta = \frac{1}{kT} = \left(\frac{\partial \text{Ln}(\Omega)}{\partial E} \right)_{N,V} \quad (8.13)$$

Equation 8.12 can be rewritten as

$$\text{Ln}[\Omega_r(E_r)] = \text{Ln}[\Omega_r(E_{rs})] - \beta E_s \quad (8.14)$$

Therefore, by taking the exponential of Equation 8.14 and by defining $C = \Omega_r(E_{rs})$,

$$\Omega_r(E_r) \sim C e^{-\beta E_s} = C \exp\left(-\frac{E_s}{kT}\right) \quad (8.15)$$

The probability that the small system occupies a single microstate with energy E_s can be found from Equations 8.10 and 8.15

$$P(E_s) = \frac{1}{Z} \exp\left(-\frac{E_s}{kT}\right)$$

where the partition function Z provides a normalization factor so that the sum of all probabilities comes out to 1. We can find Z by requiring the probability to add to one

$$1 = \sum_s P(E_s) = \frac{1}{Z} \sum_s \exp\left(-\frac{E_s}{kT}\right) \rightarrow Z = \sum_s \exp\left(-\frac{E_s}{kT}\right) \quad (8.16)$$

where the partition function Z sums over all accessible microstates of the system regardless of energy. For nondegenerate levels, the sum over “ s ” covers the different energy values. The probability $P(E_s)$ refers to the probability of that particular value of E_s . For degenerate levels, the sum extends over all states including the multiple states with the same energy. $P(E_s)$ refers to one particular microstate (out of the many degenerate states) that has energy E_s . The probability of finding the system in *any* state with energy E_s must account for a summation over all the microstates with energy E_s . For now, we continue with the nondegenerate case and handle the degenerate one with density of states functions as in Section 8.4.3.

8.3.3 DERIVATION OF BOLTZMANN DISTRIBUTION USING AN ENSEMBLE

This section uses another approach to find the Boltzmann distribution for indistinguishable classical systems. A large number $\mathcal{N}$ of systems comprises an ensemble with total energy $\mathcal{E}$ (the sum of the energy of all systems in the ensemble). The systems can exchange energy with each other and therefore the energy of each system can change. Assume that $\{E_s\}$ is the set of possible energy levels that any given system can occupy. Further assume that n_s refers to the number of systems (at a particular time) with a particular energy E_s . Let $\mathcal{U}$ be the average energy over all of the systems. The following constraints must be satisfied for all times

$$\sum_s n_s = \mathcal{N} \quad (8.17a)$$

and

$$\sum_s n_s E_s = \mathcal{E} = \mathcal{N} \mathcal{U} \quad (8.17b)$$

The set of numbers $(n_1, n_2, \dots)$ specifies a “mode” of the distribution.

Example 8.3

Figure 8.12 shows an example of $\mathcal{N} = 6$ systems (denoted by S1–S6) in an ensemble where $n_1 = 3$, $n_2 = 1$, $n_3 = 2$ of the systems have energy E_1 , E_2 , and E_3 , respectively. The total energy in the ensemble must be $\mathcal{E} = 3E_1 + E_2 + 2E_3$ and the average energy is $\mathcal{U} = \mathcal{E}/\mathcal{N} = 0.5E_1 + 0.17E_2 + 0.33E_3$. The mode of the distribution is $(n_1, n_2, n_3) = (3, 1, 2)$.

The most probable state of the system must occur most often in the ensemble (with the limitations imposed by Equations 8.17). The probability of finding a system with energy E_1 can be written as

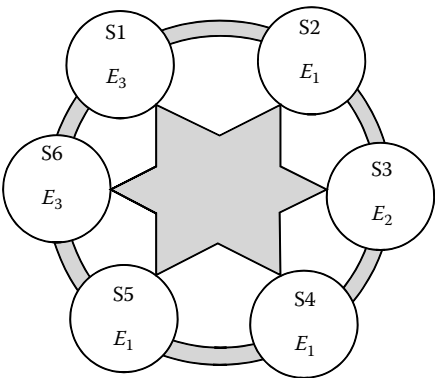


FIGURE 8.12 Systems S_1 through S_6 in the ensemble can exchange energy (though the star-shaped object). The total energy is $E = 3E_1 + E_2 + 2E_3$.

$$P(E_1) = \frac{\langle n_1 \rangle}{\mathcal{N}}$$

where the average value $\langle n_1 \rangle$ appears in the formula because for a finitely sized ensemble, the number of systems with energy E_s might slightly fluctuate in time. In general, the probability of finding a system with energy E_s must be

$$P(E_s) = \frac{\langle n_s \rangle}{\mathcal{N}} \tag{8.18}$$

The ensemble of systems will evolve to maximum entropy, which occurs when the ensemble has access to the largest number of states. The function $\Omega\{n_s\} = \Omega(n_1, n_2, \dots)$

$$\Omega = \frac{\mathcal{N}!}{n_1!n_2!\dots} \tag{8.19}$$

provides the number of states accessible to the ensemble. As reviewed in Appendix G, this formula gives the number of ways to take n_1 items of type 1, n_2 items of type 2, ... from $\mathcal{N}$ total items. The n_1 objects have energy E_1 , and n_2 have energy E_2 , and so on. There must be a set of numbers $(\bar{n}_1, \bar{n}_2, \bar{n}_3, \dots)$ that maximize Ω . The relation $\langle n_s \rangle = \bar{n}_s$ must hold. Equation 8.18 then provides the probability distribution. We assume that each individual system has only one microstate with total energy E_s (nondegenerate case). Multiple systems can be in the state E_s since each system can take on the value E_s . If one system has degenerate microstates then all of the systems in the ensemble have degenerate microstates (definition of ensemble). In such a case, we will most likely find the systems in the highly degenerate states simply because the system has so many such microstates and not solely because the system has a certain temperature. The next section discusses the degenerate case.

Example 8.4

For $\mathcal{N} = 4$ balls, find the number of different arrangements of $n_0 = 2$ orange and $n_1 = 2$ indigo balls. Equation 8.19 provides a value of 6. The arrangements are
ooii oioi ioio iioo iioo

This example could be restated to read as follows. Find the number of ways four atoms can be arranged so that two have one energy value and the other two have a second energy value.

The Boltzmann distribution can be found by maximizing Equation 8.19 while including the constraints in Equations 8.17. Note that because a natural logarithm Ln monotonically increases with

its argument, we can maximize either the argument or the Ln and get the same results. The constraints can be included by using Lagrange multipliers (refer to Appendix H). First we need to maximize Equation 8.19 without the constraints. Sterling's formula for very large numbers " n " provides

$$\text{Ln}(n!) \cong n \text{Ln}(n) - n \quad (8.20)$$

We find

$$\text{Ln}(\Omega) = \text{Ln}(\mathcal{N}!) - \sum_s \text{Ln}(n_s!) = \{\mathcal{N}\text{Ln}(\mathcal{N}) - \mathcal{N}\} - \sum_s \{n_s \text{Ln}(n_s) - n_s\}$$

Applying $\mathcal{N} = \sum_s n_s$ to the very last term in the previous equation provides

$$\text{Ln} \Omega = \mathcal{N} \text{Ln}(\mathcal{N}) - \sum_s n_s \text{Ln}(n_s)$$

The change in $\text{Ln}(\Omega)$ due to a change in n_s , denoted by δn_s , can be written as

$$\delta \text{Ln}(\Omega) = - \sum_s [\delta n_s \text{Ln}(n_s) + \delta n_s] = - \sum_s [\text{Ln}(n_s) + 1] \delta n_s \quad (8.21)$$

where $\mathcal{N}$ is constant. We want to set $\delta \text{Ln}(\Omega) = 0$. If the variations δn_s were independent then we could conclude that the expansion coefficient in Equation 8.21, namely $[\text{Ln}(n_s) + 1]$, must be equal to zero. However, the variations δn_s are not independent because of the restrictions in Equations 8.17.

The method of Lagrange multipliers allows us to set the expansion coefficients to zero by the following method. The variation of Equations 8.17 gives

$$\sum_s \delta n_s = 0 \quad \sum_s E_s \delta n_s = 0$$

The technique multiplies the sums by constants α, β

$$\alpha \sum_s \delta n_s = 0 \quad \beta \sum_s E_s \delta n_s = 0$$

and adds the resulting sums to Equation 8.21 to get

$$0 = \delta \text{Ln}(\Omega) = \sum_s - [\text{Ln}(\bar{n}_s) + 1 + (\alpha + \beta E_s)] \delta n_s$$

Choosing α, β so that the arguments (i.e., the terms in $[\]$ brackets) are zero provides

$$\text{Ln}(\bar{n}_s) = -(\alpha + 1) - \beta E_s$$

Therefore, taking the exponential and setting $C = e^{\alpha+1}$, we get

$$\bar{n}_s = \frac{1}{C} e^{-\beta E_s} \quad (8.22)$$

The value of the constant C in this last equation can be found using $\sum_s \bar{n}_s = \mathcal{N}$ to get

$$C \mathcal{N} = C \sum_s \bar{n}_s = \sum_s e^{-\beta E_s} \quad (8.23)$$

Therefore the Boltzmann probability of finding the state E_s follows from Equations 8.18, 8.22, and 8.23 to give

$$P(E_s) = \frac{e^{-\beta E_s}}{\sum_s e^{-\beta E_s}} \quad (8.24)$$

It turns out that $\beta = (1/kT)$. The partition function $Z = \sum_s e^{-\beta E_s}$ is important since the thermodynamic quantities can be obtained from it.

Example 8.5

Find the effective temperatures for the system given in the previous example if $E_1 = 1$, $E_2 = 2$, $E_3 = 3$ and $n_1 = 3$, $n_2 = 1$, $n_3 = 2$. Note the energy level #3 has a population inversion with respect to level #2 since more atoms occupy the larger energy E_3 . As a result, the system is not in thermal equilibrium.

SOLUTION

Equations 8.22 and 8.23 provides the ratios

$$\begin{aligned} \frac{P(E_1)}{P(E_2)} &= \frac{e^{-\beta E_1}}{e^{-\beta E_2}} = \frac{\bar{n}_1/\mathcal{N}}{\bar{n}_2/\mathcal{N}} = 3 \rightarrow \beta = 1.09 \rightarrow T = \frac{1}{1.09k} \\ \frac{P(E_2)}{P(E_3)} &= \frac{e^{-\beta E_2}}{e^{-\beta E_3}} = \frac{\bar{n}_2/\mathcal{N}}{\bar{n}_3/\mathcal{N}} = 0.5 \rightarrow \beta = -0.69 \rightarrow T = \frac{-1}{0.69k} \end{aligned}$$

For this problem, we cannot find a single temperature that characterizes the distribution. Notice also that population inversions produce “negative” temperatures. Lasers always need population inversions to produce stimulated emission. This means that greater numbers of electrons must occupy the conduction band than required for thermal equilibrium. One sometimes hears that population inversions occur for high temperatures but this is untrue and infact requires a negative temperature when requiring a Boltzman distribution to apply.

8.3.4 COUNTING DEGENERATE STATES

Now consider degenerate microstates that have the same energy as illustrated in Figure 8.13. The microstates have been labeled as “s,j” for convenience where “s” represents the energy (vertical scale) and “j” represents one of the states with energy E_s (i.e., one of the lines along the horizontal). All but levels #2 and #8 are degenerate. Let $g(E_s)$ denote the number of states for energy level E_s . We start with the partition function.

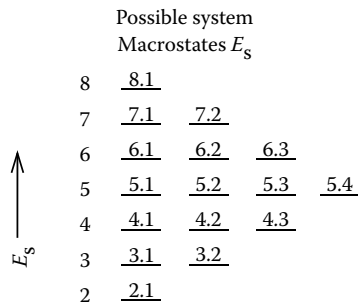


FIGURE 8.13 Microstates with the same energy are degenerate. Each little line represents a microstate.

$$Z = \sum_{\substack{\text{states} \\ s}} \exp(-\beta E_s) \quad (8.25a)$$

The summation includes all 16 states shown in Figure 8.13; that is, “s” refers to the microstates and not just the energy values. We expand the summation

$$Z = \underbrace{\exp(-\beta E_{2,1})}_{g(E_2)=1} + \underbrace{\exp(-\beta E_{3,1}) + \exp(-\beta E_{3,2})}_{g(E_3)=2} + \underbrace{\exp(-\beta E_{4,1}) + \cdots + \exp(-\beta E_{4,3})}_{g(E_4)=3} + \cdots + \quad (8.25b)$$

Therefore, rather than include all of the microstates in the summation for the partition function, we can instead write

$$Z = \sum_E g(E) \exp(-\beta E) \quad (8.25c)$$

Applying similar reasoning to the numerator of Equation 8.24, we can write the probability of the system being in any of the microstates with energy E as

$$P(E) = \frac{g(E) \exp(-\beta E)}{\sum_E g(E) \exp(-\beta E)} \quad (8.26)$$

8.3.5 BOLTZMANN DISTRIBUTION FOR DISTINGUISHABLE BOSON-LIKE PARTICLES

This section uses another approach to find the Boltzmann distribution for distinguishable classical particles. We assume these classical particles behave similar to bosons for which any number of them can occupy the same state. We first discuss the implications of the particles being distinguishable and bosonic in character. Next we derive an expression for the entropy by determining the number of possible ways to arrange $\mathcal{N}$ particles in the accessible states. For this case, we want the average number of particles per available state rather than a probability.

We want to find an expression for the entropy in terms of the number of different states accessible to the system. Each system state corresponds to a different arrangement of the $\mathcal{N}$ particles in the states of the system. Assume n_i represents the number of particles in g_i states with energy E_i . Two constraints must be applied to the system. First we assume the number of particles does not change so that

$$\mathcal{N} = \sum_{i=1}^{N_L} n_i \quad (8.27a)$$

and we assume a fixed amount of energy in the system

$$\mathcal{E} = \sum_{i=1}^{N_L} E_i n_i \quad (8.27b)$$

where N_L represents the total number of energy levels in the system.

Example 8.6

List the various parameters for the system shown in Figure 8.14.

SOLUTION

The system has $N_L = 4$ levels labeled $i = 1, 2, 3, 4$. The energy takes on the values

$$E_1 = 0.2, \quad E_2 = 0.3, \quad E_3 = 0.4, \quad E_4 = 0.5$$

The degeneracy of the levels can be written as $g_1 = 1, g_2 = 2, g_3 = 3, g_4 = 3$. The levels have $n_1 = 1, n_2 = 2, n_3 = 1, n_4 = 2$ which gives the total number of particles $\mathcal{N} = \sum_{i=1}^{N_L} n_i = 6$. The total energy for the system must be $\mathcal{E} = \sum_{i=1}^{N_L} E_i n_i = 2.2$.

We want to vary the number of particles in level E_i to find the arrangement (i.e., find n_1, n_2 , etc.) that gives the largest number of possible different arrangements of particles. For example, if all particles were confined to exactly one arrangement at the lowest levels, then we would have very low entropy; in fact, $\Omega = 1$ so that the entropy would be $S = k \ln(\Omega) = 0$. On the other hand, if we have not any particles in a level then the entropy for that level must also be zero. Therefore, some number of particles must produce a nonzero entropy and there must therefore also be some number of particles giving a maximum of the entropy. For example, Figure 8.15 shows an example for the total number of arrangements for a two-level system and how the constraint on the total number affects the calculation (the total number must be directly above the straight line). The number of accessible states depends on the total-number constraint.

Next, examine the implication of having distinguishable particles. Essentially, we claim to be able to keep track of separate particles (maybe molecules) and always be able to tell the difference between them. In this case, for every arrangement, there must always be another, different one found

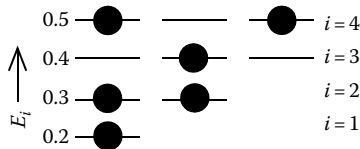


FIGURE 8.14 A particular microstate of the system.

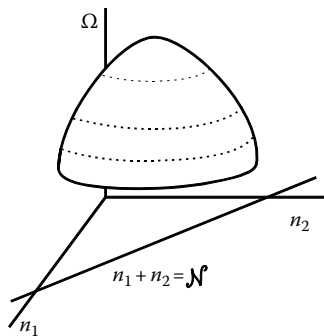


FIGURE 8.15 The number of accessible states for two levels.



FIGURE 8.16 Switching distinguishable particles between states in the same level gives another arrangement.

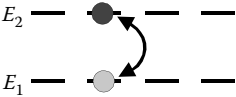


FIGURE 8.17 Switching distinguishable particles between levels produces a new arrangement.

by just switching the position of two of the particles. These must be counted as separate arrangements. For example, Figures 8.16 and 8.17 show two sets of two different particles; one set has the particles in the same energy level and the other set has the particles in separate energy levels. In both cases, the switch gives a new arrangement and must be included in the counting procedure.

For this section, the particles also have boson-like properties whereby more than one particle can reside in a state at the same time as shown in Figure 8.18. Both properties (distinguishable and boson-like) must be taken into account during the counting procedure.

Now count the number of states. Perhaps a simple example would provide the easiest entrance to deduce the most general formula while gaining some valuable insight. Consider two levels $N_L = 2$ with two degenerate states $g_1 = 2$ and $g_2 = 2$ in each level. Assume three distinguishable particles according to #1 = Blue, #2 = Green, and #3 = Red. Further assume level 2 has 1 particle (say Blue) $n_2 = 1$ and level 1 has 2 particles (Green and Red) as shown in Figure 8.19. As an easy learning tool, it might help the reader to label the levels and the particle colors in the figure.

First find the number of arrangements of the particles #2 and #3 in the lower level. The particles can occupy the same or different states. So we can assign either state to particle #1 or state to particle #2. We might think of dropping the states into a bucket with each bucket corresponding to particle #1 or #2 as schematically illustrated in Figure 8.20. This means the number of arrangements for level #1 can be written as

$$\Omega_1 = 4 = 2 \cdot 2 = g_1^2 \tag{8.28a}$$

By the way, this number includes interchanges between particles in Figure 8.19 (i.e., switching 2 and 3). Next particle #1 can occupy either of two states so that

$$\Omega_2 = 2 = g_2^1 \tag{8.28b}$$



FIGURE 8.18 Boson-like particles in the same state.

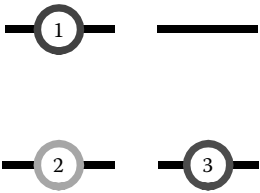


FIGURE 8.19 Initial configuration for three distinguishable particles in two levels.

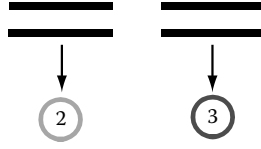


FIGURE 8.20 Either of two states can be assigned to either particle.

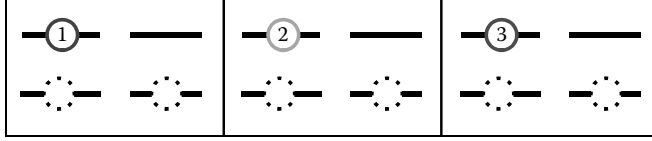


FIGURE 8.21 Switching the distinguishable particles into the upper level leads to three times the number of arrangements as without the switching.

We can multiply these together to get a total number but we still must consider the effect of switching particles between levels. Figure 8.21 shows that the three particles lead to three different arrangements. Particles #1, #2, or #3 can occupy the upper level. As a result, we must multiply the total number of arrangements by 3. Notice that rotating the particles in this manner does not change the numbers n_1 and n_2 . The total number of arrangements becomes

$$\Omega = 3\Omega_1\Omega_2 = 24$$

The previous relation can be rewritten as follows

$$\Omega = 3\Omega_1\Omega_2 = 3g_1^2g_2^1 = 3 \cdot 2 \cdot 1 \frac{g_1^2}{2} \frac{g_2^1}{1} \rightarrow \Omega = \mathcal{N}! \frac{g_1^{n_1}}{n_1!} \frac{g_2^{n_2}}{n_2!}$$

With some thought, one can define the most general formula for arranging $\mathcal{N}$ total objects (all distinguishable) with n_i in N_L levels with g_i degenerate states in each level.

$$\Omega = \mathcal{N}! \prod_{i=1}^{N_L} \frac{g_i^{n_i}}{n_i!} \quad (8.29)$$

Notice that one should expect the form of the distribution derived from Equation 8.29 to be approximately the same as that derived from Equation 8.19, reprinted below

$$W = \frac{\mathcal{N}!}{n_1!n_2!\dots}$$

One can see the similarity by assuming all the g_i are approximately equal. In such a case, the two formulas agree except for the term

$$\prod_i g_i^{n_i} \sim g_i^{n_1+n_2+\dots} = g_i^{\mathcal{N}} \quad (8.30)$$

in Equation 8.29. However, the derivatives in the maximization procedure with dn_i as the variation variables will map the constant terms $g_i^{\mathcal{N}}$ to zero. The extra terms in 8.29 have not any affect on the

form of the resulting distribution. Therefore the form of the distribution must only be sensitive to those terms affecting the number of particles in a given energy level.

Now maximize the entropy (subject to two constraints) to find the Boltzmann distribution corresponding to thermal equilibrium for the distinguishable boson-like particles.

$$S = k \ln(\Omega) = k \ln \left(N! \prod_{i=1}^{N_L} \frac{g_i^{n_i}}{n_i!} \right) = k \left\{ \ln(N!) + \sum_{i=1}^{N_L} [n_i \ln(g_i) - \ln(n_i!)] \right\} \quad (8.31a)$$

$$N = \sum_{i=1}^{N_L} n_i \quad \mathcal{E} = \sum_{i=1}^{N_L} E_i n_i \quad (8.31b)$$

The method of Lagrange multipliers found in Appendix H defines a new function $\mathcal{S}$ incorporating both the entropy and the constraints

$$\mathcal{S} = k \left\{ \ln(N!) + \sum_{i=1}^{N_L} [n_i \ln(g_i) - \ln(n_i!)] \right\} - \lambda_1 \sum_{i=1}^{N_L} n_i - \lambda_2 \sum_{i=1}^{N_L} E_i n_i \quad (8.32)$$

The new function does not have constraints and the variations dn_i must therefore be independent of each other.

In order to differentiate Equation 8.32, we should simplify the term $n_i!$ using the Sterling approximation.

$$\ln(n!) = n \ln(n) - n$$

Substituting the Sterling approximation into the new function in Equation 8.32 produces

$$\mathcal{S} = k \left\{ \ln(N!) + \sum_{i=1}^{N_L} [n_i \ln(g_i) - n_i \ln(n_i) + n_i] \right\} - \lambda_1 \sum_{i=1}^{N_L} n_i - \lambda_2 \sum_{i=1}^{N_L} E_i n_i \quad (8.33)$$

We maximize the new function $\mathcal{S}$ in Equation 8.33. Setting the differential to zero produces

$$0 = d\mathcal{S} = k \sum_{i=1}^{N_L} [dn_i \ln(g_i) - dn_i \ln(n_i) - dn_i + dn_i] - \lambda_1 \sum_{i=1}^{N_L} dn_i - \lambda_2 \sum_{i=1}^{N_L} E_i dn_i$$

Canceling terms and factoring out the variations dn_i , we find

$$0 = d\mathcal{S} = \sum_{i=1}^{N_L} [k \ln(g_i) - k \ln(n_i) - \lambda_1 - \lambda_2 E_i] dn_i$$

Because the variations dn_i can be chosen independently, the coefficients must be zero.

$$k \ln(g_i) - k \ln(n_i) - \lambda_1 - \lambda_2 E_i = 0$$

This last equation can be rearranged to produce the Boltzmann distribution

$$F_B(E_i) = \frac{n_i}{g_i} = \exp \left(-\frac{\lambda_1 + \lambda_2 E_i}{k} \right) \quad (8.34a)$$

Substituting this last expression into the constraints gives the values of the two undetermined constants. We would find (with some effort) that $\lambda_2 = 1/T$.

$$F_B(E_i) = \frac{n_i}{g_i} = C \exp\left(-\frac{E_i}{kT}\right) \quad (8.34b)$$

where $C = \text{Exp}(-\lambda_1/k)$. We still need to determine the remaining constant C . If we use $C = 1$ then this last equation gives the number of particles per state which for boson-like particles cannot be interpreted as a probability (not normalized). The Boltzmann distribution has the form

$$F_B(E) = \exp\left(-\frac{E}{kT}\right) \quad (8.34c)$$

where the superfluous index “ i ” has been dropped. Equation 8.34c does not exactly represent a probability because any number of particles can occupy a state according to the counting methods. Later sections will show how the Boltzmann distribution corresponds to the Fermi–Dirac distribution for electrons (as Fermions) when E is large compared with the Fermi level.

Next we determine the Boltzmann probability distribution. We return to Equation 8.34b and substitute it into the constraint Equation 8.31b

$$N = \sum_{i=1}^{N_L} n_i$$

to find the constant C .

$$C = \frac{N}{Z} \quad \text{where} \quad Z = \sum_{i=1}^{N_L} g_i \exp\left(-\frac{E_i}{kT}\right) \quad (8.35)$$

which defines the so-called partition function Z . Equation 8.34b becomes

$$P_B(n_i) = \frac{n_i}{N} = \frac{g_i}{Z} \exp\left(-\frac{E_i}{kT}\right) \quad (8.36)$$

This last equation defines a probability with Z being the normalization so that the probabilities add to 1. Notice especially the E_i refers to an entire level and not just one microstate in the level.

8.3.6 INDEPENDENT, DISTINGUISHABLE SUBSYSTEMS

When two subsystems cannot exchange energy, they must be independent systems. The subsystem can be as small as a single atom. Assume the subsystems consist of a single atom. For example, two atoms each confined to separate Styrofoam containers, would be thermally noninteracting. The subsystems remain independent so long as they do not interact with each other. If each independent atom has Hamiltonian $\hat{H}_a$ then the total Hamiltonian for the total system can be written as

$$\hat{H} = \sum_a \hat{H}_a$$

Each Hamiltonian $\hat{H}_a$ operates on its own Hilbert space. As opposed to independent subsystems, interacting ones have terms in the full Hamiltonian that link the Hilbert spaces. In this section,

we work solely with the independent subsystems. If the eigenvalues for subsystem #a are $\{\varepsilon_1^{(a)}, \varepsilon_2^{(a)}, \dots\}$ then, because the subsystems are identical in structure, the other subsystems must have the same set. For example, atom #b has eigenvalues $\{\varepsilon_1^{(b)}, \varepsilon_2^{(b)}, \dots\}$ with $\varepsilon_i^{(a)} = \varepsilon_j^{(b)}$. For convenience, we assume the subsystems states are not degenerate so that $\varepsilon_i^{(a)} \neq \varepsilon_j^{(a)}$ so long as $i \neq j$. The probability that atom #1 occupies state $\varepsilon_i^{(1)}$, atom #2 occupies state $\varepsilon_j^{(2)}$ (and so on) can be written as

$$P(\varepsilon_i^{(1)}, \varepsilon_j^{(2)}, \dots) = \frac{e^{-\beta E_s}}{\sum_s e^{-\beta E_s}} = \frac{\exp[-\beta(\varepsilon_i^{(1)} + \varepsilon_j^{(2)} + \dots)]}{\sum_{i,j,\dots} \exp[-\beta(\varepsilon_i^{(1)} + \varepsilon_j^{(2)} + \dots)]}$$

The partition function Z for the whole system can be written as

$$Z = \prod_a Z_a = \prod_a \left[\sum_i \exp(-\beta \varepsilon_i^{(a)}) \right]$$

where Z_a is the partition function for subsystem #a. Therefore, the probability can be written as

$$P(\varepsilon_i^{(1)}, \varepsilon_j^{(2)}, \dots) = \frac{\exp(-\beta \varepsilon_i^{(1)})}{Z_1} \frac{\exp(-\beta \varepsilon_j^{(2)})}{Z_2} \dots = P(\varepsilon_i^{(1)}) P(\varepsilon_j^{(2)}) \dots$$

Therefore, the probability of a particular configuration for the whole system must be the same as the product of probabilities for the subsystems.

Consider another question regarding independent subsystems. What is the probability that subsystem #1 occupies state $\varepsilon_i^{(1)}$ regardless of the state occupied by the others? This answer can be found by summing over the states for the other atoms.

$$\sum_{j,k,\dots} P(\varepsilon_i^{(1)}, \varepsilon_j^{(2)}, \dots) = \frac{\exp(-\beta \varepsilon_i^{(1)})}{Z_1} \sum_j \frac{\exp(-\beta \varepsilon_j^{(2)})}{Z_2} \sum_c \frac{\exp(-\beta \varepsilon_k^{(2)})}{Z_3} \dots = P(\varepsilon_i^{(1)})$$

8.4 INTRODUCTION TO FERMI-DIRAC DISTRIBUTION

To successfully design and operate electronic devices, it is necessary to focus attention on the behavior of the electrons and holes for producing current and signals. Many contemporary devices operate by controlling the flow of current. Applying electric fields to bend the bands whether for a pn junction or the field effect can control the number of electrons or holes in a band. One therefore needs to understand the bands and the band states along with the method of populating those bands with mobile charge. The band states correspond to the Bloch plane waves. The density of states function $g(E)$ describes the number of states at energy E . The average number of electrons in the conduction band can be determined from the Fermi-Dirac distribution (Fermi function) that describes the probability of electrons occupying states as a function of the energy. Fermi energy (approximately, the chemical potential) for thermal equilibrium situations and the quasi-Fermi levels for nonequilibrium situations characterize the Fermi function. This section discusses the form and role of the Fermi function in solid state electronics while the next section derives the functional form by maximizing the entropy.

8.4.1 FERMİ–DIRAC DISTRIBUTION

The Fermi–Dirac distribution $F(E)$ gives the probability of an electron occupying a given state with energy E . In actuality, the distribution function describes the number of particles occupying a given state. However, with at most one electron per state, the Fermi function can be given the probability interpretation. The Fermi function for electrons has the form

$$F(E) = \frac{1}{e^{\frac{E-E_F}{kT}} + 1} \quad (8.37)$$

where

E_F represents the Fermi energy

k denotes the Boltzmann constant

T denotes the temperature in degree Kelvin

Figure 8.22 shows the Fermi distribution as a function of energy and parameterized by temperature. Those states located more than a few kT from the Fermi energy most likely remain empty whereas those more than a few kT below the Fermi level remain filled. The Fermi level E_F represents the energy of those states with 50% chance of being filled or empty. We can easily see this by setting $E = E_F$ in Equation 8.37 for then $F(E) = 0.5$. The a priori probability inherent to the Fermi function does not depend on there being states available at energy E . The Fermi function measures the average number of electrons that would occupy a state at energy E if the state exists.

Sometimes people say that all of the states below E_F must be occupied and those above must empty. This approximation can sometimes be useful. However, intrinsic semiconductors (i.e., undoped) would not conduct current if it were true. We need electrons in the conduction band (and holes in the valence band) if we expect to see any current flow. However, looking at the $T = 0$ plot shows that, in fact, all states below E_F must be filled while those above E_F must be empty. At $T = 0$, the Fermi function tells us that the lattice does not contain enough thermal energy (phonons) to transition electrons from one state to another (for example, from the valence to conduction band).

For states with energy slightly larger (a few kT) than the Fermi level E_F , we can approximate the Fermi function (Fermi–Dirac distribution) by the Boltzmann distribution.

$$F(E) = \frac{1}{e^{\frac{E-E_F}{kT}} + 1} \cong e^{-\frac{E-E_F}{kT}} \quad (8.38)$$

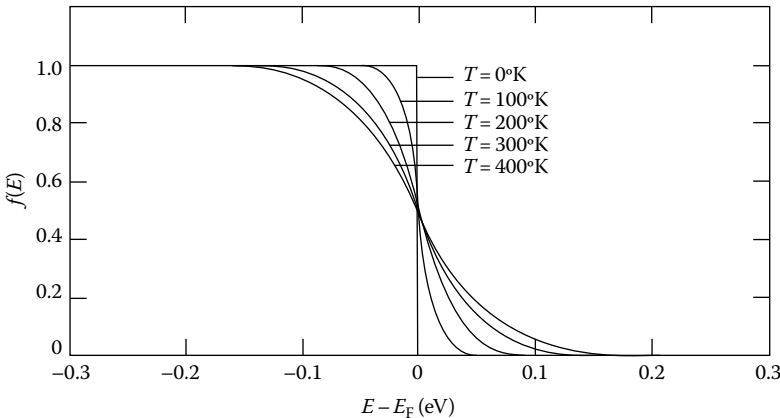


FIGURE 8.22 The Fermi function for various temperatures.

which is then essentially a Boltzmann distribution. However, for $E < E_F$, it is inaccurate since it predicts more than one electron per state for some E (and T).

The Fermi function $F_e = F(E)$ gives the probability of an electron occupying a state in either the conduction or valence band. We can alternatively discuss the probability of a hole occupying a state. This must be the probability of finding an empty state. Therefore, the probability of a hole occupying a state must be $F_h = 1 - F(E)$.

The value of the Fermi function ranges from 0 to 1 for each state in a band (due to the Pauli exclusion principle) and it can therefore represent a probability. However, typically, a distribution function represents the number of particles per state so that the sum over all energies for the valence and conduction band comes out to N the number of states in one of the bands. We can see this by considering the case of $T=0$. In this case, $F(E)=1$ for the valence band and $F(E)=0$ for the conduction band. If there are N states in the valence band (N represents the number of primitive cells in the crystal) then we must have

$$\sum_{vb} F(E) = N F(E) = N$$

Essentially, the Fermi–Dirac distribution gives the average number of electrons occupying a state at energy E . Because the total number of electrons remains at N for temperatures other than 0 K (one of the constraints used in the derivation in Section 8.3), the total sum over $F(E)$ must still produce N .

8.4.2 DENSITY OF CARRIERS

The Fermi function and the density of states determine the density of electrons in the conduction band and holes in the valence band. Additionally, they determine the occupancy of traps under conditions of thermal equilibrium. The density of electrons per unit energy n_E in the conduction band can be found using simple dimensional analysis as follows

$$n_E = \frac{\#e^-}{\text{energy} \cdot \text{vol}} = \frac{\# \text{ states}}{\text{energy} \cdot \text{vol}} * \frac{\#e^-}{\text{state}} = g_e(E) F_e(E)$$

where g_e represents the density of states. The number of electrons Δn in the interval dE must be

$$\Delta n = n_E dE = g_e(E) F_e(E) dE \rightarrow n = \int_{E_c}^{\infty} g_e(E) F_e(E) dE \quad (8.39a)$$

where the symbol n represents the total number of electrons (per unit volume) in the conduction band. A similar expression must hold for holes in the valence band.

$$p = \int_{-\infty}^{E_v} g_h(E) F_h(E) dE \quad (8.39b)$$

where $F_h = 1 - F_e$.

We need both the Fermi function $F(E)$ and the density of states $g(E)$ to find the number of carriers in a band. The Fermi functions appear in Figure 8.23. Many devices operate near room temperature of $T = 300$ K. Most of the carriers occupy states within a few kT of the band edges. At room temperature, kT represents about 25 meV (approximately 1/40 eV)—a very small energy compared with the 1 eV (or more) band gap of many semiconductors. Therefore, we realize the carrier distributions accumulate at the band edges (i.e., the minimum of the CB and maximum of the VB). Assuming sufficient energy separation between the band edge and the Fermi energy, the electron Fermi function can be approximated by the Boltzmann distribution

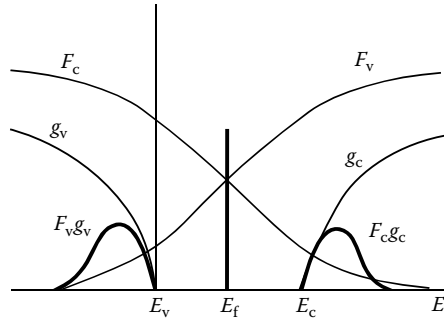


FIGURE 8.23 The carrier distribution $F(E)g(E)$ for each band.

$$F_c(E) = e^{-\frac{E-E_f}{kT}} = e^{-\frac{E_c-E_f}{kT}} e^{-\frac{E-E_c}{kT}} = \text{Const} * e^{-\frac{E-E_c}{kT}} \quad (8.40a)$$

For energy larger than the conduction band edge, the probability of a state being occupied exponentially decreases with energy. Similar comments apply to the Fermi function for holes near the valence band where we have

$$F_h(E) = \text{Const} * e^{-\frac{E_v-E}{kT}} \quad (8.40b)$$

Finally, we need the density of states to complete the calculation of the carrier density in Equations 8.39. Recall that the density of state functions (Chapter 7) for the conduction and valence band increase as the square root of the energy from the band edge

$$g_{\text{elect}}(E) = \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \sqrt{E - E_c} \quad g_{\text{hole}}(E) = \frac{1}{2\pi^2} \left(\frac{2m_h}{\hbar^2} \right)^{3/2} \sqrt{E_v - E} \quad (8.41)$$

where the symbols m_e and m_h represent the effective masses of the electrons and holes, E_c and E_v denote the minimum and maximum for the conduction and valence bands, respectively. Equation 8.41 includes the electron spin degeneracy. The density of states function $g(E)$ has units of the number of states per unit energy per unit volume. Notice that the energies in the density of state functions are referenced to the conduction band minimum E_c and the valence band maximum E_v , where $E_c = E_v + E_g$ and E_g represents the band gap energy. Figure 8.23 shows very sharp cutoffs for the density of state functions at the band edges. For intrinsic semiconductors, the Fermi level E_f sits approximately midway between the conduction band CB and valence band VB.

Now we can calculate the density of carriers n_E as the product of the density of states $g(E)$ and the electron Fermi function $F(E)$. Figure 8.23 shows the results. As E approaches the band edge, the density of states $g(E)$ rapidly decreases. Likewise, the carrier distributions n_E and p_E rapidly decrease as E moves away from the edge because the Fermi function exponentially decreases. Therefore, the electrons tend to accumulate near the lowest portion of the conduction band and we only need to consider those states within a few kT of the band edge. Similar comments apply to holes in the valence band. As a note, we sometimes refer to the number of electrons

$$n = \int_{E_c}^{\infty} g(E) F_c(E) dE$$

for intrinsic (undoped) semiconductors as the “intrinsic” number of electrons n_i (and not to be confused with the number of electrons in energy level $\#i$).

8.4.3 COMMENTS

For an intrinsic semiconductor at thermal equilibrium, the number of electrons in the conduction band must be the same as the number of holes in the valence band. In other words, we use $n = p = n_i$, where n, p, n_i represent the density of electrons, holes, and intrinsic carriers, respectively. We find the apparently trivial relation $np = n_i^2$. Its real significance becomes apparent when we realize that it also applies to doped semiconductors. In this case, none of the quantities n, p, n_i have the same value. While n_i remains the number of electrons (or holes) for the intrinsic semiconductor at temperature T , the n and p now refer to the number of electrons and holes in the doped semiconductor at temperature T . Later sections in the present chapter further discuss the law of mass action.

The Fermi–Dirac distribution describes the system of electrons in thermal equilibrium. All portions of the crystal must be at the same temperature for the system to be in thermal equilibrium. Shining light on the crystal (if it is absorbed) spoils the thermal equilibrium. For example, assume the crystal has a temperature of 0 K. Further suppose that we continuously illuminate the crystal with photons having energy significantly larger than the band gap. Then the absorption process produces holes in the valence band and electrons in the conduction band. However, the Fermi function (at $T = 0$) does not allow electrons in the conduction band. Therefore, we have created a situation where the actual electron distribution cannot be described by the Fermi distribution. In fact, the continuous absorption (of a single small wavelength for example) places electrons well above the conduction band edge and leaves holes well below the valence band edge because of the large photon energy. So not even a temperature T different from zero can describe the actual electron distribution since thermal distributions always place the electrons in available states with the lowest energy. We therefore see that light absorption represents a nonequilibrium process. We must use quasi-Fermi levels to describe the actual electron and hole distributions for nonequilibrium conditions. The single Fermi level splits into two quasi-Fermi levels, one to describe the number of electrons in the conduction band and another to describe the number of holes in the valence band. One should note that the quasi-Fermi levels in conjunction with the Fermi distribution will not properly describe the carrier distribution caused by the absorption of a single wavelength of light prior to carrier thermalization. The carriers initially occupy levels beyond the minimums for the band whereas the distributions require more carriers at the bottom (regardless whether one uses quasi-Fermi levels or not). However, if the carriers thermalize in the sense of assuming an exponential distribution within their respective bands, then it is possible to apply the quasi-Fermi levels.

8.5 DERIVATION OF FERMİ–DIRAC DISTRIBUTION

The present section derives the Fermi–Dirac distribution from first principles which accounts for the Pauli exclusion principle and the fact that one cannot distinguish between electrons. The procedure maximizes the entropy function using Lagrange multipliers.

8.5.1 PAULI EXCLUSION PRINCIPLE

The Fermi–Dirac distribution applies to quantum mechanically indistinguishable Fermions such as electrons. First consider the Fermion aspects and then the effects of the particles being indistinguishable.

Fermion particles obey the Pauli exclusion principle whereby two or more Fermions cannot occupy a single state at the same time. In general, Fermion particles have half-integer spin ($1/2, 3/2, \dots$). As discussed in Chapter 5, one might view spin from the classical point of view as the angular momentum associated with a rotating object; however, true point particles would not have classical spin angular momentum. Quantum mechanics allows only certain values of angular momentum.

The Pauli exclusion principle for Fermions can be linked to anti-commutation relations for the Fermion creation $\hat{f}^+$ and annihilation $\hat{f}^-$. Suppose a ket $|\rangle$ represents an electron state. The creation operator places a single electron in the state according to $|1\rangle = \hat{f}^+|0\rangle$, where $|0\rangle$ represents the empty state. Likewise, the annihilation operator removes a single particle according to $|0\rangle = \hat{f}^-|1\rangle$ and where $\hat{f}^-|0\rangle = 0$. The Fermion creation and annihilation operators obey anticommutation relations

$$[\hat{f}^-, \hat{f}^+]_+ = 1 \quad [\hat{f}^+, \hat{f}^+]_+ = 0 \quad [\hat{f}^-, \hat{f}^-]_+ = 0 \quad (8.42)$$

where the anticommutator is defined by

$$[\hat{A}, \hat{B}]_+ = \hat{A}\hat{B} + \hat{B}\hat{A} \quad (8.43)$$

One can illustrate the relation between the commutation relations and the Pauli exclusion principle. Let the anticommutator $[\hat{f}^+, \hat{f}^+]_+ = 0$ operate on the vacuum state $|0\rangle$. Then

$$[\hat{f}^+, \hat{f}^+]_+|0\rangle = 0|0\rangle = 0 \quad \text{or} \quad [2\hat{f}^+\hat{f}^+]|0\rangle = 0$$

Therefore, we see that trying to place two Fermions in the same state results in zero

$$\hat{f}^+\hat{f}^+|0\rangle = 0$$

In contrast to the Fermion, any number of Boson particles with their integer spins (0, 1, 2, ...) can occupy a single state at one time. For example, any number of photons (spin 1) can occupy the fundamental Fabry-Perot resonator mode. This means the fundamental sine wave can have any amplitude as determined by the number of photons in the mode. The fact that an unlimited number of bosons can occupy a single mode can be linked to the commutation relations for the Boson creation $\hat{b}_{\mathbf{k}s}^+$ and the annihilation $\hat{b}_{\mathbf{k}s}^-$ operators

$$[\hat{b}^-, \hat{b}^+] = 1 \quad [\hat{b}^-, \hat{b}^-] = 0 \quad [\hat{b}^+, \hat{b}^+] = 0 \quad (8.44)$$

Indistinguishable quantum particles obey different statistics than their indistinguishable classical counterparts. Quantum mechanically, we cannot distinguish between two Fermions in different states with the same energy nor between two Fermions occupying two states with different energy as shown in Figure 8.24. Switching the particles in either position or energy does not result in a new thermodynamic microstate. We can see this behavior from the electron wave function. Suppose a system consists of exactly two Fermions. The wave function for the system with one particle in state “a” and the other in state “b” comes from the Slater determinant.

$$\psi = C \begin{vmatrix} \psi_a(x_1) & \psi_b(x_1) \\ \psi_a(x_2) & \psi_b(x_2) \end{vmatrix} \sim \psi_a(x_1)\psi_b(x_2) - \psi_a(x_2)\psi_b(x_1) \quad (8.45)$$

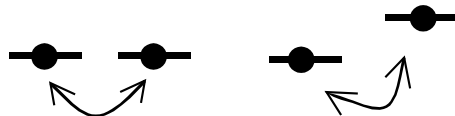


FIGURE 8.24 Indistinguishable Fermions.

Switching “ a ” and “ b ” does not affect the wave function (except for a minus sign). Also, if $a = b$ then the wavefunction is zero (no more than one particle per energy state). These considerations change the counting procedure.

8.5.2 BRIEF REVIEW OF MAXWELL–BOLTZMANN DISTRIBUTION

Recall, entropy $S = k\text{Ln}(\Omega)$ measures “disorder” and must be related to probability. Here Ω represents the number of different arrangements of particles consistent with the particle properties such as indistinguishability. Figure 8.25 shows an example of 6 electrons in 6 states. The system has a very high degree of order. If one removes the barrier between the two sides, then the electrons diffuse to the right-hand side; this motion produces current and results in higher entropy. Removing the barrier increases the number of states available to the system of electrons and therefore increases the entropy.

The Boltzmann distribution can be derived by maximizing the number of combinations Ω for $\mathcal{N}$ classically distinguishable particles with n_i of the particles occupying energy level E_i with degeneracy g_i .

$$\Omega = \mathcal{N}! \prod_{i=1}^{N_L} \frac{g_i^{n_i}}{n_i!}$$

The derivation uses two constraints

$$\sum_{i=1}^{N_L} n_i = \mathcal{N} \quad \text{and} \quad \sum_{i=1}^{N_L} E_i n_i = \mathcal{N} \mathcal{E}$$

where $\mathcal{E}$ represents the total energy of the ensemble and the average system energy, respectively.

The Maxwell–Boltzmann distribution for classically indistinguishable particles (Section 8.3) gives the probability of finding a system at temperature T in a particular state “ s ” with energy E_s .

$$P(s|E_s) = \frac{e^{-\beta E_s}}{\sum_s e^{-\beta E_s}}$$

The partition function $Z = \sum_s e^{-\beta E_s}$ sums over all microstates regardless of energy. Recall that the partition function, besides being important for determining microscopic quantities, provides the normalization constant to ensure that the probabilities add up to one. The probability distribution implicitly includes the degenerate states in the summation. If $g(E_s)$ represents the number of states with energy E_s then the probability distribution can also be written as

$$P(E_s) = \frac{g(E_s) e^{-\beta E_s}}{\sum_{E_s} g(E_s) e^{-\beta E_s}}$$

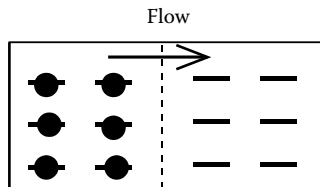


FIGURE 8.25 Electrons flow to maximize entropy.

where now the summation is over the energy values rather than each microstate, and $\beta = 1/(kT)$. The Maxwell–Boltzmann distribution assumes that a system has equal probability of occupying any microstate with the same energy E_s .

8.5.3 FERMİ–DIRAC AND BOSE–EINSTEIN DISTRIBUTIONS

The Fermi–Dirac distribution describes thermal equilibrium for electrons in the semiconductor crystal and it can be derived by maximizing the entropy

$$S = k \text{Ln}(\Omega) \tag{8.46}$$

where k denotes Boltzmann’s constant. The symbol Ω represents the number of microstates available to the system, which can be determined by the number of different arrangements of N electrons subject to two constraints. Each possible arrangement of the electrons represents a microstate. The procedure explicitly uses the fact that electrons cannot be distinguished and no more than one of them can occupy a single state at any given time.

The counting procedure for the indistinguishable Fermions differs from the distinguishable boson-like particles. First examine the fact that the Fermions must be indistinguishable as shown in Figure 8.26. Unlike the Boltzmann particle, we do not include extra factors for switching two Fermions between states. Similarly, we do not include extra factors for switching between energy levels as shown in Figure 8.27. Fermions obey the Pauli exclusion principle and we cannot place two of them in the same state at the same time.

We now find the number of possible arrangements Ω of N electrons with n_i in level i with energy E_i and degeneracy g_i . Imagine that the number n_i remains fixed for the counting. After counting, the numbers n_i can be varied to find the numbers n_i that gives the largest number of arrangements and therefore the largest probability of occurring.

We first deduce the general formula for the number of arrangements Ω starting with a simple system (Figure 8.28) consisting of two levels $N_L = 2$, each having four degenerate levels $g_i = 4$, with two electrons in each level $n_i = 2$ for $i = 1, 2$. First look at the level E_1 . Figure 8.29 shows the

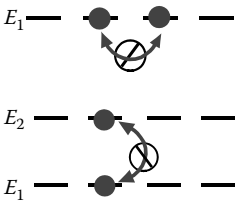


FIGURE 8.26 Interchanging electrons does not produce a new microstate.

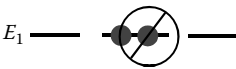


FIGURE 8.27 Two Fermions cannot occupy the same state at the same time.

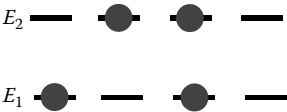


FIGURE 8.28 The simple system.

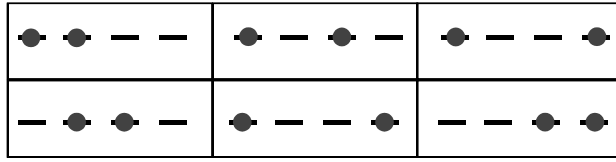


FIGURE 8.29 Six distinct arrangements for the first level E_1 .

possible arrangements of the $n_1 = 2$ electrons in the $g_1 = 4$ states. None of the microstates have two electrons in a single state and none of the microstates repeat. The number of microstates (i.e., the number of different arrangements) is 6. As discussed in Appendix G, the number can also be calculated by taking 2 objects from 4 without regard to order (i.e., indistinguishable).

$$\Omega_1 = \binom{g_1}{n_1} = \binom{4}{2} = \frac{4!}{(4-2)! 2!} = 6$$

Similarly level E_2 must also have the same number of possible arrangements

$$\Omega_2 = \binom{g_2}{n_2} = \binom{4}{2} = \frac{4!}{(4-2)! 2!} = 6$$

Unlike the Boltzmann distribution, we do not switch electrons between levels and therefore do not include another factor of $2 \cdot 2 = 4$. The total number of arrangements must be

$$\Omega = \Omega_1 \Omega_2 = \binom{g_1}{n_1} \binom{g_2}{n_2} = 6 \cdot 6 = 36$$

This last formula can then be generalized for the total number of arrangements for the N_L levels as

$$\Omega = \prod_{i=1}^{N_L} \binom{g_i}{n_i} = \prod_{i=1}^{N_L} \frac{g_i!}{(g_i - n_i)! n_i!} \quad (8.47)$$

We maximize the entropy by maximizing the number of arrangements Ω . Figure 8.30 shows an example of multiple levels. Assume that the total number N of electrons in the crystal and the total energy E_T of the electrons remain constant in time.

$$N = \sum_{i=1}^{N_L} n_i \quad E_T = \sum_{i=1}^{N_L} n_i E_i \quad (8.48)$$

where the symbol N_L denotes the number of energy levels. For example, Figure 8.30 shows four electrons. The total energy in the system represented by the figure comes to $3 + 4 + 4 + 5 = 16$.

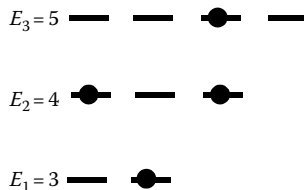


FIGURE 8.30 Multiple energy levels.

So the electrons can rearrange themselves in many ways so long as the total number of electrons remains 4 and the total energy remains 16. Mathematically, we vary the number of electrons n_i in each level in order to find the occupation distribution with the greatest number of arrangements and therefore with the greatest probability (leading to the greatest entropy).

Maximize the entropy with respect to the number of electrons in each level with energy E_i

$$S = k \ln(\Omega) = k \ln \left[\prod_{i=1}^{N_L} \frac{g_i!}{n_i!(g_i - n_i)!} \right] = k \sum_{i=1}^{N_L} [\ln(g_i!) - \ln(n_i!) - \ln((g_i - n_i)!)] \quad (8.49)$$

subject to two constraints. The total number of electrons must remain constant at $N = \sum_{i=1}^{N_L} n_i$ and the total energy must remain at $E_T = \sum_{i=1}^{N_L} n_i E_i$. The natural log terms in Equation 8.49 can be simplified by using the sterling approximation for large values of n

$$\ln(n!) = n \ln(n) - n$$

Therefore, Equation 8.49 becomes

$$S = k \ln(\Omega) = \sum_i \{ \ln(g_i!) - n_i \ln(n_i) + n_i - (g_i - n_i) \ln(g_i - n_i) + (g_i - n_i) \} \quad (8.50)$$

The maximum entropy is found by differentiating S and setting the result to zero: $0 = dS$. Using the fact that g_i is constant, we can write

$$0 = dS = k \sum_i [-\ln(\bar{n}_i) - 1 + \ln(g_i - \bar{n}_i) + 1] \delta n_i = k \sum_i \ln\left(\frac{g_i - \bar{n}_i}{\bar{n}_i}\right) \delta n_i \quad (8.51)$$

We should make a comment on the procedure taking us from Equation 8.49 to Equation 8.51. The entropy in Equation 8.49 has the form

$$S = \sum_i \ln[f(n_i)] \quad (8.52a)$$

where f represents the more complicated function in the summation. The entropy S is maximized by setting the differential to zero

$$0 = dS = \sum_j \frac{\partial S}{\partial n_j} dn_j \quad (8.52b)$$

Notice the use of “ j ” instead of “ i ” for the summation even though S involves the “ i ”. Inserting 8.52a into 8.52b, we find

$$0 = dS = \sum_{j,i} dn_j \frac{\partial}{\partial n_j} \ln[f(n_i)] \quad (8.52c)$$

The last equation only produces a contribution to the derivative when $i = j$. Therefore, the differential in 8.52c reduces to

$$0 = dS = \sum_i dn_i \frac{\partial}{\partial n_i} \ln[f(n_i)] \quad (8.52d)$$

as required.

If all of the variations δn_i in Equation 8.51 could be taken as independent, then their coefficients could be set to zero. However, the maximization must account for the two equations of constraint (Equation 8.48)

$$N = \sum_{i=1}^{N_L} n_i \quad E_T = \sum_{i=1}^{N_L} n_i E_i \quad (8.53)$$

that reduce the number of independent variations to $N_L - 2$. We can use the method of Lagrange Multipliers (see Appendix H) to treat all of the variations as independent. The method of Lagrange multipliers incorporates the constraints into Equations 8.51. Taking the variation of the constraints in Equations 8.53 provides

$$\sum_i \delta n_i = 0 \quad \text{and} \quad \sum_i E_i \delta n_i = 0 \quad (8.54)$$

Multiplying the first one by α and the second by β , and adding the results into Equation 8.51 provides

$$\sum_i \left\{ \text{Ln} \left(\frac{g_i - \bar{n}_i}{\bar{n}_i} \right) - \alpha - \beta E_i \right\} \delta n_i = 0 \quad (8.55)$$

where α and β incorporate the Boltzmann constant k .

The method of Lagrange multipliers allows us to conclude that

$$\text{Ln} \left(\frac{g_i - \bar{n}_i}{\bar{n}_i} \right) - \alpha - \beta E_i = 0$$

and hence

$$\bar{n}_i = \frac{g_i}{\exp(\alpha + \beta E_i) + 1} \quad (8.56)$$

After some work, the constraints give the values of α, β as

$$\alpha = -\beta E_f \quad \beta = \frac{1}{kT}$$

where E_f represents the Fermi level that approximates the chemical potential.

$$\bar{n}_i = \frac{g_i}{\exp[\beta(E_i - E_f)] + 1} \quad (8.57)$$

Therefore, the number of electrons per state can be written as

$$F(E) = \frac{\bar{n}_i}{g_i} = \frac{1}{\exp[\beta(E - E_f)] + 1} \quad (8.58)$$

Notice that the degeneracy divides out to define the Fermi function. For this reason, semiconductor models must include the density of states as a separate factor. In fact, Equation 8.57 looks just like the product of the Fermi function with the density of states.

The existence of the Fermi level E_f serves as a reminder that a system maximizes the entropy by allowing the constituent particles to move around. For example, placing p-type and n-type semiconductors in contact produces diffusion currents in order to ensure the Fermi level is flat (with respect to the vacuum level) through both pieces of material. The diffusion currents flow until the built-in field increase.

8.6 EFFECTIVE DENSITY OF STATES, DOPING, AND MASS ACTION

Dopant atoms affect the electrical properties of a crystal. As mentioned in Chapter 1, n-type dopants have one more valence electron than required for bonding. For example, phosphorus serves as an n-type dopant for silicon (see Figure 8.31). Not all of the dopant's valence electrons participate in bonding. For temperatures larger than the ionization temperature (i.e., more than about $T_i = 100$ K) the dopants ionize and the outer valence electrons can move about the crystal. Much larger temperatures can cause significant transfer of the electrons from the bonds (and eventually melting). For a dopant, the core has a charge of $+e$. Also of special note, the ionized dopant atom has an available (localized) state for an electron. As a result, this doping state must show up on the energy band ($E-k$) and band-edge diagrams.

The p-Type dopants have one less electron in the valence shell than do the atoms in the host material. For example, boron serves as a p-type dopant for silicon. A p-type dopant is neutral at low temperatures but, when “ionized,” it acquires an extra electron and becomes negatively charged $-e$. At 0 K, all of the crystal electrons reside in the VB and the acceptor states do not have electrons and therefore have stationary holes.

Consider n-type dopants. For temperatures less than the ionization temperature, the extra valence electron orbits about the dopant core forming a hydrogen-like atom. However, unlike the hydrogen atom, the binding energy of the electron (ionization energy) remains relatively small (on the order of a few milli-eV). The electrons form large orbits about the core enclosing a large number of atoms. The reasons for the small binding energy (i.e., small energy between the dopant state and the conduction band minimum) include a small effective mass for the electron (or hole) and the dielectric constant associated with the collection of atoms falling within the orbit (the dielectric constant reduces the electric field and decreases the binding force). The binding energy becomes

$$E_b = \frac{m^* e^4}{2(4\pi K \epsilon_0 \hbar^2)} = \frac{13.6}{K^2} \frac{m^*}{m_0}$$

where

E_b denotes the binding energy measured in eV

e denotes the elementary charge

K denotes the dielectric constant

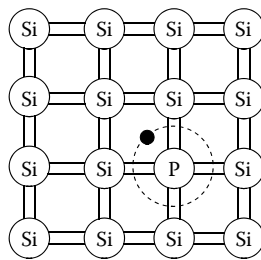


FIGURE 8.31 A cartoon representation of an n-type dopant atom embedded in a silicon host crystal. The dopant atom loosely binds the outermost valence electron. For the actual situation, the electron orbit will encompass many atoms that will further lower the binding energy of the electron (through the dielectric response).

ϵ_0 represents the permittivity

m^* represents the effective mass (electron or hole)

m_0 represents the free mass of the electron

Doping a semiconductor increases the concentration of carriers (holes or electrons) and can create internal electric fields for diodes (homojunctions). Many devices including photodetectors, solar cells, LEDs, semiconductor lasers, and transistors use the diode structure in one form or another. Some devices use homojunctions formed by two identical materials brought into contact but each having different types of doping. For the case of the n–p homojunction, the electrons, and holes diffuse to the other sides forming a built-in field (a.k.a., space charge region). Other devices use heterojunctions formed from two different types of materials such as GaAs–AlGaAs. The PIN structure represents a derivative of the pn diode. The intrinsic layer (I, undoped) can be used as an absorption region for photodetectors. Other devices, such as Ohmic contacts, intentionally circumvent any interfacial junctions; they incorporate highly doped regions (10^{19} dopant atoms per cm^3). For Ohmic contacts, certain metals are evaporated onto the semiconductor and then annealed (heated) to form a junction with linear I – V characteristics (and presumably very low resistance).

8.6.1 CARRIER CONCENTRATIONS

The present section calculates the electron density (number per volume) for the conduction band and the hole density for the valence band for doped semiconductors. Density comes from the product of the density of states and the Fermi function according to

$$n = \int_{E_c}^{\infty} g_e(E) F_e(E) dE \quad p = \int_{-\infty}^{E_v} g_h(E) F_h(E) dE \quad (8.59)$$

where

$$g_{\text{elect}}(E) = \frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \sqrt{E - E_c} \quad g_{\text{hole}}(E) = \frac{1}{2\pi^2} \left(\frac{2m_h}{\hbar^2} \right)^{3/2} \sqrt{E_v - E} \quad (8.60)$$

$$F_e(E) = \left[e^{\frac{E - E_F}{kT}} + 1 \right]^{-1} \quad F_h(E) = 1 - F_e(E) \quad (8.61)$$

where

m_e and m_h denote the effective masses of electrons and holes

E_c and E_v represent the conduction and valence band edges

Recall the effective masses can be an average over the components of the tensor effective masses.

The band-edge diagrams represent a most useful approximation. The conduction and valence bands appear as two levels in plots of energy versus spatial position. These diagrams work because electrons tend to accumulate within a couple of kT of the band edge. The value of kT is on the order of 25 m-eV which is small compared with the relatively large band gap energy on the order of 1 eV. Therefore only those states near the band edge have any significance for the conduction process. The effective density of states provides a description of those important states.

To find the effective density of states, first calculate the number of electrons in the conduction band.

$$n = \int_{E_c}^{\infty} g_e(E) F_e(E) dE = \int_{E_c}^{\infty} dE \left[\frac{1}{2\pi^2} \left(\frac{2m_e}{\hbar^2} \right)^{3/2} \frac{\sqrt{E - E_c}}{e^{\frac{E - E_F}{kT}} + 1} \right] = \frac{1}{2\pi^2} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} \int_0^{\infty} d\xi \frac{\sqrt{\xi}}{e^{\xi + \frac{(E_c - E_F)}{kT}} + 1} \quad (8.62)$$

where the last integral uses the change of variables $\xi = (E - E_c)/kT$, which references the energy to the edge of the conduction band in units of kT . The coefficient of the integral can be related to the “effective density of states” N_C by

$$\frac{1}{2\pi^2} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} = \frac{2}{\sqrt{\pi}} \left[2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \right] = \frac{2}{\sqrt{\pi}} N_C$$

where the factor of $2/\sqrt{\pi}$ remains in order to cancel a factor from the integral. The effective density of states must be

$$N_C = 2 \left(\frac{m_e kT}{2\pi\hbar^2} \right)^{3/2} \quad (8.63)$$

The Fermi–Dirac integral (of order 1/2) in Equation 8.62 can be written as

$$F_{1/2}(\eta_c) = \int_0^\infty d\xi \frac{\sqrt{\xi}}{e^{\frac{\xi - \eta_c}{kT}} + 1} \quad (8.64)$$

where $\eta_c = (E_F - E_c)/kT$ measures the energy difference between the Fermi level E_F and the bottom of the conduction band E_c in units of kT . The Fermi–Dirac integral only depends on η_c . This means that we can interpret the product in Equation 8.62 as if there were a single energy level near the conduction band edge with the Fermi–Dirac integral acting as an “effective Fermi function” for those levels. To see this, rewrite Equation 8.62 as

$$n = \frac{1}{2\pi^2} \left(\frac{2m_e kT}{\hbar^2} \right)^{3/2} \int_0^\infty d\xi \frac{\sqrt{\xi}}{e^{\xi + \frac{(E_c - E_F)}{kT}} + 1} = \frac{2}{\sqrt{\pi}} N_C F_{1/2}(\eta) \quad (8.65)$$

and evaluate the integral using a Boltzmann approximation as follows.

$$n = N_C \frac{2}{\sqrt{\pi}} \int_0^\infty d\xi \frac{\sqrt{\xi}}{e^{\xi + \frac{(E_c - E_F)}{kT}} + 1} \cong N_C \frac{2}{\sqrt{\pi}} \int_0^\infty d\xi \sqrt{\xi} e^{-\xi - \frac{(E_c - E_F)}{kT}} = N_C e^{-\frac{(E_c - E_F)}{kT}} \frac{2}{\sqrt{\pi}} \int_0^\infty d\xi \sqrt{\xi} e^{-\xi}$$

The approximation assumes that $E_c - E_F > 3kT$. By the way, when E_F comes within $3kT$ of the conduction band, the semiconductor must be very heavily doped (degenerately doped). The integral can be computed by numerical methods. We find

$$n = N_C e^{-\frac{(E_c - E_F)}{kT}} \quad (8.66)$$

We see that the number of energy states (per unit volume) at the edge of the conduction band (at E_c) must be N_C and the probability of electrons occupying a given state must be $\exp[-(E_c - E_F)/(kT)]$. The charge resides within several kT of the band edge, which justifies the use of band-edge diagrams.

Similar calculations can be made for the holes in the valence band. Using Equations 8.59 through 8.61:

$$p = \int_{-\infty}^{E_v} dE g_h(E) [1 - F_e(E)] = N_v e^{\frac{E_v - E_F}{kT}} \quad (8.67)$$

where the effective density of states for the conduction band is given by

$$N_v = 2 \left(\frac{m_h kT}{2\pi\hbar^2} \right)^{3/2} \quad (8.68)$$

The number of electrons and holes depends on the position of the Fermi level as illustrated by Equations 8.66 and 8.67. In fact, as E_f comes closer to the conduction band, the density of electrons n (number per unit volume) increases while the number of holes decreases. We will see this again with the law of mass action. Here is the real question. What determines the position of the Fermi level? We discuss this more fully later. For now, we show that the parameter E_f , the Fermi energy, can be replaced by either n or p .

8.6.2 LAW OF MASS ACTION

The law of mass action states

$$np = n_i^2 \quad (8.69)$$

where n_i denotes the “intrinsic” electron density at temperature T . For a given temperature, n_i is a constant and can be related to the effective density of states in the VB and CB, and to the gap energy E_g . The CB electron density n and the VB hole density p in Equation 8.69 apply to either intrinsic or extrinsic (doped) material. For example, for n-type material there must be more electrons than holes. The Fermi level moves closer to the conduction band and the Fermi function has the value of approximately “1” for the valence states. As a result, electrons fill essentially all of the valence band and partially fill the conduction band. Essentially, the extra electrons (from doping) fall into the VB and fill it up. Therefore, doping increases n but also decreases p so that the product np remains constant as given by Equation 8.69.

We proceed in two steps to prove the law of mass action. The first step gives a value for np in the intrinsic case. We already know for intrinsic semiconductors that the number of holes equals the number of electrons $n = p \equiv n_i$. For the intrinsic case, Equations 8.66 and 8.67 provide

$$n_i^2 = n_i p_i = N_C e^{-\frac{(E_C - E_f^{(i)})}{kT}} N_V e^{\frac{E_V - E_f^{(i)}}{kT}} = N_C N_V e^{-\frac{E_G}{kT}} \quad (8.70)$$

where $E_f^{(i)}$ is the Fermi level for the intrinsic semiconductor.

Next, calculate the product np for the case of a nondegenerately doped semiconductor. Again using Equations 8.66 and 8.67 we find

$$np = N_C e^{-\frac{(E_C - E_f)}{kT}} N_V e^{\frac{E_V - E_f}{kT}} = N_C N_V e^{-\frac{E_G}{kT}} \quad (8.71)$$

where now E_f denotes the Fermi level for the doped semiconductor. The right-hand sides of Equations 8.70 and 8.71 have the same value. Therefore, the law of mass action holds for the case of nondegenerately doped semiconductors.

$$np = n_i^2$$

8.6.3 ELECTRIC FIELDS

The introduction showed that the dopant atom (n or p) is neutrally charged at 0 K. At higher temperatures, the electron leaves the donor atom or a hole leaves the acceptor atom. In either case the atom acquires a charge. The electron leaves the donor atom and leaves behind a stationary core

with charge $+e$. The acceptor atom gains an electron and becomes negatively charged. Given that the operation of most semiconductor devices depend on the number of free carriers (holes and electrons) as well as the electric fields in the material (think of the pn diode or the pin photodetector), it is necessary to account for the number of dopant atoms that are actually ionized at a given temperature.

The static electric field can be related to the charges in the material. Let us consider “charge neutrality.” Take a chunk of semiconductor and ask what the electric field is surrounding the chunk. We know that there can be no electric field around the chunk if the net charge is zero. Assume that we do not artificially charge the material. The atoms in the crystal have as many electrons as protons; this is true for the dopant atoms as well. As a result, there are as many positive as negative charges and so the chunk of semiconductor must have a net charge of zero (charge neutrality).

Next look at the semiconductor on a microscopic scale and again discuss the charge neutrality relation. Assume the following definitions (number/vol)

N_d	Number of donor atoms	N_d^+	Number of ionized donors
N_a	Number of acceptors	N_a^-	Number of ionized acceptors
n	Number of CB electrons	p	Number of VB holes

The electric field at any point in the semiconductor must be related to the number of charged particles (whether they be holes, electrons, or ionized dopants). In a volume ΔV of crystal, we can write the charge density (charge per unit volume) as

$$\rho_{\text{net}} = +e(p + N_d^+) - e(n + N_a^-) \quad (8.72a)$$

The little chunk ΔV of semiconductor must be neutrally charged if $\rho_{\text{net}} = 0$, that is

$$-n + p + N_d^+ - N_a^- = 0 \quad (8.72b)$$

Of course this means that the electric field due to the chunk ΔV must also be zero as can be seen from Gauss’ law:

$$\int_{\Delta V} d\vec{a} \cdot \vec{E} = \frac{1}{4\pi\epsilon} \int_{\Delta V} dV \rho_{\text{net}} = 0$$

Any electric fields in the volume ΔV must arise from external agents or from charge imbalance from neighboring volumes ΔV . Any states within the semiconductor need to be included in the calculation of the total charge. This can include traps and recombination centers in the bulk or on the surface. In fact, it is possible for a crystal to be electrically neutral overall, but to have charged trapped on the surface that produces an electric field (dipole field) into the bulk of the material where the opposite charge resides.

An imbalance of charge from Equation 8.72 would then be written as

$$\rho_{\text{net}} = +e(p + N_d^+) - e(n + N_a^-) = e \left\{ N_v e^{-\frac{E_f - E_v}{kT}} + \frac{N_d}{1 + g_d e^{\frac{E_f - E_d}{kT}}} - N_c e^{-\frac{E_c - E_f}{kT}} - \frac{N_a}{1 + g_a e^{\frac{E_v - E_f}{kT}}} \right\}$$

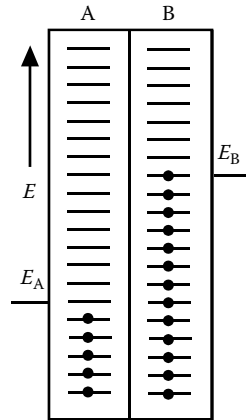


FIGURE 8.32 Two materials with different Fermi levels.

8.6.4 SOME COMMENTS

For a pn diode (see Chapter 1, Section 4), the total charge for the entire device adds to zero even though either side of the diode becomes charged. This occurs when the Fermi levels for n and p line up with the simultaneous transfer of electrons from the n side to the p side. However, for the device as a whole, the charge is conserved and the total charge must be zero.

A final comment concerns the reason for charge transfer. Suppose two materials A and B are brought into contact at time $t=0$ such as in Figure 8.32 (at $T=0$ K). The Fermi levels for type A and B materials at $t=0$ are positioned at the top energy for the occupied states $E_A < E_B$. The figure could correspond to semiconductors having band gap states. Consider small enough times so that the Fermi levels have not changed much from their initial energy with respect to each other. In such a case, there are empty states in type A with energy lower than E_B . Assuming normal physical conditions and good electrical contact between types A and B, the states in A in the range $E_A < E < E_B$ should have as many electrons as type B since there is a high probability (after contact) that they will be filled. That is, the contact causes the probability of occupation to agree between the two sides. Upon contact, electrons diffuse from B to A to occupy available states which increases the entropy. For semiconductors, this means an electric field is established at the same time which also alters the relation between the energy on either side. Here again, the electric field establishes since the transferred electrons leave behind positively charged cores. The field exists between these transferred electrons and the cores. As an important exercise, the reader should consider how the electric field in turn affects the relative position of the two Fermi-levels.

8.7 DOPANT IONIZATION STATISTICS

Many devices have doped semiconductors to increase the conductivity of the material. However, the dopants remain ionized only for temperature larger than a minimum value. The present section derives the Fermi function for dopants in order to determine the occupation probability as a function of temperature. Sufficiently low temperatures push the electron and hole populations into an intrinsic region of operation which most often prevents proper device operation.

8.7.1 DOPANT FERMION FUNCTION

We already know the Fermi distribution for valence band holes and conduction band electrons, but what Fermi function should be used for the dopants (assuming the same Fermi energy as for the conduction and valence bands)?

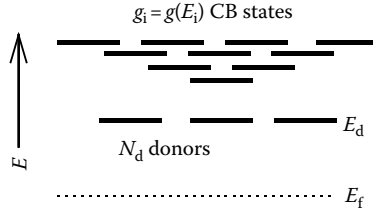


FIGURE 8.33 The number N_d donors with energy E_d below the conduction band (CB) states.

This section of discussion provides an outline of the procedure for calculating the probability of an electron occupying a dopant level; most of the calculation will be left as an exercise for the reader. The result provides the average number of ionized donors as a function of temperature. The same procedure can be used for the acceptor levels. The calculation requires (Figure 8.33) the dopant energy levels E_d (donors) or E_a (acceptors), the total number of donors N_d , the total number of acceptors N_a , and the Fermi level E_f . The calculation yields the number of ionized donors N_d^+ and the number of ionized acceptors N_a^+ according to the following results

$$\frac{N_d^+}{N_d} = \frac{1}{1 + g_d e^{\frac{E_d - E_f}{kT}}} \quad \frac{N_a^-}{N_a} = \frac{1}{1 + g_a e^{\frac{E_a - E_f}{kT}}} \quad (8.73)$$

Notice that Equation 8.73 for N_d^+ for example, provides a Fermi function $F_d(E)$ (number of donors occupied by an electron) given by

$$1 - F_d(E_d) = \frac{N_d^+}{N_d} \quad \rightarrow \quad F_d(E_d) = \frac{N_d - N_d^+}{N_d} = \frac{n_d}{N_d} \quad (8.74)$$

where $n_d = N_d - N_d^+$ must be the average number of occupied donors (i.e., the number of electrons in donor states). Similar considerations hold for the acceptor levels. The “degeneracy factors” g_d and g_a typically have values of $g_d = 2$ and $g_a = 4$. The term “degeneracy factor” does not refer to degenerately doped semiconductors nor do the symbols g_d and g_a refer to the density of states used in the derivation of the Fermi distribution. The value of $g_d = 2$ comes from the fact that a donor can have only one electron per state but the state can accommodate either spin up or spin down (not both simultaneously). The value of $g_a = 4$ occurs because each acceptor can have spin up or down and also because there are heavy and light-hole bands.

8.7.2 DERIVATION

In this section, we indicate how to find Equation 8.73 for the donors. The calculation starts by finding the Fermi function given in Equation 8.74 and then working backwards to find N_d^+ .

The Fermi function $F_d(E_d)$ can be found from the total number of distinct combinations given by

$$\Omega' = \Omega \Omega_d \quad (8.75)$$

where Ω represents the number of distinct combinations for the conduction/valence band (Section 8.5) and Ω_d represents that for the donors

$$\Omega_d = \frac{g_d^{N_d - N_d^+} N_d!}{(N_d - N_d^+)! N_d^+!} = \frac{g_d^{n_d} N_d!}{n_d! (N_d - n_d)!} \quad (8.76)$$

The constraints become

$$N = \sum_{i=1}^{N_L} n_i + n_d \quad E = \sum_{i=1}^{N_L} E_i n_i + E_d n_d \quad (8.77)$$

Setting the derivative of the Entropy to zero produces the additional terms in n_d (having taken into account the constraints by the method of Lagrange multipliers). Assuming the variations dn_d are independent of the dn_i we have

$$0 = dn_d \left\{ \text{Ln}(g_d) - \frac{d}{dn_d} [n_d \text{Ln}(n_d) - n_d] - \frac{d}{dn_d} [(N_d - n_d) \text{Ln}(N_d - n_d) - (N_d - n_d)] - \alpha - \beta E_d \right\}$$

where Sterling's approximation has been used $\text{Ln}(n!) = n \text{Ln}(n) - n$. Carrying out the differentiation provides

$$\text{Ln}(g_d) - \text{Ln}(n_d) + \text{Ln}(N_d - n_d) - \alpha - \beta E_d = 0$$

Taking the exponential of both sides gives

$$\frac{g_d(N_d - n_d)}{n_d} = e^{\alpha + \beta E_d} \quad \rightarrow \quad \frac{n_d}{(N_d - n_d)} = g_d e^{-(\alpha + \beta E_d)} \quad \rightarrow \quad \frac{g_d e^{-(\alpha + \beta E_d)}}{1 + g_d e^{-(\alpha + \beta E_d)}} = \frac{n_d}{N_d} = F_d(E_d)$$

Then using the definition of F_d from Equation 8.74 we find

$$\frac{N_d^+}{N_d} = 1 - F_d(E_d) = 1 - \frac{g_d e^{-(\alpha + \beta E_d)}}{1 + g_d e^{-(\alpha + \beta E_d)}} = \frac{1}{1 + g_d e^{-(\alpha + \beta E_d)}} \quad (8.78)$$

as required.

8.8 pn JUNCTION AT EQUILIBRIUM

The notes in this section sketch the physical principles involved with establishing a pn junction. As previously mentioned, many modern devices rely on a pn junction for their operation. The present section provides an introduction to the pn junction. For more information, the reader should refer to the references at the end of the chapter (c.f., Sze's book).

The pn junction forms when p-type and n-type semiconductors come into sufficient contact to allow the transfer of electrons. The dopants provide mobile electrons and holes in the conduction and valence bands, respectively. We first examine the conditions of thermal equilibrium requiring the electron occupation to conform to the Fermi–Dirac distribution. For a semiconductor at equilibrium, the Fermi–Dirac distribution must be obeyed at each point in space regardless of the composition of the material. If two semiconductors come into contact, then the Fermi–Dirac distribution must eventually hold even for different types of doping or materials. However, there can be a transient response where the distributions in the two materials approach thermal equilibrium.

8.8.1 INTRODUCTORY CONCEPTS

A cartoon representation of the conduction and valence bands versus distance into a material appears in Figure 8.34. The energy-position of the Fermi energy level in a material indicates the predominant type of carrier. For p-type, the Fermi level F lies closer to the valence band and the

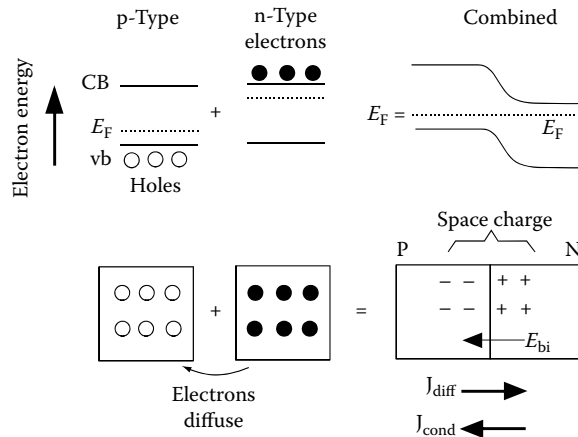


FIGURE 8.34 Combining two initially isolated doped semiconductors produces a pn junction with a built-in voltage (top). The built-in voltage is associated with a space charge region produced by drift and diffusion currents.

material has a larger number of free holes than free electrons. Similarly, a Fermi level F closer to the conduction band implies a larger number of conduction electrons.

Initially, the p-type and n-type materials are spatially separated and electrically isolated. Both materials must be in thermal equilibrium as indicated by the existence of the Fermi level. However, the energy between the vacuum level and the two Fermi levels must be different. Bringing the two materials into contact allows diffusion current to flow during an initial transient period to increase the entropy of the combined system. After the transient period, the electrons, and holes must once again conform to thermal equilibrium. However, the transfer of charge establishes a built-in field, which bends the bands and forms a barrier to further diffusion. The built-in field forces the two Fermi levels to coincide across the interface. We can see why the Fermi levels must coincide based on fundamental principles (see Section 8.6.4). Suppose for the sake of argument, the Fermi levels remain separated even though the two materials come into contact. Further, assume states exist at the Fermi level on either side of the interface (contrary to the figure). Half of the states at the Fermi level must be filled on either side of the interface. The electrons on the n-side would have larger energy than the electrons on the p-side. As a result, the electrons at the n-side Fermi level could lose the extra energy by transferring to the states at the p-side Fermi level. The process could continue so long as the Fermi levels remain separated. Eventually (without a bias current) the states on the n-side must empty and the states on the p-side must become full. The Fermi levels must therefore change from the initial configuration to describe the change in occupation. The n-side Fermi level must decrease while the p-side level must increase to bring them into coincidence. Equivalently said, the transfer of charge establishes a built-in field that changes the energy from the Fermi levels to the vacuum level and brings the levels into coincidence.

To establish thermal equilibrium, the holes and electrons transfer between the n-type and p-type materials. The diffusing electrons attach themselves to the p-dopants on the p-side but they leave behind positively charged cores. The separated charge forms a dipole layer and electric field. The built-in electric field prevents the diffusion process from continuing indefinitely.

During the transient period, as the junction establishes the built-in field, currents flow to establish thermal equilibrium. We define the diffusion current J_d to be the flow of positive charge under the force of diffusion alone (Figure 8.34 shows positive charge diffuses to the right across the junction). The conduction current J_c consists of charge flowing in response to an electric field alone. Equilibrium occurs when $J_c = J_d$. The particles stop diffusing when the built-in field establishes the electrostatic barrier at the junction. Electrons on the n-side of the junction would be required to

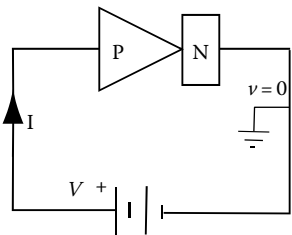


FIGURE 8.35 Forward biasing the pn junction. The n-side is grounded.

surmount the barrier to reach the p-side by diffusion; for this to occur, energy would need to be added to the electron. Other books such as by Sze or Parker discuss the nonequilibrium processes and the resulting quasi-Fermi levels. Sustained current flow through a diode constitutes a nonequilibrium process. Figure 8.35 shows the typical forward bias circuit for the diode with positive current flowing from the p- to n-side of the diode. As an important practical matter, a current limiting device such as a resistor should often be placed in the circuit to prevent damage to the diode! Diodes exhibit “turn-on” voltages produced by the built-in electric field that depends on the material composition. The magnitude of the built-in voltage is roughly the same as the band gap energy (in eV). The bias voltage must be larger than the turn-on voltage to obtain significant current flow (see the “dark” curve in Figure 8.36). The turn-on voltage is not very well defined because it is related to the Fermi-Dirac distribution for the electrons, which has an exponential dependence on energy (or voltage).

Small bias voltages approximately maintain thermal equilibrium allowing only small currents to flow. Applying large voltages reduces the built-in field and produces exponentially larger currents. Obviously, the number of electrons and holes must have changed from the equilibrium values in some region of space to produce the disproportionately larger currents. Therefore, diodes under forward bias do not obey the Fermi–Dirac distribution (with a single Fermi level) and the diode cannot be in thermal equilibrium. However, the forward biased diode can be in steady-state when the current remains constant in time. Obviously the terms “steady-state” and “equilibrium” refer to two separate conditions.

Later results show that the current density (amps per unit area) can be written as

$$J = en_i^2 \left(\frac{D_n}{L_n N_a} + \frac{D_p}{L_p N_d} \right) \left(e^{\frac{eV}{kT}} - 1 \right) \tag{8.79}$$

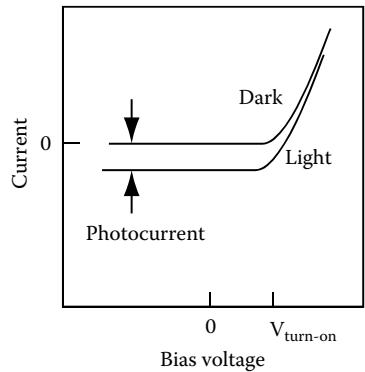


FIGURE 8.36 Current–voltage characteristics for a pn junction.

where $L_n, L_p, D_n, D_p, N_a, N_d$ represent the diffusion lengths for electrons on the p-side and holes on the n-side, the diffusion length of electrons and holes and the density of acceptor atoms (per volume) on the p-side and the density of donor atoms on the n-side, respectively. Notice in reverse bias $V < 0$, the exponential becomes small with respect to 1 and drops out of the equation. In this case, the reverse current saturates and must be independent of the applied voltage. The coefficient

$$J_s = en_i^2 \left(\frac{D_n}{L_n N_a} + \frac{D_p}{L_p N_d} \right) \quad (8.80)$$

gives the saturation current. This current arises from carriers thermally generated within the built-in field. The built-in field (and any applied field) sweeps the carriers out through the electrodes. The applied field is generally small compared with the built-in field and has very little control over the rate at which the charge is removed there (sufficiently large junction fields will sweep-out the electrons and holes before they can recombine there). Therefore, only the rate of thermal generation controls the saturation current.

8.8.2 QUICK CALCULATION OF BUILT-IN VOLTAGE OF PN JUNCTION

The thermal equilibrium condition determines the carrier concentrations (i.e., density). The concentration at a particular energy level only depends on the energy separating that level from the Fermi level (Figure 8.37).

We want to find the number of electrons and holes as a function of position. We (1) assume nondegenerate doping, (2) approximate the Fermi–Dirac distribution with the Boltzmann distribution, and (3) use the effective density of states approximation. Equations from the previous sections provide the electron $n(x)$ and hole $p(x)$ density as a function of position

$$n(x) = N_c e^{-\frac{E_c - E_f}{kT}} \quad p(x) = N_v e^{-\frac{E_f - E_v}{kT}}$$

where N_c, N_v, E_c, E_v , and E_f represent the effective conduction density of states, the effective valence density of states, the minimum of the conduction band, the maximum of the valence band, and the Fermi level. As usual, T represents the temperature in Kelvin.

Diffusion produces a built-in field and therefore a built-in voltage $V_b(x)$ that modifies the energy of all the CB and VB energy levels. The carrier densities become

$$n(x) = N_c e^{-\frac{(E_c - eV_b) - E_f}{kT}} \quad p(x) = N_v e^{-\frac{E_f - (E_v - eV_b)}{kT}} \quad (8.81)$$

The minus sign occurs between $E_{c,v}$ and V_b because the energy $E_{c,v}$ is a measure of electron energy rather than the usual energy of positive charges. For example, on the p-side, if $V_b > 0$ then the VB level moves further away from the Fermi level and the number of holes decreases. On the n-side, $V_b > 0$ requires the energy level to move closer to the Fermi level and increase the electron density.

We find the voltage $V_b(x)$ along the length of the diode by making some assumptions. For large positive x , the voltage on the n-side must be zero $V_n = 0$ since the n-side is tied to ground. For large negative x , the p-side has voltage $V_p = V_n + V_b$. Assume that deep within the bulk of the n- and p-type materials that the voltage is constant—the voltage only changes in the vicinity of the junction. It should be noted that the electric field is confined to the junction region since at equilibrium, the junction as a whole remains neutral. This means that the voltage changes only near the junction.

The total “electron” current density J_{tot} consists of the sum of the drift and diffusion currents

$$J_e = en\mu_n E + eD_n \frac{dn}{dx} = -en\mu_n \frac{dV}{dx} + eD_n \frac{dn}{dx} \quad (8.82)$$

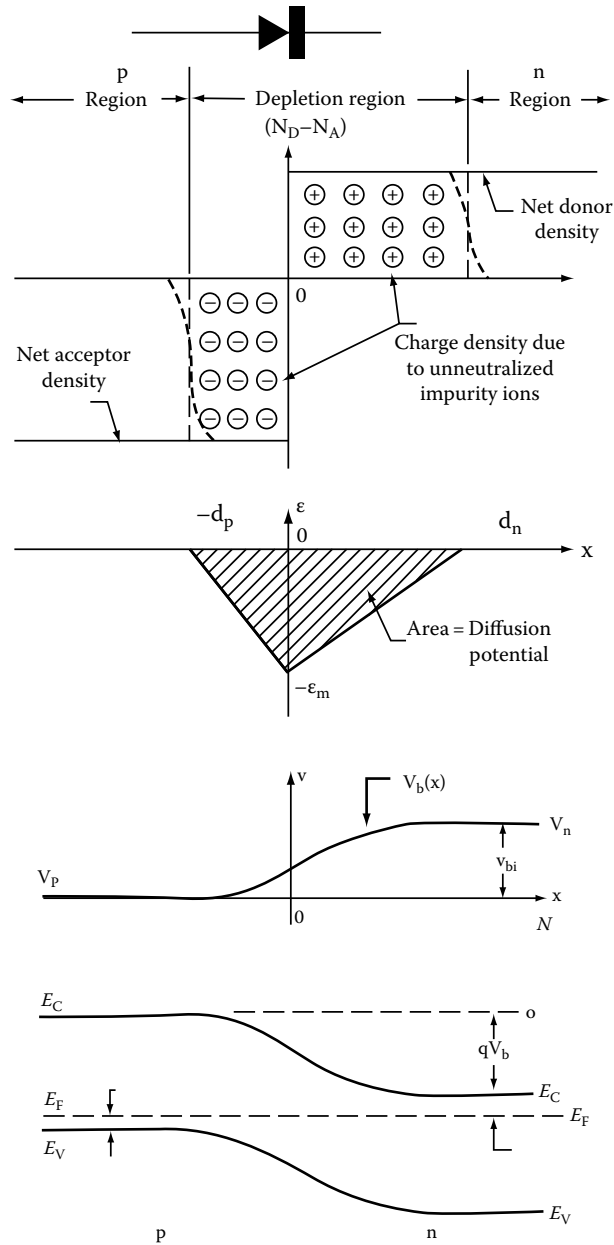


FIGURE 8.37 The diode internal voltages and acceptor–donor distributions. (From Sze, S.M., *Physics of Semiconductor Devices*, 2nd Edn., John Wiley & Sons, New York, 1981. With permission.)

where μ_n is electron drift mobility. In equilibrium, the total electron (and hole) current must be zero. The previous equation provides

$$en\mu_n \frac{dV}{dx} = eD_n \frac{dn}{dx} \tag{8.83}$$

Integrating both sides from $-\infty$ (the p-side where the voltage is V_p) to $+\infty$ (the n-side where the voltage is $V_n = 0$), we find

$$\int_{-\infty}^{\infty} e\mu_n \frac{dV}{dx} dx = \int_{-\infty}^{\infty} eD_n \frac{1}{n} \frac{dn}{dx} dx \rightarrow \int_{V_p}^{V_n} e\mu_n dV = \int_{n_p}^{n_n} eD_n \frac{dn}{n} \quad (8.84)$$

where n_p , n_n are the electron densities in the p-type and n-type material, respectively, far away from the junction. The integration gives us

$$e\mu_n(V_p - V_n) = eD_n \text{Ln}\left(\frac{n_p}{n_n}\right) \quad (8.85)$$

The electron density far into the n-type material is approximately equal to the donor density

$$n_n = N_d \quad (8.86)$$

Far to the left on the p-side, the density of holes is approximately equal to the acceptor density N_a so that the electron density can be found from the law of mass action to be

$$n = n_i^2/p = n_i^2/N_a \quad (8.87)$$

Therefore, Equation 8.85 can be written as

$$V_b = V_p - V_n = \frac{D_n}{\mu_n} \text{Ln}\left(\frac{n_i^2}{N_d N_a}\right) \quad (8.88)$$

8.8.3 JUNCTION FIELDS

The carrier densities are

$$n(x) = N_c e^{\frac{[E_c - eV_b(x)] - E_f}{kT}} \quad qp(x) = N_v e^{\frac{E_f - [E_v - eV_b(x)]}{kT}} \quad (8.89)$$

The potential can be determined self-consistently. Assume that the Fermi level is flat across the two regions. Assume that the junction is at $x = 0$ with p-type on $x < 0$ and n-type in the $x > 0$ region. For notation, $V_b(x)$ is the built-in field as a function of x . The symbol V_b means $V_b \equiv V_p(-\infty) - V_n(+\infty)$. The voltage at $x = +\infty$ is $V_n = 0$ since the n-side is tied to ground. The voltage at $x = -\infty$ is $V_b(-\infty) \equiv V_p = V_b$ which is the built-in voltage. If we do not tie the n-side to ground, we have $V_b = V_p - V_n$. Assume far from the junction that

$$n(+\infty) = N_d = N_c e^{\frac{[E_c - eV_b(+\infty)] - E_f}{kT}} = N_c e^{\frac{(E_c - eV_n) - E_f}{kT}} = N_c e^{\frac{E_c - E_f}{kT}} \quad (8.90a)$$

and

$$p(-\infty) = N_a = N_v e^{\frac{E_f - [E_v - eV_b(-\infty)]}{kT}} = N_v e^{\frac{E_f - [E_v - eV_p]}{kT}} = N_v e^{\frac{E_f - [E_v - eV_b]}{kT}} \quad (8.90b)$$

where N_d and N_a are the density of donors and acceptors.

Solving Equations 8.89 and 8.90 provides

$$-eV_b = E_g + kT \text{Ln}\left[\frac{N_d N_a}{N_c N_v}\right] \quad (8.91)$$

where

N_d and N_a are the donor and acceptor densities

N_c and N_v are the conduction band and valence band effective density of states

E_g is the gap energy where $E_g = E_c - E_v$

Equations 8.89 and 8.90 can be combined to yield

$$n(x) = N_d e^{-\frac{eV_n - eV_b(x)}{kT}} \quad p(x) = N_a e^{-\frac{[eV_b(x) - eV_p]}{kT}} \quad (8.92)$$

We find the potential $V_b(x)$ using Poisson's Equation

$$\frac{d^2}{dx^2} V_b = -\frac{\rho(x)}{\epsilon} \quad (8.93)$$

where ϵ is the dielectric constant and ρ is the charge density. The charge density is given by

$$\rho(x) = e[N_d(x) - N_a(x) - n_c(x) + p_v(x)] \quad (8.94)$$

To simplify matters, assume the field changes over the “junction region” $x = -d_p$ to $x = +d_n$. Now, using the fact that

$$V_b(x < -d_p) \cong V_b(-\infty) = V_p \quad V_b(x > d_n) \cong V_b(\infty) = V_n \quad (8.95)$$

Equations 8.92 provide

$$p(x) = N_a \quad x < -d_p \quad n(x) = N_d \quad x > d_n \quad (8.96)$$

Also, in the p-type semiconductor for $x < -d_p$, the number of holes is negligible by the law of mass action; similar comments apply to the holes in the n-type material. Therefore, the charge density is zero for the regions outside the junction region and cannot give rise to any electric field.

$$\rho = 0 \quad \text{for } x < -d_p \text{ and } x > d_n \quad (8.97)$$

The electric field is confined to the junction region.

We still need to calculate the charge density within the junction region. Inside the junction region $-d_p < x < d_n$ the voltage varies by many orders of magnitude from kT , that is

$$e|V_b(x) - V_{n,p}| \gg kT \quad (8.98)$$

Equation 8.90 then provides

$$n = 0 = p \quad -d_p < x < d_n \quad (8.99)$$

Therefore in the junction region $-d_p < x < d_n$ the charge density in Equation 8.94 is

$$\rho(x) = e[N_d(x) - N_a(x)] \quad -d_p < x < d_n \quad (8.100)$$

Note that

$$N_d(x) = 0 \quad x < 0 \quad \text{and} \quad N_a(x) = 0 \quad x > 0 \quad (8.101)$$

Therefore, Poisson's Equation 8.93 provides

$$\frac{d^2}{dx^2} V_b = \begin{cases} 0 & d_n < x \\ -\frac{eN_d}{\epsilon} & 0 < x < d_n \\ \frac{eN_a}{\epsilon} & -d_p < x < 0 \\ 0 & x < -d_p \end{cases} \quad (8.102)$$

Equations 8.102 can be integrated twice to give

$$V_b = \begin{cases} V_n & d_n < x \\ V_n - \left(\frac{eN_d}{2\epsilon}\right)(x - d_n)^2 & 0 < x < d_n \\ V_p + \left(\frac{eN_a}{2\epsilon}\right)(x + d_p)^2 & -d_p < x < 0 \\ V_p & x < -d_p \end{cases} \quad (8.103)$$

where V_n, V_p, d_n, d_p are the integration constants. One of the constants d_n, d_p can be eliminated by requiring dV_b/dx to be continuous at $x=0$, which gives

$$N_d d_n = N_a d_p \quad (8.104)$$

Equation 8.104 gives the total charge on either side of the junction (multiply each side by the cross sectional area to see this). Also, continuity of $V_b(x)$ at $x=0$ gives

$$V_b = V_p - V_n = \left(\frac{e}{2\epsilon}\right) \left(N_d d_n^2 + N_a d_p^2\right) \quad (8.105)$$

Combining Equations 8.104 and 8.105, we find

$$d_{n,p} = \left\{ \frac{(N_a/N_d)^{\pm 1}}{(N_a + N_d)} \frac{2V_b}{e} \right\}^{1/2} \quad (8.106)$$

The thickness $d_n + d_p$ delineates the width of the space charge region, also known as the depletion region.

8.9 REVIEW EXERCISES

- 8.1** Complete the derivation for the statistics of the dopant states. Include the constraints for all electrons and the total energy. Include all electrons in the variation and not just those occupying donor sites.
- 8.2** Suppose $f(x, y) = (x - 1)^2 + y^2$ subject to the constraint $C(x, y) = x + y = 1$. Find the points where f is minimum using the method of Lagrange multipliers.
- 8.3** An engineering student plans to grow semiconductor material in a homemade growth chamber. This particular material has hole mobility μ_p and electron mobility μ_n such that $\mu_p < \mu_n$. The student has an idea to reduce the conductivity of the material by slightly doping the material p-type.
 - a.** Write the conductivity in terms of the hole density p and the two mobilities (along with some constants).
 - b.** Find the hole density that gives the minimum conductivity.
 - c.** Find the minimum value of the conductivity.

8.4 Suppose N electrons can only have spin up or down (i.e., pointing along the positive or negative z -axis). Apply a magnetic field $\vec{B} = B_0 \hat{z}$ along the z -axis. Assume the interaction energy $E = c \vec{S} \cdot \vec{B}$ where $\vec{S} = \pm \frac{1}{2} \hat{z}$ depending on the spin direction.

- Find the separation in energy between spin up and spin down for a single electron.
- If the N electrons are in thermal equilibrium according to a Boltzmann distribution, find the average number in the lower level and the average number in the upper level.

8.5 Suppose phonons obey the Bose–Einstein distribution

$$f_{\text{BE}}(E) = \frac{1}{\exp\left(\frac{\hbar\omega}{kT}\right) - 1} = \frac{1}{e^{\frac{E}{kT}} - 1}$$

- For $kT = 0.1$ and 1 eV plot a graph of the distribution vs. energy. Recall that any number of Bose particles can occupy a single state. In this case, the distribution gives the average number of particles in a given state, which is different from a probability.
- Based on the graph in Part a, what does the distribution look like for $T \sim 0$.

8.6 Starting with the number of combinations given in Appendix J

$$W = \prod_i \frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!}$$

show the Bose–Einstein distribution

$$f_{\text{BE}}(E) = \frac{\bar{n}_i}{g_i} = \frac{1}{z^{-1} \exp[\beta E] - 1} \quad \text{or} \quad f_{\text{BE}}(E) = \frac{1}{\exp\left(\frac{\hbar\omega}{k_b T}\right) - 1} = \frac{1}{e^{\frac{E}{k_b T}} - 1}$$

8.7 For the density operator given in Appendix K

$$\hat{\rho}_r = \frac{1}{Z} \exp\left(-\frac{\hat{H}_r}{k_B T}\right)$$

Find the average $\langle \hat{H}_r \rangle$ for a system with distinct energy levels $E_1 < E_2 < E_3 < \dots$

8.8 An electron moves under an impressed electric field through a material.

- Demonstrate the formula for electron mobility

$$\mu = \frac{q}{2m_e} \frac{1}{(c_d N_d + c_p N_p)}$$

where

N_d and N_p represent the density of phonons and crystal defects
 c_d and c_p are constants

Hint: Assume the rate of collisions must be proportional to the density of defects, etc.

- At temperature T , what is the density of phonons (number per volume)? For simplicity, use only acoustic phonons for a monatomic 3-D crystal. Allow only one mode for each of the three directions. Assume a Bose–Einstein distribution given by

$$f_{\text{BE}}(E) = \frac{1}{\exp\left(\frac{\hbar\omega}{kT}\right) - 1} = \frac{1}{e^{\frac{E}{kT}} - 1}$$

- Predict the mobility as a function of temperature.

- 8.9** Suppose an electron sits in a trap at an energy $\Delta E \gg kT$ below the conduction band. Assume the number of collision per second of phonons with the electron must be proportional to the density of phonons $N_p(E)$ (#/Vol/energy). Therefore, the time between collisions must be the reciprocal of the collision rate. Find the length of time required for the electron to wait before a phonon releases it from the trap. Hint: only incident phonons with energy greater than ΔE can release the electron.

REFERENCES AND FURTHER READINGS

General References

1. Brennan K.F., *The Physics of Semiconductors with Applications to Optoelectronic Devices*, Cambridge University Press, Cambridge, U.K. (1999).
2. Bhattacharya P., *Semiconductor Optoelectronic Devices*, 2nd ed., Prentice Hall, Upper Saddle River, NJ (1997). Good general reference on most aspects of solid state including fabrication, electronic processes, bands, junctions, and optoelectronic devices.
3. Pierret R.F., *Advanced Semiconductor Fundamentals*, Volume VI, *Modular Series on Solid State Devices*, R.F. Pierret and G.W. Neudeck, eds., Addison Wesley Publishing, Reading, MA (1989). Very thin and readable text.
4. Rosenberg H.M., *The Solid State: An Introduction to the Physics of Crystals for Students of Physics, Materials Science, and Engineering*, 2nd ed., Clarendon Press, Oxford, U.K. (1984).
5. Fraser D.A., *The Physics of Semiconductor Devices*, 3rd ed., Clarendon Press, Oxford, U.K. (1985).
6. Blakemore J.S., *Solid State Physics*, 2nd ed., W.B. Saunders Company, Philadelphia, PA (1974).
7. Ashcroft N.W. and Mermin N.D., *Solid State Physics*, Holt, Rinehart & Winston, New York (1976).
8. Kittel C., *Introduction to Solid State Physics*, 5th ed., John Wiley & Sons, New York (1976).

Statistical Mechanics

9. Reif F., *Statistical Physics, Berkeley Physics Course*, Volume 5, McGraw-Hill Book Company, New York (1965).
10. Pathria R.K., *Statistical Mechanics, International Series in Natural Philosophy*, Volume 45, Butterworth-Heinemann Ltd., Oxford. First printing 1972 and reprinted through 1995. This is one of the most readable treatments.
11. Datta S., *Quantum Transport: Atom to Transistor*, Cambridge University Press, Cambridge, U.K. (2005).
12. Gasser R.P.H. and Richards W.G., *An Introduction to Statistical Thermodynamics*, World Scientific, River Edge, NJ (1995).
13. Tolman R.C., *The Principles of Statistical Mechanics*, Dover Publications Inc., New York (1979).
14. Hill T.L., *An Introduction to Statistical Thermodynamics*, Dover Publication Inc., New York (1986).
15. Huang K., *Statistical Mechanics*, John Wiley & Sons, Inc., New York (1963).

Doping and Diodes

16. Sze S.M., *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, New York (1981).
17. Tang C.L., *Fundamentals of Quantum Mechanics for Solid State Electronics and Optics*, Cambridge University Press, Cambridge, U.K. (2005).
18. Shur M., *GaAs Circuits and Devices*, Plenum Press, New York (1987).
19. Klingshirn C.F., *Semiconductor Optics*, Springer, New York (1997).

Lagrange Multipliers

20. Arfken G., *Mathematical Methods for Physicists*, 2nd ed., Academic Press, New York (1970).
21. Cushing J.T., *Applied Analytical Mathematics for Physical Scientists*, John Wiley & Sons, Inc., New York (1975).
22. Dahlquist G. and Bjorck A., *Numerical Methods*, Dover Publications, Inc., Mineola (2003).
23. Byron F.W. Jr. and Fuller R.W., *Mathematics of Classical and Quantum Physics*, Dover Publications, Inc., New York (1992).

Appendix A: Growth and Fabrication Methods

The present appendix discusses the fundamentals of GaAs growth and provides example fabrications procedures.

A.1 INTRODUCTION TO EQUIPMENT

The present section details some of the typical apparatus used to fabricate devices (such as lasers) in heterostructure and the clean room equipment used to pattern the wafers (i.e., fabricate devices). The fabrication and growth of heterostructure has several phases. Once the design for the material is finished, molecular beam epitaxy (MBE) or metal organic chemical vapor deposition (MOCVD) most typically grows the heterostructure. However, to a lesser extent, liquid phase epitaxy might be employed. As the semiconductor materials are readied, the designer makes explicit drawings of the integrated optoelectronic circuits on computer-aided design (CAD) software. The CAD process requires the broadest knowledge of the fabrication sequence usually manifested through “design rules.”

As discussed in the next section, a typical ridge-guided laser requires many fabrication steps. The CAD drawings are transferred to quartz plates (termed masks or reticles) for the photolithography. Reactive ion etchers (RIEs), chemically assisted ion beam etchers (CAIBE), electron cyclotron resonance (ECR) etchers are all equipment used to remove material from the semiconductor wafer to form mirrors, waveguides, and other components. Usually the CAIBE and ECR are the preferred equipment for preparing flat, vertical surfaces because of the high degree of etch anisotropy (see Figure A.1). Sometimes wet chemical etches (dipping the wafer in an acid-oxidizer solution) can provide the desired sloping surfaces. However, recent progress shows that wet etching augmented by optical illumination can produce sidewalls with angles ranging from 0° to 90° and have successfully produced laser mirrors (refer to the Yi and Parker publications). Focused ion beam (FIB) etchers also have uses. Metal electrical contacts are evaporated onto the wafer using either thermal or electron-beam evaporators. Layers intended for electrical isolation (i.e., dielectric layers) can be deposited by plasma-enhanced chemical vapor deposition (PECVD) or spun-on to the wafer. The spin-on process consists of placing a drop of liquid onto the wafer surface and then spinning the wafer at several thousand RPM to evenly distribute the liquid. The spin-on process is appropriate for liquid plastics like polyimide.

The next few sections introduce some of the equipment used to grow the semiconductor material and fabricate the devices. The subsequent section discusses some of the processing steps required.

A.1.1 MOLECULAR BEAM EPITAXY

MBE is a method of using material from multiple sources and depositing it onto a wafer as shown schematically by Figure A.2. The molecules travel in straight lines from the source to the target through the highly evacuated (ultrahigh vacuum—UHV) chamber. The wafer epitaxially grows as the molecules deposit on the wafer (i.e., grows as a single crystal). The heated effusion cells evaporate highly purified materials. When the shutters open, the evaporated materials can travel to the substrate. Most often the substrate rotates to ensure uniform deposition. The quality of the crystal depends critically on the growth temperature (i.e., the temperature of the rotating substrate). Good quality GaAs requires around 640°C whereas “low-temperature” GaAs (noncrystalline) grows at temperatures on the order of 200°C . The high temperatures allow the deposited molecules

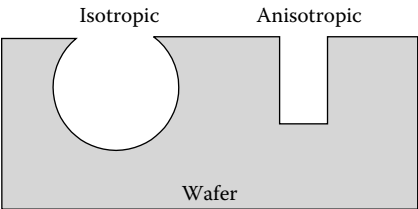


FIGURE A.1 Isotropic and anisotropic etches.

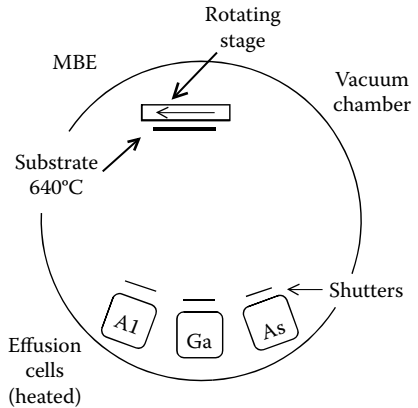


FIGURE A.2 Block representation of the MBE system.

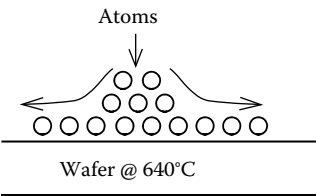


FIGURE A.3 Atoms easily diffuse at high temperatures.

to diffuse to the proper locations on the wafer (see Figure A.3) so as to produce a crystal. The quality of the crystal can be monitored as it grows by using a RHEED monitor. The RHEED system analyzes the electron diffraction patterns obtained by reflecting electrons from the growing layers. Good interference patterns indicate highly crystalline growth.

A.1.2 REACTIVE ION ETCHER AND PLASMA-ENHANCED CHEMICAL VAPOR DEPOSITION

The reactive ion etcher (RIE) and the plasma-enhanced chemical vapor deposition (PECVD) have very similar construction. As previously mentioned, the RIE units remove material from the grown wafer (the reverse of growth) whereas the PECVD is a method of growing materials (usually noncrystalline). Figure A.4 shows a typical block diagram.

The wafer is mounted in a vacuum chamber between two capacitor plates. Vacuum pumps evacuate the chamber while gases are allowed to flow. An RF generator (typically 13 MHz at 100 W) excites and decomposes the gas. The gases either selectively remove or deposit the desired materials (depending on the type of gasses). The quality of the grown films or the etch rate depends on the temperature of the substrate.

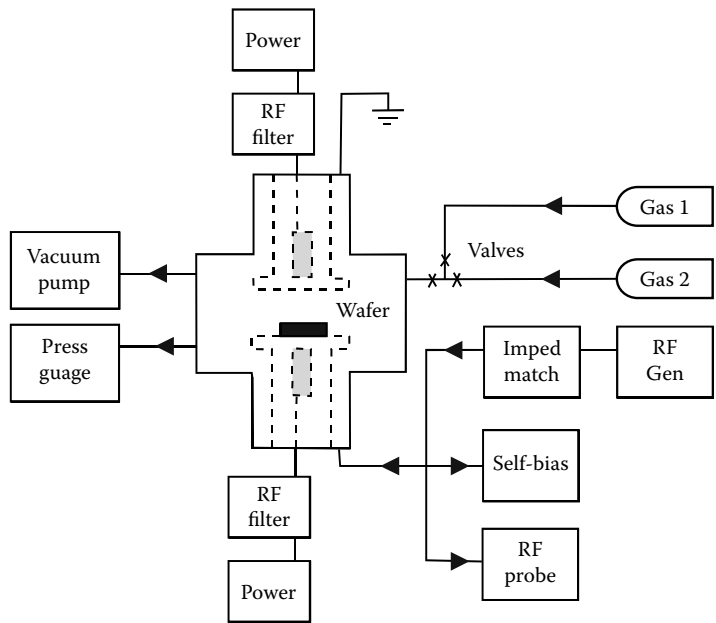


FIGURE A.4 Block diagram of RIE and PECVD systems.

The self-bias probe is essential. The ionized gases actually tend to “rectify” the RF fields. As a result, a DC bias (self-bias) develops across the two capacitor plates. The self-bias fields direct the ionized gas molecules toward the electrode with the wafer. The RF bias is a good measure of etch or deposition rate. As a note, it is possible to shine a laser beam on heterostructure and monitor the etch or deposition rates (refer to Parker and Yi publications and their references) by monitoring interference fringes.

A.1.3 THE ELECTRON CYCLOTRON RESONANT ETCHER

The ECR etcher has a low-pressure chamber with a chuck to hold a wafer for etching. Figure A.5 shows a block diagram. A microwave generator and a set of upper magnets excite and contain a

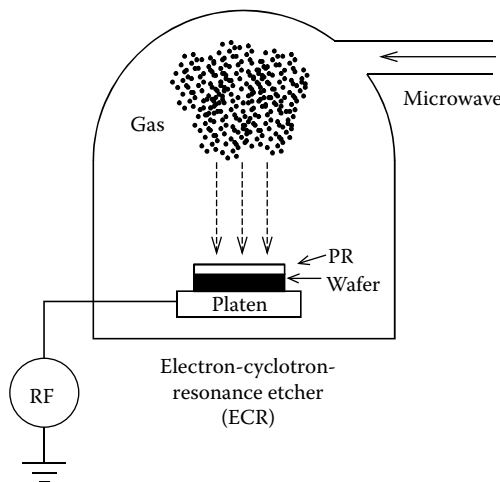


FIGURE A.5 Block diagram of the electron-cyclotron-resonant etcher.

plasma of reactive gasses above the wafer. A set of lower magnets focuses the ions near the wafer surface. A 13.566 MHz RF signal applied to the wafer chuck produces the self-bias potential that accelerates the ions toward the wafer. Unlike the RIE, it is possible to independently adjust the gas pressure and flow rate, the plasma density, the focus, and the accelerating potential. The ECR produces excellent quality sidewalls (highly anisotropic etches).

A.1.4 EVAPORATORS

In order to place metals onto the wafer, a metal evaporator system must be employed. There are thermal and electron-beam (e-beam) evaporators. The thermal evaporators are less expensive but can only be used for metals with lower evaporation temperatures. The e-beam evaporators can be used for all metals.

Figure A.6 shows a diagram of the thermal evaporator. The wafer hangs in such a way that the surface to be metallized faces the source. The source consists of a piece of metal placed in a “wire basket” (sometimes called a “boat”) made of multiple turns of tungsten wire. The chamber is evacuated. High amperage current through the wire basket melts and evaporates the metal flakes. The evaporated atoms travel in straight lines through the vacuum to the wafer and stick. Sometimes the wafer rotates to increase the uniformity of the metal film. The wafer can also be angled so only certain sidewalls receive the metal. Unfortunately, the tungsten wire in the basket tends to evaporate near the evaporation temperature of some metals. As a result, the e-beam evaporator is used instead.

As a metal layer grows on the wafer, the thickness of the layer is monitored using a crystal (part of a resonant circuit) which is placed next to the wafer. As metal accumulates, the resonant frequency of the crystal changes (mass increases) as an indication of the layer thickness.

The e-beam evaporator also deposits metal on a wafer and has a bell geometry similar to the thermal evaporator (see Figure A.7). The e-beam transfers its kinetic energy to the metal so as to melt and evaporate it. The metal is held in a ceramic crucible. This unit is primarily used to evaporate “hard” metals such as titanium, platinum, tungsten, etc.

A.2 TYPICAL FABRICATION STEPS FOR GAAS OPTICAL CIRCUITS

This chapter discusses the design and fabrication of an example GaAs-based laser. In the first section, an overall view is given of the fabrication process and the CAD phase of the design. The remaining sections provide fabrication data for the clean room phase. One should be aware that

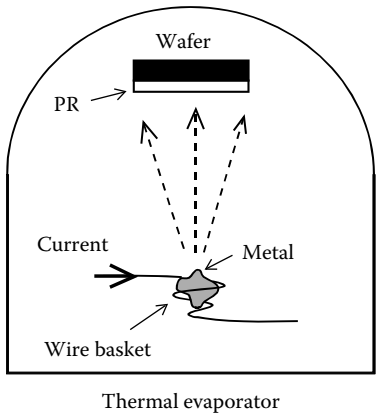


FIGURE A.6 Thermal evaporator.

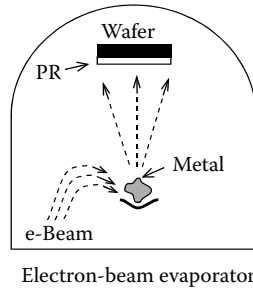


FIGURE A.7 The e-beam evaporator.

the sequence is only one example out of many possible process sequences. The author highly recommends the Ralph E. Williams classic book on the fabrication of GaAs devices.

A.2.1 THE CAD PHASE OF DESIGN

The fabrication process for the semiconductor GaAs lasers has four phases for this example: (1) design and growth of the laser heterostructure, (2) design of the devices using CAD software, (3) fabrication of the wafer using clean room facilities, and (4) postprocessing such as cleaving and mounting the devices. Obviously, the CAD design must take into account the other three phases of the fabrication process. The heterostructure design, for example, influences the number and type of masking steps drawn on the CAD. The equipment and processing steps in phase 3 determine the tolerance in the overlap of mask patterns as well as the number and type of such patterns. The number and placement of the wafer cleaves along with the electrical contact points used in phase 4 must be drawn into the CAD diagram. Thus the CAD phase presumes knowledge of the entire fabrication process as well as the operational theory of the devices.

To fabricate GaAs laser devices, either four or five masking layers can be used depending on whether n-type or semi-insulating substrates are used, respectively. Suppose the example device is an etched-ridge waveguide laser with etched mirrors and the heterostructure is grown on an n-type substrate (see Figure A.8). There are four masking levels: (1) p-contact, (2) deep etch, (3) via, (4) top metal. The fabrication sequence appears in Figure A.8. Subfigures C, E, G, and H show the CAD sequence while B, D, and F show the physical results after the corresponding fabrication. The first masking level defines the Ohmic contacts to the top p-type material and the waveguides. The CAD diagram appears as in Figure A.8C. Photolithography and metallization steps produce the physical structure shown in Figure A.8D. Next, the areas for deep etching are drawn on the CAD as shown in Figure A.8E. The etch forms two separate regions during fabrication as depicted in Figure A.8F. One region of the wafer is etched below the active layer so as to form the mirror surfaces at the ends of the waveguide. The other region is a shallow etch that defines the ridge for the waveguides. The shallow etch appears everywhere on the surface of the wafer except where the metal masks the surface. After the deep etches, oxygen can be implanted for electrical isolation and a layer of polyimide can be applied across the entire topside. The polyimide improves the coupling between waveguides in the actual circuits and also provides electrical isolation for metal pads. Holes must be opened up in the polyimide above the waveguides for electrical connections as shown in Figure A.8G for the vias. Finally, a top-metal mask defines the contact pads (Figure A.8H). The metal covers portions of the polyimide and makes contact with the exposed waveguides.

Electrical crossovers can be made in the integrated circuit using the polyimide as the insulator between the two electrical traces. As a final note, the CAD design for semi-insulating substrates uses a fifth masking level for a wet etch to the n-type contact.

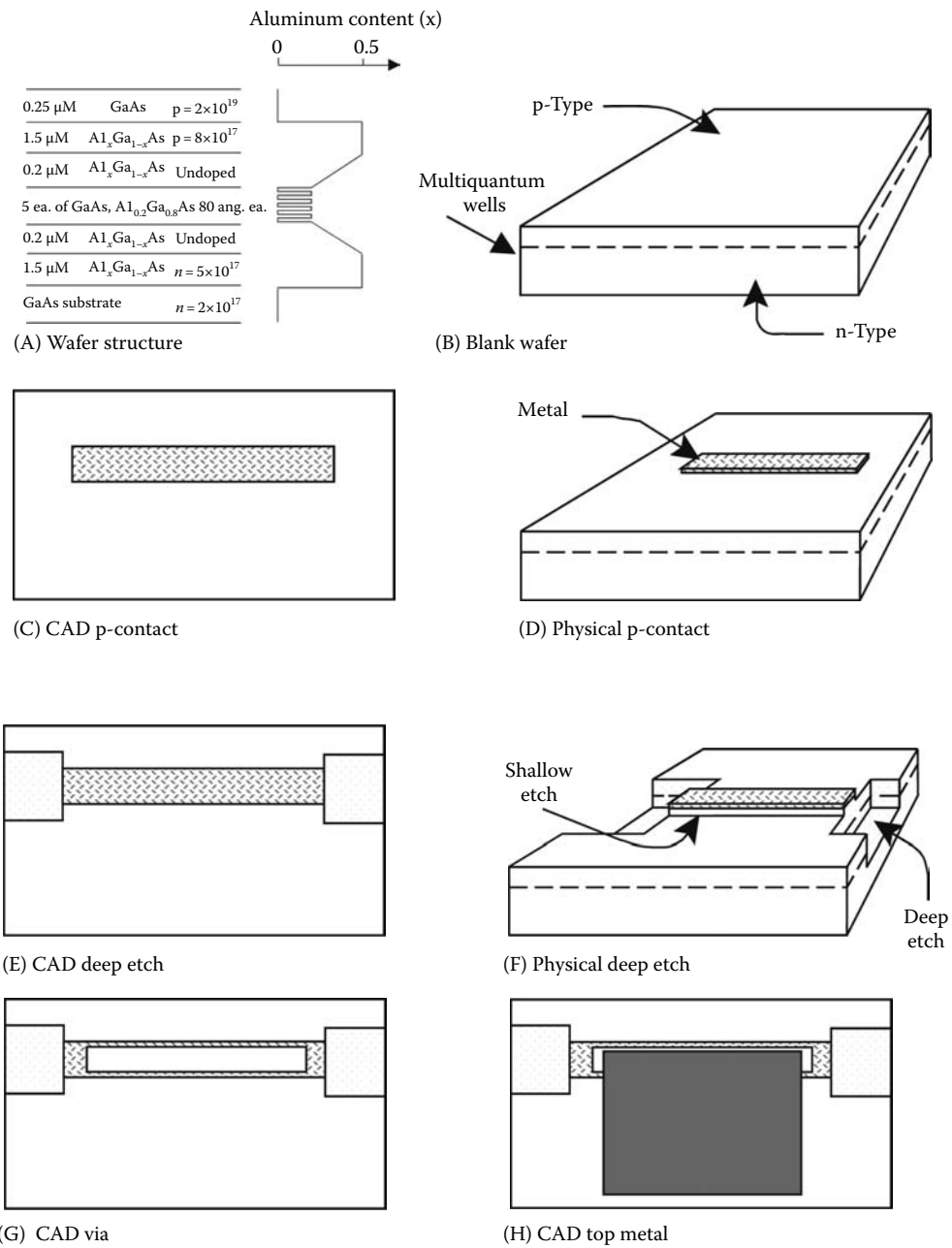


FIGURE A.8 The processing sequence: subfigures C, E, G, H show the CAD design and subfigures B, D, F show the results of the processing. Subfigure A shows the wafer structure.

A.2.2 WAFER CLEAVING AND CLEANING

After scribing and cleaving the epitaxially grown MQW wafer to a manageable size (typically 1 cm^2), the sample is cleaned with a specific sequence of solvents: trichloroethane (TCE), acetone, methanol, deionized water, and finally isopropanol. The wafer is first rinsed with TCE, which removes grease from the wafer surface. Acetone is the next solvent used; it dissolves organic compounds, but leaves a residue behind as it dries. Methanol is therefore used to rinse off the acetone, but it also

leaves a film behind. Deionized water is used to drive off the methanol. The final rinse is with isopropanol, which dries in sheets, rather than droplets. The isopropanol is blown off the wafer surface with a nitrogen pressure gun, leaving a surface clean of dust, and organic contaminants. Photolithography and material deposition require baking the wafer above the vapor point of the solvents to remove any residual moisture. The wafer is placed in a convection oven at 100°C–150°C for at least 30 min and commonly up to 8 h.

A.2.3 PHOTOLITHOGRAPHY AND IMAGE REVERSAL

The CAD designs are transferred to the wafer using photolithography. Photolithography using UV light can produce feature sizes slightly smaller than 0.5–0.7 μm . X-ray lithography can produce still smaller sizes. e-Beam lithography uses electrons rather than photons for feature sizes as small as 100 Å. The photolithography process consists of coating a semiconductor wafer with photoresist (spinning it on) and exposing it through a set of quartz plates (mask plates) with chrome patterns matching the CAD drawings. The mask plates are made directly from the CAD diagrams. Figure A.9 shows the general arrangement.

The photolithography can be used in a positive or negative mode; that is, the photoresist (PR) will remain after developing where the resist was exposed to light or where it was not exposed to light, respectively. The light either breaks the link between PR or cross-links them. Some devices or components, such as electrical contacts, require a negative process. The applicability of positive or negative photolithography really comes down to whether or not a “liftoff” process will be used. A liftoff process (refer to Figure A.10 and the next paragraph) is one where the photoresist is developed (thereby exposing portions of the wafer to air while other portions remain protected by resist), another layer (such as a metal or dielectric) is deposited on top (overexposed and protected regions alike), and then the wafer is placed in a solution to remove the resist. As a result, the extra layer remains wherever the photoresist has been removed by the developing process.

An image reversal process is performed on the exposed photoresist in preparation for liftoff. If the photoresist were to be developed immediately following exposure (the normal photolithography process), the sidewalls of the pattern would be vertical (or shallowly sloped), as shown in Figure A.10A. This profile is not acceptable for metal liftoff, as it allows the metal to cover the wafer in a continuous film. During liftoff, this tends to cause peeling of the metal from surfaces where adhesion is intended. The image reversal process causes an undercut slope in the photoresist sidewalls, as shown in Figure A.10B, resulting in better metal liftoff. The image reversal is accomplished by performing an ammonia diffusion into the photoresist, followed by a 90 s ultraviolet flood exposure of the entire wafer surface. The ammonia diffusion causes a chemical reaction in the photoresist (molecular cross-linking), hardening the areas which have been previously exposed, and leaving unaffected the unexposed areas. The flood exposure has no effect on the hardened resist, but exposes the remaining wafer area, which can be subsequently removed by developer.

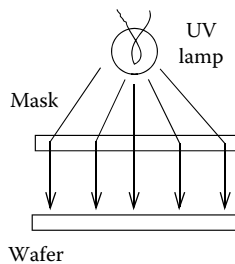


FIGURE A.9 Photolithography transfers the mask pattern to the wafer.

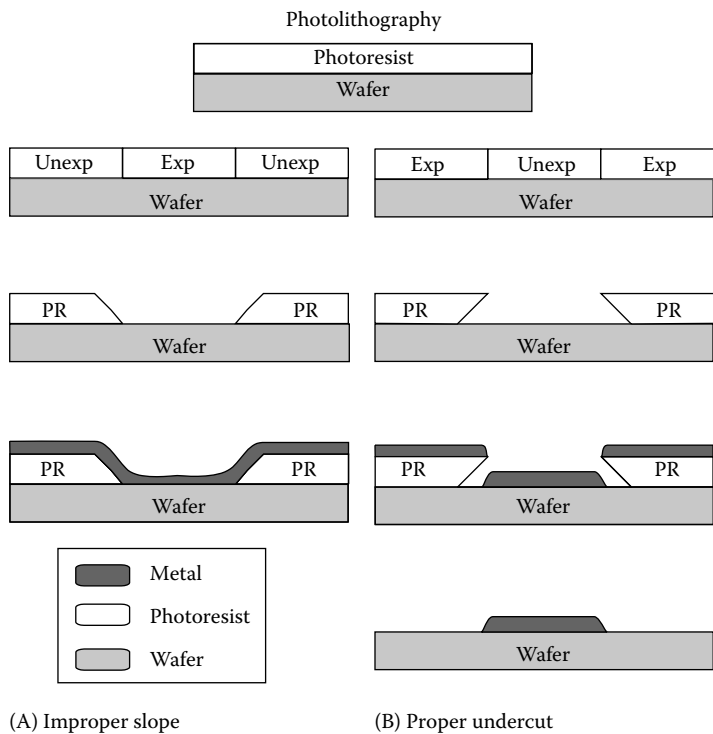


FIGURE A.10 The photolithography process. The left side (A) shows a view of the results for a normal photolithography process, while the right side shows the image-reversed photolithography process. The metal liftoff works more efficiently with the proper undercut (B) than it does with the process depicted on the left side. [Exp = exposed, Unexp = unexposed].

A.2.4 A CASE STUDY OF ETCHING

As previously mentioned, the etching removes material from the wafer. For waveguides or etched mirrors, the material must be removed from select portions of the wafer. The selected regions are defined by masking the wafer using either PR or other masking material (such as an oxide or silicon nitride) and then patterning the layer using photoresist. The wafer will be etched wherever it is exposed (unprotected by photoresist or other mask). The masking material is selected to be inert (or nearly inert) to the etching.

Consider an example using a chemically-assisted ion beam etcher (CAIBE) etcher from the early 1990s. The process begins by coating the entire wafer surface with an insulating layer of SiO₂. Chrome is used as the etch mask. Windows are then etched in this layer for p-contacts and mirror facets. After metallization, the CAIBE is used to perform a vertical etch. The chrome etch rate in the CAIBE is much less than the etch rate of SiO₂, which is much less than that of GaAs. A single CAIBE etch could then produce a three-level topography. The CAIBE uses both a chemical etching through the use of chlorine as a gas and also a mechanical etch through the impact of argon and other molecules on the surface of the wafer.

This process was advantageous for simple geometries and devices, because the CAIBE etching can be performed in a single step. There are a few disadvantages to this method, however, as the device complexity increased. The first difficulty was alignment, since the processing sequence requires tight tolerances in overlaying the metal waveguides on top of the SiO₂ windows. Second, the CAIBE etch rate can widely vary, and it is difficult to predict the required etch duration to achieve the necessary depths. This lack of consistency can usually be compensated

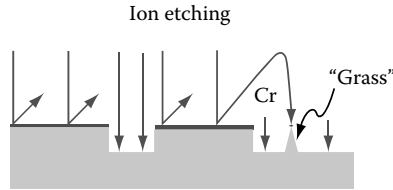


FIGURE A.11 The grass is growing: ions striking the metal cause clumps of metal to jump off and deposit on the wafer.

by first etching a blank wafer as a calibration die. Third, in the absence of a neutralizer filament in the CAIBE, electrical arcing becomes a problem on the wafer surface. Surface charge can build up on the sample as it is bombarded with ions, since the SiO_2 was restricting charge dissipation through the wafer. This surface charge collects at windows in the SiO_2 (mirror facets), and arcs to the GaAs substrate, destroying the wafer surface. These problems have significant consequences, since the damage to the wafer occurs after the majority of the fabrication had already been completed. To bypass these processing problems, a new CAIBE sequence using different masking materials can be used.

Another process can be devised as a method for fabricating devices on semi-insulating substrates with contacts to both the p- and n-doped epilayers. It can be used for n-type substrate devices as well. It solves the alignment problem by putting down the chrome metal layer first. The deep-etch mask is then aligned to the metal, a step which does not require a time-consuming image reversal, and does not require such tight tolerances. The CAIBE etch rate variation is corrected by replacing the single-step CAIBE procedure with a two-step process, whereby the first etch can be used to calibrate the required duration of the final etch. The arcing problem can be averted by switching from an insulating SiO_2 layer to a photoresist etch mask with enhanced conduction to ground or the wafer body. Finally, the fabrication sequence should be changed so that the CAIBE etch comes earlier in the overall fabrication process; as a result, any wafer damage in the CAIBE is less costly in terms of time and expense.

There is still a drawback to using the CAIBE. Under certain circumstances, "grass" grows on the wafer. Figure A.11 shows an example of "grass" (consisting of a large number of spiked objects). During the etch, clumps of chrome are knocked off the metal strips by the colliding ions. Some clumps adhere to the wafer and form a mask. Then incident reactive ions etch away the GaAs everywhere except at the places where the clumps of Cr adhere.

From the author's point of view, the ECR etchers are easier to use and have better control over the parameters such as RF self-bias. The next section briefly discusses the ECR and its etch monitor.

A.2.5 REVISITING THE ELECTRON CYCLOTRON RESONANT ETCHER

High quality etched laser mirrors remain a priority for monolithic photonic device integration. While cleaved facets provide consistent, high quality in-plane laser mirrors, cleaves can only be made at the edges of the die, where the output light exits. Cleaves cannot be used for integrated laser mirrors for which the output light remains within the monolithic chip. In this letter, we report on the fabrication and characterization of laser mirrors in GaAs using an ECR etcher. The ECR etcher offers several advantages over competing CAIBE and RIE for fabricating etched mirrors including a higher etch rate, control over more adjustable parameters, and higher etch selectivity between GaAs and the etch mask. These are important consideration for optical devices, for example, where the optical scattering and reflectivity of the laser mirrors are direct measures of the quality of the etch.

The ECR etcher has a low-pressure chamber with a chuck to hold a wafer for etching. A microwave generator and a set of upper magnets excites and contains a plasma in reactive gasses above the wafer. A set of lower magnets focuses the ions near the wafer surface. A 13.566 MHz RF signal applied to

the wafer chuck produces the self-bias potential that accelerates the ions toward the wafer. It is possible to adjust the gas pressure and flow rate, the plasma density, the focus, and the accelerating potential independently. The microwave forward power can be 400 W or more and the RF forward power is 80 W or more. The upper and lower magnets on the etcher are typically set to 16 and 20 A, respectively. A low-pressure helium source pushes the rim of a 3 in. silicon wafer up against temperature-controlled cooling stage. An o-ring prevents the He backpressure from escaping into the chamber. A masked wafer for etching (1 cm^2) can be mounted on top of the silicon wafer using resist. The silicon wafer is then place into the etcher through a load lock.

The vertical surface of an etched mirror (for example) cannot be made smoother than the edge of the metal mask delineating the mirror etch. For this reason, minimizing the grain size of the metal is important for obtaining high quality mirrors. The metals are deposited using an electron-beam evaporator. Cooling the substrate during metal deposition can reduce the grain sizes by up to a factor of 10 and smooth the edges of the etch mask. The improved quality of the edges is significant for an 850 nm laser with an effective wavelength in the semiconductor of $\sim 240\text{ nm}$.

Dry etching of semiconductor heterostructures is common for fabricating optoelectronic components such as semiconductor lasers and transistors. However, it is difficult to accurately and repeatably etch to a specific depth. Several methods are available to achieve the desired etch profile. The most common method is to calibrate the time required to etch a specific depth into the material, using a sacrificial test sample with an identical material structure. However, the etch parameters must be tightly controlled for subsequent etches and the wafer structure must be known a priori. This knowledge cannot be guaranteed without extensive wafer testing since the thickness tolerance during wafer growth can be as high as 20%. In some cases, it is possible to use an optical or a mass spectrometer to monitor the gas species in the etch-chamber as the etch progresses, determining etch depth by the sudden appearance of etch-products from a specific heterostructure layer of differing composition. Alternatively, thin etch-stop layers are sometimes added to the heterostructure, and the proper combination of reactant gasses is used to preferentially etch the heterostructure. The addition of etch-stop layers into the heterostructure, however, fixes the etch depth at specific values, eliminating the flexibility to optimize device performance.

A laser reflectometer can be used as an in situ monitor for ECR etching. The technique can be applied to a GaAs–AlGaAs laser heterostructure (see Figure A.12) to ascertain (1) the etch depth, (2) the etch rate, (3) the quality of the postetch surface, and also to (4) determine or verify the composition of a wafer without knowledge of its precise structure (i.e., “reverse engineering”).

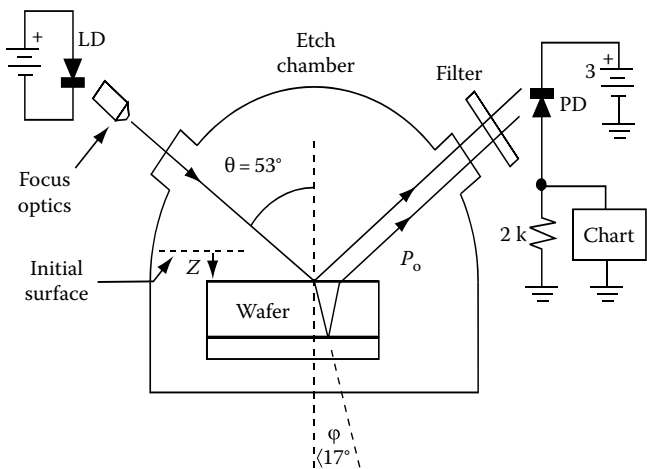


FIGURE A.12 The reflectometer and the ECR.

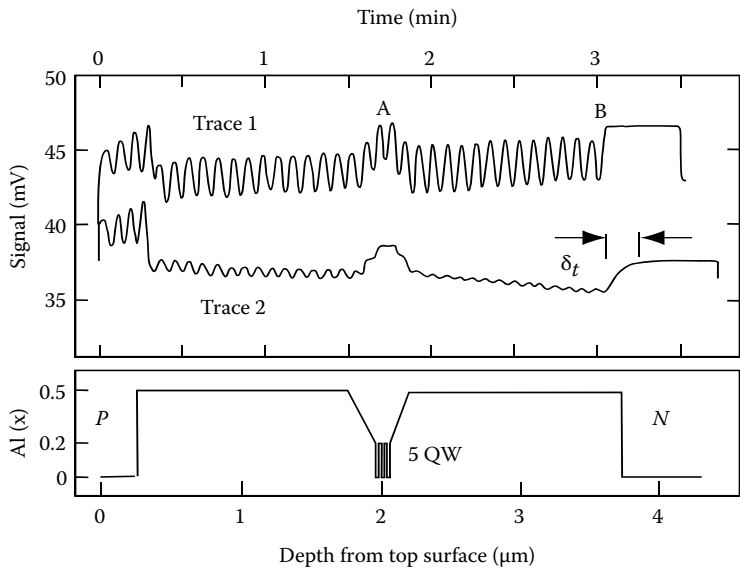


FIGURE A.13 Bottom plot specifies the amount of aluminum “ x .” Top two plots are the signal from the photodetector.

Figure A.12 shows the setup. A laser beam (670 nm) is focused onto the etching $\text{Al}_x\text{Ga}_{1-x}\text{As}$ wafer and reflected into a photodetector. The output can be digitized and recorded. The idea is that as the wafer etches, light reflects off the etching surface and also from interfaces buried within the heterostructure. The reflected beams interfere at the photodetector, which on a chart recorder for example (Figure A.13), shows the interference fringes.

The chart in Figure A.13 shows the reflected signal versus etch time for a five quantum-well laser heterostructure. The bottom plot shows the aluminum content “ x ” as a function of distance from the initial top surface. Notice that the overall shape of the reflected signal (ignoring the small oscillations) matches the amount of aluminum present (changing x , changes the amount of aluminum and therefore the reflectivity). The overall course structure of the top plots is due to the signal reflected from the etching surface. The smaller fringes on both of the upper plots are caused by the interference between the signals reflected from an internal interface and the etching surface. The fringes can be counted to determine a depth to well within $0.1\text{ }\mu\text{m}$. The interference fringes on the lower plot have smaller amplitude due to roughness on the exposed etching surface for that particular sample.

A.2.6 P-TYPE OHMIC CONTACTS

Figure A.8C and D shows that the waveguides with the metal contact on top are the first fabrication level. Besides being the Ohmic p-type contact for the laser diode devices, this metal serves as an etch mask during etches to form the waveguide, as well as an implant mask during an oxygen ion implantation. The processing sequence to apply the p-contacts begins with a shallow zinc diffusion to improve the ohmic contacts. Photolithographic pattern transfer and an image reversal of the photoresist follow this last step. The image reversal is required to produce an undercut in the photoresist for easy removal of the unwanted metal areas (see Figure A.9B). Next, the metals are applied by electron-beam evaporation, and finally a liftoff technique is used to remove the unwanted metal.

Although MOCVD growth of GaAs can dope to $2 \times 10^{19}\text{ cm}^{-3}$, the GaAs surface can be more highly p-doped by diffusing zinc into the lattice. This creates slightly better Ohmic contacts for the

waveguides. This is the first process step, after initially cleaning the wafer, so the zinc diffusion is not masked. The entire wafer surface thus becomes more highly doped. This does not cause neighboring devices to interfere with each other, however, because the wafer surface between the devices is later etched to a depth surpassing the diffusion depth of the zinc. Also, the effect of the zinc diffusion on the back surface of the wafer is unimportant, since the backside is lapped down later in the fabrication process. The diffusion is performed by placing the wafer in a carbon susceptor with a solid zinc-arsenide source in a 650°C oven, with a hydrogen atmosphere, for 9 min, resulting in a diffusion depth of about 300 nm.

To prepare the GaAs surface for metal adhesion, the wafer is dipped in a 10% solution of ammonia hydroxide for 10 s. This chemical removes surface oxides, after which the wafer is immediately loaded into the electron-beam evaporator. The p-contact metals are evaporated onto the wafer in a specific sequence to promote adhesion, and alloyed later in the fabrication process for optimal conductivity.

The p-contact metallization consists of 400 Å titanium, 200 Å platinum, 3000 Å gold, 1500 Å chromium, 500 Å nickel, and another 1500 Å chromium. The titanium is a buffer layer which adheres well to the GaAs surface. Platinum is used as a diffusion barrier for the gold during the alloy step. Gold is the primary metal for conduction and the main mask for the oxygen implant. Because gold is rapidly removed in some etchers (like the CAIBE), however, a chrome or nickel layer is added as the etch mask.

The metallization pattern is completed by performing a liftoff step. The photolithographic image reversal creates undercut sidewalls in the photoresist. This causes gaps between the metal on the GaAs surface and the metal overlaying the photoresist, facilitating the removal of the unwanted metal areas. Soaking the wafer in acetone, which dissolves the photoresist and carries with it any overlaying metal, performs the liftoff. It is often necessary to encourage the chemical process by agitating the surface in an ultrasonic bath. Since the photoresist is hardened by the temperatures in the evaporator, an oxygen plasma descum in the RIE is necessary to completely remove the photoresist from the wafer surface.

A.2.7 OXYGEN ION IMPLANT

Devices on the wafer are electrically isolated in two ways. The first isolation method is to make a shallow etch between the devices to physically delineate the device structures. Better electrical isolation, however, is achieved by implanting oxygen or hydrogen ions into the p-type semiconductor. With the shallow etch, the ions can be implanted down to the quantum-well region without destroying the metallization for the p-type Ohmic contacts. This has the effect of greatly increasing the resistivity of the semiconductor material between devices.

A.2.8 APPLICATION OF N-CONTACT METAL (SEMI-INSULATING SUBSTRATES)

When fabricating devices on a “semi-insulating” substrate (see Figure A.14), it is necessary to make metal contacts to the n-doped epilayer. The term semi-insulating refers to the fact that GaAs is

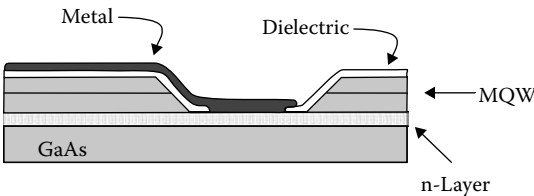


FIGURE A.14 The wet selective etch creates tapered sidewalls for electrical contact between the top-metal and the n-layer.

highly electrically insulating when it is undoped. As indicated in the figure, the device heterostructure resides between the top surface and the buried n-layer. An electrically insulating layer of intrinsic GaAs separates the n-layer from anything grown or attached below it.

To make contact with the n-layer, vias (openings) must first be etched down to the n-doped layer before the metallization is performed. “n-Substrate” wafers, on the other hand, do not require this procedure, since the entire back plane is n-doped and serves as a common ground plane. This section, therefore, applies only to fabrication on semi-insulating substrates.

The pattern for the n-contact vias is transferred to the wafer by photolithography. The same procedure is used as for the deep-etch windows, leaving holes in the photoresist where n-contact vias are desired. After an oxygen plasma descum to ensure clean windows, a wet chemical etch is performed which selectively etches AlGaAs faster than GaAs. The wet etch produces tapered sidewalls in the GaAs, facilitating metallization between the top surface and the n-doped layer, as shown in Figure A.14. The faster etch rate for AlGaAs ensures that the etch will stop when it reaches the GaAs n-doped layer. A final advantage of the wet etch is that it creates an undercut in the GaAs beneath the photoresist which facilitates metallization liftoff.

As a note, it is possible to construct a laser reflectometer to measure the etch rate in the wet etch. In this case, the wafer is submerged in a cylindrical beaker with the etchant. A laser beam enters the beaker at right angles to the glass and fluid. It reflects off the etching surface and leaves the beaker at right angles to the glass. Motion of the fluid surface does not affect the laser beam in this setup.

The metallization procedure is nearly identical to the procedure outlined above for the p-contact metal, the only exception being the types of metals used. Immediately prior to metallization, an ammonia hydroxide dip is performed to remove surface oxides. The metallization sequence for n-type Ohmic contacts is 100 Å nickel, 400 Å germanium, 800 Å gold, 500 Å silver, and a final cap of 800 Å gold. Germanium diffuses into the GaAs during a later alloying step as an n-type dopant for better Ohmic contacts. Nickel keeps the metal from forming “puddles” during the alloy. Gold and silver are the primary low-resistance metals for good conductivity. The unwanted metal is then removed with a liftoff procedure. The same photoresist n-contact mask which was used for the wet etch is used as the soluble liftoff layer.

A.2.9 POLYIMIDE INSULATING LAYERS

A layer of polyimide can be added to the wafer surface for two reasons. First, it acts as an electrical insulator between the underlying structures and the contact pads which are to be placed on the uppermost layer. It also planarizes the surface by smoothing out the multilevel topography of the device structures for better metallization. The polyimide material must be chosen to have the same thermal expansion coefficient as GaAs to prevent strain during device processing and operation.

A.2.10 APPLICATION OF A TOP-METAL CONTACT PAD

With the insulating polyimide layer covering all underlying structures, contact pads can be placed on top of the polyimide without interfering with the devices below. The contact pads, shown in Figure A.8H, only make contact to the underlying p-type or n-type metals where vias are etched through the polyimide. By strategically placing the vias, contact pads can be enlarged to cover the entire wafer area, facilitating electrical probing during test and evaluation. Additionally, if electrical crossovers are required in the device design, one leg can be routed up to the top-metal plane to bypass the other leg. The top-metal contact pads are added to the wafer with photolithography and the liftoff process.

The top-metal contacts are applied with a metallization sequence of 500 Å chromium, 3000 Å silver, and 1000 Å gold. The chrome is used to make the contacts stick to the polyimide surface. The thick silver layer is used primarily for economy, and the gold cap is applied to prevent oxidation and

for later wire bonding. The unwanted metal between contact pads is then removed with the standard liftoff process of washing in acetone followed by the solvent sequence and an oxygen plasma descum.

A.2.11 LAPPING THE WAFER TO FINAL THICKNESS

After completing all of the front-face patterning, the wafer is lapped down to an appropriate thickness. Thinning the wafer facilitates cleaving of the individual devices. On n-substrate wafers, thinning also lowers the series resistance of the devices by shortening the conduction distance through the semiconductor. Finally, lapping also improves heat dissipation, by bringing the devices closer to the heat sink. The final wafer thickness is determined by the feature size. Normally, wafers are lapped to a thickness of 10 mils (250 μm). If the cleaves are to be less than 500 μm apart, however, the wafer thickness is taken down to 4 mils (100 μm).

The wafer is mounted face-down on a chuck with white wax. The backside of the wafer is then manually circulated on a glass counter covered with a 3 μm alumina grit. The thickness of the wafer is gradually reduced, with periodic measurements of progress. Normally this is one of the last steps in the fabrication sequence since the wafer no longer has parallel surfaces which makes lithography difficult at best.

As outlined above, the n-contacts for devices on semi-insulating substrates are metallized from the top. n-Contacts for devices on n-substrate wafers, however, cover the entire backside of the wafer. Although this facilitates metallization of the n-contact, a disadvantage is that the n-contact cannot be patterned. The n-substrate is thus used only when a continuous ground plane can be used for all devices.

The n-contact metallization for n-substrate devices is accomplished following the lapping process. The metallization is essentially the same as for the semi-insulating substrate devices, except that there is no patterning or liftoff; the metallization is applied to the entire backside permanently. The sequence of metals used for the n-contact is the same as for the semi-insulating n-contacts: 100 Å nickel, 400 Å germanium, 800 Å gold, 500 Å silver, and a final cap of 800 Å gold.

A.2.12 ALLOY THE METALS

At the completion of the final metallization procedure, an alloy step is performed to eliminate the Schottky barrier in the n-contact, making it Ohmic. The alloy is accomplished by placing the wafer in a 360°C oven for 60 s. This causes the germanium to diffuse into the GaAs at the n-contacts.

REFERENCES AND FURTHER READINGS

Fabrication Books and Related

1. Williams R., *Modern GaAs Processing Methods*, Artech House, Boston, MA (1990). One of the best books for fabrication.
2. Jager R.C., *Introduction to Microelectronic Fabrication*, 2nd ed., *Modular Series on Solid State Devices*, Vol. 5, Prentice Hall, Upper Saddle River, NJ (2002).
3. Campbell S.A., *The Science and Engineering of Microelectronic Fabrication*, 2nd ed., Oxford University Press, New York (2001).
4. Campbell S.A., *Fabrication Engineering at the Micro- and Nanoscale*, 3rd ed., Oxford University Press, New York (2008).
5. Bhattacharya P., *Semiconductor Optoelectronic Devices*, 2nd ed., Prentice Hall, Upper Saddle River, NJ (1997). Good general reference on most aspects of solid state including fabrication, electronic processes, bands, junctions, contacts and optoelectronic devices.
6. Sze S.M., *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, New York (1981).

Wet Etching and Monitors—Journal Publications

7. Swanson P.D., Parker M.A., Kimmet J.S., Shire D.B., Tang C.L., and Michalak R.J., Electron cyclotron resonance etching of laser mirrors for ridge guided lasers, *IEEE Photonics Technol. Lett.* 7, 605 (1995).
8. Parker M.A., Kimmet J.S., Michalak R.J., Wang H.S., Shire D.B., Tang C.L., and Drumheller J., Accurate electron-cyclotron-resonance etching of semiconductor laser heterostructures using a simple laser reflectometer, *Photonics Technol. Lett.* 8(#6), 818–820 (1996).
9. Parker M.A., Michalak R.J., Kimmet J.S., Pirich A.R., and Shire D.B., Etched-surface roughness measurements from an in-situ laser reflectometer, *Appl. Phys. Lett.* 69(#10), 1459–1461 (September 1996).
11. Thakurdesai M., Parker M.A., A handheld smart wet etch monitor: Theory, design and test, *IEEE Trans. Instrum. Measure.*, 55(5), 1814–1822 (October 2006). Visual Basic and Bascom software posted on web site ece.rutgers.edu/~maparker.
12. Yi E.H., Akdogan I.G., and Parker M.A., Measurements of wet etch dynamics using an in-situ optical monitor, *IEEE Electrochem. Solid-State Lett.* 6(5) G75–77 (2003).
13. Yi H.T., Thakurdesai M., and Parker M.A., Control of sidewall angles using UV LEDs during wet etching of GaAs, *IEEE Electrochem. Solid State Lett.*, 7, C137–C139 (2004).
14. Yi E.H. and Parker M.A., Photo-dynamics of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ heterostructure dissolution: Experiments and applications, *ECS Trans.*, 6, 525–534 (2007).
15. Yi E.H. and Parker M.A., Photo-assisted wet-etched III-V heterostructure laser mirrors and waveguides, *Photonics Tech. Lett.* (2008).

Appendix B: Dirac Delta Function

The Dirac delta function (also called the impulse function) arises in many fields of engineering and physics. In many respects, the Dirac delta function can be thought of as a function. The Dirac delta function departs from classical mathematical theory and must be defined as the limit of a sequence of functions. Distribution function theory provides a firm basis for the Dirac delta function. This section provides a number of representations of the Dirac delta function. We will find that every basis set of functions provides another representation. This section also discusses the idea of principal part.

B.1 INTRODUCTION TO THE DIRAC DELTA FUNCTION

We often think of the Dirac delta function $\delta(x - x_0)$ as a function with exactly one infinite value at the point x_0 and zero everywhere else (Figure B.1).

$$\delta(x - x_0) = \begin{cases} \infty & x = x_0 \\ 0 & x \neq x_0 \end{cases} \quad (\text{B.1})$$

The function must be infinitely large at x_0 but infinitely narrow so that the area under the function equals to 1. Apparently, integrals over the delta function have wonderful properties.

We might also consider an alternate definition of the Dirac delta function by the effect it has on integrals. Define the delta function by

$$\int_a^b dx f(x) \delta(x - x_0) = \begin{cases} f(x_0) & x_0 \in (a, b) \\ \frac{1}{2} f(x_0) & x_0 = a \text{ or } b \\ 0 & \text{else} \end{cases} \quad (\text{B.2})$$

Notice that if $f(x) = 1$ then Equation B.2 provides

$$\int_a^b dx \delta(x - x_0) = \begin{cases} 1 & x_0 \in (a, b) \\ \frac{1}{2} & x_0 = a \text{ or } b \\ 0 & \text{else} \end{cases} \quad (\text{B.3})$$

The integral of the delta function has the value of one when the point of discontinuity x_0 appears entirely inside the integration interval. When you encounter a delta function in an equation, you should consider it an “invitation to integrate.” The next section shows how the Dirac delta function really comes from the limit of a sequence of functions, which substantiates Equations B.2 and B.3 (Figure B.1).

Example B.1

What is $\int_{10}^{50} \frac{\sin x}{\sqrt{1 + 3x^2}} \delta(x - 45^\circ) dx$?

$$\int_{10}^{50} \frac{\sin x}{\sqrt{1 + 3x^2}} \delta(x - 45^\circ) dx = \left. \frac{\sin x}{\sqrt{1 + 3x^2}} \right|_{x=45^\circ} = \frac{\sin 45}{\sqrt{1 + 3(45)^2}}$$

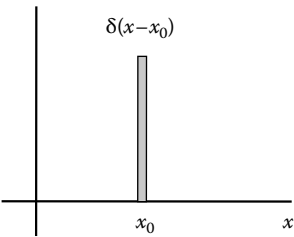


FIGURE B.1 Representation of the delta function as a narrow spike.

B.2 THE DIRAC DELTA FUNCTION AS LIMIT OF A SEQUENCE OF FUNCTIONS

The Dirac delta function should really be defined as the limit of a sequence of functions S_n according to the definition

$$\int_{-\infty}^{\infty} dz \, \delta(z - z_0) = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} dz \, S_n(z - z_0) \tag{B.4}$$

The order of the limit, integral, and function S_n should be carefully noted. Many different sequences of functions S_n will work even those that cannot be differentiated everywhere.

Example B.4

Figure B.2 shows a sequence of functions $S_n(z - z_0)$ given by

$$\begin{aligned} S_1 &= 1/6 & x \in (0, 6) \\ S_2 &= 1/2 & x \in (2, 4) \\ S_3 &= 1 & x \in (2.5, 3.5) \\ &\vdots & \vdots \end{aligned}$$

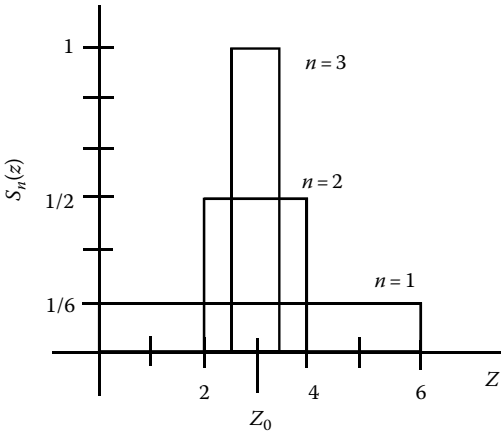


FIGURE B.2 A sequence of functions with a “limit” that represents the Dirac delta function.

Notice that the area under each function S_n equals to 1. We can then trivially write

$$\lim_{n \rightarrow \infty} \int_0^9 dz S_n(z - z_0) = \lim_{n \rightarrow \infty} 1 = 1 = \int_0^9 dz \delta(z - z_0)$$

This last example brings up an important point regarding the definition for the integral of the delta function in terms of the limit of a sequence of functions.

$$\lim_{n \rightarrow \infty} \int_a^b dx f(x) S_n(x - x_0) \equiv \int_a^b dx f(x) \delta(x - x_0)$$

The integral of each function S_n does not need to equal unity; however, at the very least, the integral of S_n should approach “1” as “ n ” becomes large. In many cases, we require each function in the sequence S_n to be everywhere differentiable. For example, S_n might be Gaussian-shaped functions.

Many books use a shorthand notation for the Dirac delta function. For example, looking at the defining relation

$$\int_{-\infty}^{\infty} dz \delta(z - z_0) = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} dz S_n(z - z_0) \quad (\text{B.5})$$

we might be tempted to make the identification

$$\delta(z - z_0) = \lim_{n \rightarrow \infty} S_n(z - z_0) \quad (\text{B.6})$$

However, this can only be correct when interpreted as in Equation B.5. We can easily see the problem with directly integrating Equation B.6. Setting the Dirac delta function δ directly equal to the limit of a sequence of functions produces a limit function equal to zero everywhere except at one point. This limit function matches the intuitive view of the Dirac delta function. Taking the integral of this limit function must produce zero because a Riemann integral is insensitive to a single point. The integral of the limit function does not produce a value equal to unity contrary to the definition of the Dirac delta function.

Now let us discuss why the first integral property in Equation B.2 holds, namely

$$\int_a^b dx f(x) \delta(x - x_0) = \begin{cases} f(x_0) & x_0 \in (a, b) \\ \frac{1}{2} f(x_0) & x_0 = a \text{ or } b \\ 0 & \text{else} \end{cases} \quad (\text{B.7})$$

Figure B.3 shows a sequence of functions S_n all enclosing unit area. The first graph shows that $f(z)$ varies along the nonzero portion of S_1 . The middle picture shows a case with $f(z)$ almost constant over the width of S_2 . Finally, the last graph shows a function $S_3(z - z_0) \cong \delta(z - z_0)$ sufficiently narrow to provide a very good approximation $f(z) \cong f(z_0)$ over the nonzero width of $S_3(z - z_0)$. As a result of this intuitive approach, we can write

$$\int_{-\infty}^{\infty} dz f(z) \delta(z - z_0) \cong \int_{-\infty}^{\infty} dz f(z) S_3(z - z_0) \cong \int_{-\infty}^{\infty} dz f(z_0) S_3(z - z_0) = f(z_0) \int_{-\infty}^{\infty} dz S_3(z - z_0) = f(z_0)$$

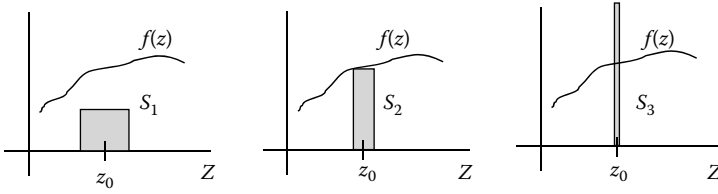


FIGURE B.3 Making n sufficiently large makes S_n sufficiently narrow so that $f(z)$ does not vary along the nonzero portion of S_n . In this case, we can take $\delta(z - z_0) \cong S_3(z - z_0)$.

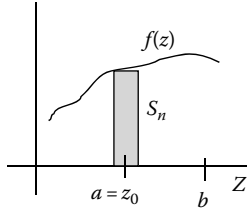


FIGURE B.4 The integral covers only “half” of the delta function.

which demonstrates the first of the integrals. This last approximation also works for functions that are not delta functions so long as they are very sharply peaked; however, the result must be multiplied by a constant equal to the integral over the function.

Now what about the property in Equation B.2 for $z_0 = a$, namely

$$\int_a^b dz f(z) \delta(z - z_0) = \frac{1}{2} f(z_0)$$

This property holds because the integral covers only half of the delta function. Using Figure B.4 and a fairly narrow S_n (as shown), we can again write

$$f(z) S_n(z) \cong f(z_0) S_n(z)$$

and the integral becomes

$$\begin{aligned} \int_a^b dz f(z) S_n(z - z_0) &\cong \int_a^b dz f(z_0) S_n(z - z_0) \quad \text{or} \\ \int_a^b dz f(z) S_n(z - z_0) &= f(z_0) \int_a^b dz S_n(z - z_0) \end{aligned}$$

Now, because $a = z_0$, the integral covers only half of the width of S_n , and the integral becomes

$$\int_{a=z_0}^b dz S_n(z - z_0) = 1/2$$

Finally, including $f(z)$

$$\int_{z_0}^{b \gg z_0} dz f(z) \delta(z - z_0) = \frac{1}{2} f(z_0)$$

B.3 THE DIRAC DELTA FUNCTION FROM THE FOURIER TRANSFORM

The Dirac delta function is most often first encountered with Fourier transforms. The following derivation shows how this comes about. Start with the Fourier integral

$$f(x) = \int_{-\infty}^{\infty} dk \frac{e^{ikx}}{\sqrt{2\pi}} f(k)$$

and then substitute the Fourier transform for $f(k)$

$$f(k) = \int_{-\infty}^{\infty} dX \frac{e^{-ikX}}{\sqrt{2\pi}} f(X)$$

to find

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} dk \frac{e^{ikx}}{\sqrt{2\pi}} f(k) = \int_{-\infty}^{\infty} dk \frac{e^{ikx}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dX \frac{e^{-ikX}}{\sqrt{2\pi}} f(X) \\ &= \int_{-\infty}^{\infty} dX \int_{-\infty}^{\infty} dk \frac{e^{-ik(X-x)}}{2\pi} f(X) = \int_{-\infty}^{\infty} dX f(X) \int_{-\infty}^{\infty} dk \frac{e^{-ik(X-x)}}{2\pi} \end{aligned}$$

Comparing both sides of the equation we see that the second integral must be related to a Dirac delta function in order that $f(X)$ becomes $f(x)$. Therefore,

$$\delta(x - X) = \int_{-\infty}^{\infty} dk \frac{e^{-ik(x-X)}}{2\pi}$$

and similarly

$$\delta(k - K) = \int_{-\infty}^{\infty} dx \frac{e^{-i(k-K)x}}{2\pi}$$

which can be proved in the same manner as for x -Delta function but starting with $f(k)$ instead of $f(x)$.

B.4 OTHER REPRESENTATIONS OF THE DIRAC DELTA FUNCTION

This section lists some common sequences for the Dirac delta function.

- 1. The previous section discusses the sequence of rectangles defined by

$$S_{\alpha} = \begin{cases} 1/\alpha & |x| \leq \alpha/2 \\ 0 & |x| \geq \alpha/2 \end{cases}$$

Note that $S_{\alpha}(x - x_0)$ is obtained by replacing x with $x - x_0$ in the formula (Figures B.5).

- 2. The Gaussian probability density function (Figure B.7)

$$g_{\sigma}(x - x_0) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(x - x_0)^2}{2\sigma^2} \right]$$

represents a delta function when the standard deviation σ approaches 0. These distribution functions can be written in terms of the integer “ n ” by setting $\sigma = 1/n$ for example. The delta function can be written as

$$\lim_{\sigma \rightarrow 0} g_{\sigma}(x - x_0) = \delta(x - x_0)$$

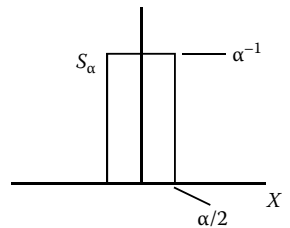


FIGURE B.5 Sequence of rectangles.

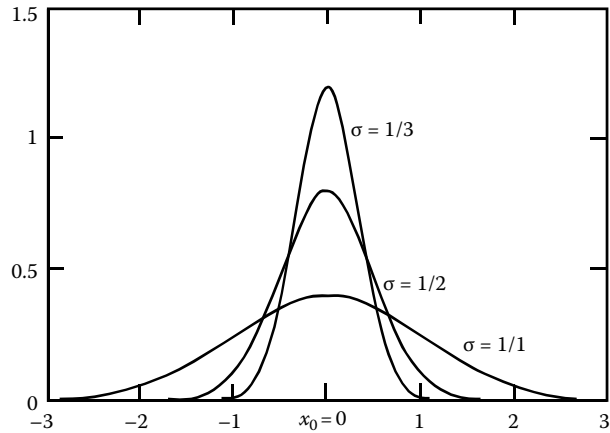


FIGURE B.6 The limit of the Gaussian probability distribution approaches the Dirac delta function.

with the understanding that this means

$$\lim_{\sigma \rightarrow 0} \int_a^b dx f(x) g_{\sigma}(x - x_0) \equiv \int_a^b dx f(x) \delta(x - x_0)$$

Without the integral, the limit of the sequence of distribution functions g_{σ} would be zero at all points except at x_0 where the limit of the distribution is infinite. The point x_0 is at the center of the distribution and σ is the standard deviation.

$$3. \delta(x) = \lim_{\varepsilon \rightarrow 0} S_{\varepsilon}(x) = \lim_{\varepsilon \rightarrow 0} \frac{1}{\pi} \frac{\varepsilon}{x^2 + \varepsilon^2}$$

4. The theory of Fourier transforms provides an integral representation (see Section B.3 above)

$$\delta(x) = \lim_{\kappa \rightarrow \infty} \int_{-\kappa}^{\kappa} dk \frac{e^{ikx}}{2\pi} = \int_{-\infty}^{\infty} dk \frac{e^{ikx}}{2\pi} \quad (\text{B.8})$$

which can be written in two other forms

$$\delta(x) = \lim_{\kappa \rightarrow \infty} \int_0^{\kappa} dk \frac{\cos(kx)}{\pi} \quad (\text{B.9})$$

and

$$\delta(x) = \lim_{\kappa \rightarrow \infty} \frac{\sin(\kappa x)}{\pi x} \quad (\text{B.10})$$

Equation B.9 is related to the “sinc” function. Figure B.7 shows how increasing the value of “ κ ” causes the function “ $\sin(\kappa x)/\pi x$ ” to become sharper and more narrow; the height of

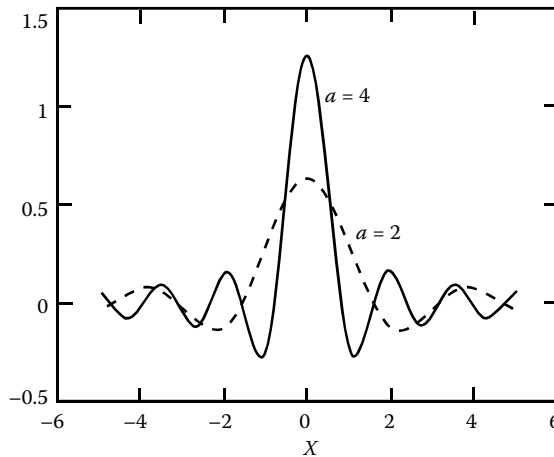


FIGURE B.7 A plot of Equation B.10 for two values of κ .

the function is “ κ/π ” and the distance from $x = 0$ to the first zero is “ π/κ ”. Equation B.9 follows from Equation B.8

$$\delta(x) = \int_{-\infty}^{\infty} dk \frac{e^{ikx}}{2\pi} = \int_{-\infty}^0 dk \frac{e^{ikx}}{2\pi} + \int_0^{\infty} dk \frac{e^{ikx}}{2\pi} = \int_0^{\infty} dk \frac{e^{-ikx}}{2\pi} + \int_0^{\infty} dk \frac{e^{ikx}}{2\pi} = \int_0^{\infty} dk \frac{\cos(kx)}{\pi}$$

where the integral is divided into two (one over negative k and the other over positive k), replacing k with $-k$ in one of them (the one for negative k) and then recombining the two integrals using one of Euler's equations $\cos(kx) = [e^{ikx} + e^{-ikx}]/2$. Equation B.10 follows from Equation B.8 as follows

$$\delta(x) = \lim_{\kappa \rightarrow \infty} \int_{-\kappa}^{\kappa} dk \frac{e^{ikx}}{2\pi} = \lim_{\kappa \rightarrow \infty} \left[\frac{e^{i\kappa x} - e^{-i\kappa x}}{2\pi i x} \right] = \lim_{\kappa \rightarrow \infty} \frac{\sin(\kappa x)}{\pi x}$$

Note that the $\sin(\kappa x)/x$ appears as a sequence in “ κ ” just like the previous examples while Equations B.8 and B.9 have the parameter as the bounds on an integral.

B.5 THEOREMS ON THE DIRAC DELTA FUNCTIONS

There are some useful theorems on the Dirac delta function that allow a person to simplify expressions. G. Barton's book “Elements of Green's Functions and Propagation” published by Oxford Science Publications in 1989 provides a good reference.

1. $\delta(x - \xi) = \delta(\xi - x)$
2. $\delta(ax) = \frac{1}{|a|} \delta(x)$
3. If $g(x)$ has real roots x_n (that is $g(x_n) = 0$) then

$$\delta[g(x)] = \sum_n \frac{\delta(x - x_n)}{|g'(x_n)|} \quad \text{where} \quad g'(x) = dg/dx$$

4. For $\xi \in (a, b)$,

$$\int_a^b dx f(x) \delta'(x - \xi) = -f'(\xi)$$

This property is important because it allows for a weak identity that is exceedingly useful

$$f(x) \delta'(x - \xi) = -f'(\xi) \delta(x - \xi)$$

B.6 THE PRINCIPAL PART

If half the range of “ k ” is left off the integral in Equation B.8 then a function $\zeta(x)$ can be defined by

$$\zeta(x) = -i \lim_{\kappa \rightarrow \infty} \int_0^{\kappa} dk \frac{e^{ikx}}{2\pi}$$

where an extra $i = \sqrt{-1}$ is added for later convenience. Integrating provides

$$\zeta(x) = \frac{1}{2\pi} \lim_{\kappa \rightarrow \infty} \left(\frac{1 - e^{i\kappa x}}{x} \right) = \frac{1}{2\pi} \lim_{\kappa \rightarrow \infty} \left(\frac{1 - \cos(\kappa x)}{x} - i \frac{\sin(\kappa x)}{x} \right)$$

where the last step obtains using $e^{i\kappa x} = \cos(\kappa x) + i \sin(\kappa x)$. Half the range of the integral in Equation B.8 is removed to obtain the expression for $\zeta(x)$. The reader should realize that for Equation B.9, half the range of the integral was not removed from Equation B.8; the range was folded up (so to speak) into the cosine term. Now for $\zeta(x)$, define the principal part $\wp\left(\frac{1}{x}\right) = \frac{\wp}{x}$

$$\frac{\wp}{x} = \lim_{\kappa \rightarrow \infty} \frac{1 - \cos(\kappa x)}{x}$$

as the principal part of $1/x$. The imaginary part of $\zeta(x)$ is related to the Dirac delta function as shown in #4 above. Now it is possible to write an alternate expression for $\zeta(x)$ as

$$\zeta(x) = \lim_{\kappa \rightarrow \infty} \frac{1 - \cos(\kappa x)}{2\pi x} - i \lim_{\kappa \rightarrow \infty} \frac{\sin(\kappa x)}{2\pi x} = \frac{\wp}{2\pi x} - i \frac{\delta(x)}{2}$$

Restricting the range of “ k ” for the integral is therefore seen to give something that differs from the delta function by the value of the principal part.

What is $P(1/x) = \lim_{\kappa \rightarrow \infty} \frac{1 - \cos(\kappa x)}{2\pi x}$? As a function of x , taking the limit literally, only $x = 0$ is defined since $\cos(\kappa x)$ does not have a limit (with κ as the limit variable) where $x \neq 0$. At $x = 0$, the limit becomes (by Taylor expanding the cosine function)

$$P(1/x) = \lim_{\kappa \rightarrow \infty} \frac{1 - \cos(\kappa x)}{2\pi x} \cong \lim_{\kappa \rightarrow \infty} \lim_{x \rightarrow 0} \frac{1 - \left[1 - \frac{(\kappa x)^2}{2!} + \dots \right]}{2\pi x} = 0$$

by LaHospital’s rule. Now, because the principal part occurs in the same equation as the Dirac delta function, the reader should anticipate that the principal part has special integral properties. The integral of the terms in $\zeta(x)$ is found before taking the limit (the limit is understood to be outside the integral).

The integral of $P(1/x)$ requires some explanation. Consider two cases for the integration interval of $[a, b]$. First assume that $a > 0$ and $b > 0$ and second, assume that $a < 0$ and $b > 0$. Consider case 1 for $a > 0$ and $b > 0$. Figure B.8 shows a plot of $[1 - \cos(\kappa x)]/x$ (solid curve) for a “fixed” κ and also a plot of $1/x$ (dotted curve). Notice how $1/x$ appears as a “local” average for the curve. To evaluate the integral, divide the interval $[a, b]$ into smaller intervals $[a_i, b_i]$ such that

1. $[a, b] = \cup_{i=1}^n [a_i, b_i]$, where $a_i = b_{i-1}$ and $\cup_{i=1}^n [a_i, b_i]$ means the union of the subintervals.
2. The function $1/x$ does not vary appreciably over $[a_i, b_i]$

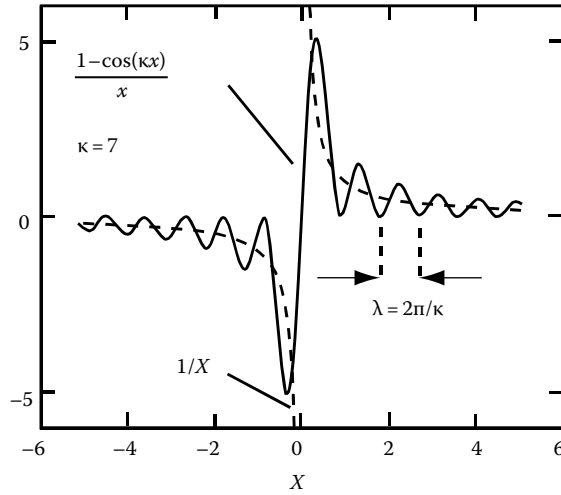


FIGURE B.8 The function “ $1/x$ ” is an average of $[1 - \cos(\kappa x)]/x$.

3. $[1 - \cos(\kappa x)]$ passes through many cycles over each $[a_i, b_i]$; this is certainly the case for large κ when $b_i - a_i \gg \lambda$. (see λ in Figure B.8).

Using the first property, the integral can be rewritten as

$$\int_a^b = \sum_{i=1}^n \int_{a_i}^{b_i}$$

We also need the mean value theorem from calculus which can be written as

$$\int_{a_i}^{b_i} dx f(x) = \langle f(x) \rangle (b_i - a_i)$$

Now, applying the mean value theorem to $[1 - \cos(\kappa x)]/x$ keeping in mind that $1/x$ is a local average, we find

$$\int_{a_i}^{b_i} dx \frac{1}{x} = \lim_{\kappa \rightarrow \infty} \int_{a_i}^{b_i} dx \frac{1 - \cos(\kappa x)}{x} = \lim_{\kappa \rightarrow \infty} \left\langle \frac{1 - \cos(\kappa x)}{x} \right\rangle (b_i - a_i) = \lim_{\kappa \rightarrow \infty} \frac{1}{x} (b_i - a_i) = \int_{a_i}^{b_i} dx \frac{1}{x}$$

The third and last terms were found by applying the mean value theorem. The limit in the fourth term does not matter and can be dropped. How is $\left\langle \frac{1 - \cos(\kappa x)}{x} \right\rangle$ found? This can be seen in two ways. For the first way, $1/x$ was already noted to be the average of $[1 - \cos(\kappa x)]/x$ for small enough intervals. For the second way, we can write

$$\int_{a_i}^{b_i} dx \frac{1 - \cos(\kappa x)}{x} \cong \frac{1}{x} \int_{a_i}^{b_i} dx [1 - \cos(\kappa x)] = \frac{b_i - a_i}{x} - \frac{\sin(\kappa x)}{\kappa} \Big|_{a_i}^{b_i} \frac{1}{x} \cong \frac{b_i - a_i}{x}$$

Thus for case 1, we can make the replacement

$$\int_{a_i}^{b_i} dx f(x) \frac{\wp}{x} \Rightarrow \int_{a_i}^{b_i} dx \frac{f(x)}{x}$$

so long as $f(x)$ is slowly varying. The original integral can be written as

$$\int_a^b dx f(x) \frac{\wp}{x} = \sum_{i=1}^n \int_{a_i}^{b_i} dx \wp \frac{f(x)}{x} = \sum_{i=1}^n \int_{a_i}^{b_i} dx \frac{f(x)}{x} = \int_a^b dx \frac{f(x)}{x}$$

for $a, b > 0$. For this case, the principal part has no effect. Also notice that the sine term (i.e., the delta function) in

$$\zeta(x) = \lim_{\kappa \rightarrow \infty} \frac{1 - \cos(\kappa x)}{2\pi x} - i \lim_{\kappa \rightarrow \infty} \frac{\sin(\kappa x)}{2\pi x} = \frac{\wp}{2\pi x} - i \frac{\delta(x)}{2}$$

is approximately zero since the point of discontinuity is outside the interval (i.e., $a > 0, b > 0$).

Consider the second case of $a < 0$ and $b > 0$. Again divide up the interval into small subintervals satisfying the properties on the previous page. Those subintervals that do not contain zero are handled just like case 1. Therefore consider the subinterval $[-\varepsilon, \varepsilon]$ where ε is a small number. As discussed above $P(1/x) \cong 0$ for “ x ” near 0. The integral over the ε subinterval becomes

$$\int_{-\varepsilon}^{\varepsilon} dx f(x) \wp \left(\frac{1}{x} \right) = f(x) \wp \left(\frac{1}{x} \right) \Big|_{-\varepsilon}^{\varepsilon} \cong 0$$

The smaller the value of ε , the better the approximation. The original integral becomes

$$\begin{aligned} \int_a^b dx f(x) \wp \left(\frac{1}{x} \right) &= \int_a^{-\varepsilon} dx f(x) \wp \left(\frac{1}{x} \right) + \int_{-\varepsilon}^{\varepsilon} dx f(x) \wp \left(\frac{1}{x} \right) + \int_{\varepsilon}^b dx f(x) \wp \left(\frac{1}{x} \right) \\ &= \int_a^{-\varepsilon} dx \frac{f(x)}{x} + 0 + \int_{\varepsilon}^b dx \frac{f(x)}{x} \end{aligned}$$

Some people define the principal part of the integral as

$$\wp \int_a^b = \int_a^{-\varepsilon} + \int_{\varepsilon}^b$$

B.7 CONVERGENCE FACTORS AND THE DIRAC DELTA FUNCTION

In many cases, the form of the Dirac delta function (for a given Hilbert space) is surmised from the closure relation. This section discusses one method of showing that the area under a Dirac

delta function is equal to 1. Consider the Fourier representation of the Dirac delta function $\delta(k - 0)$ given by

$$I(k) = \int_{-\infty}^{\infty} dx \frac{e^{ikx}}{2\pi} \quad (\text{B.11})$$

The integral can be evaluated by including a “convergence” factor $e^{\pm\alpha x}$ with $\alpha > 0$. The “positive” sign in $e^{\pm\alpha x}$ is used when “ x ” is negative and the “negative” sign in $e^{\pm\alpha x}$ is used when “ x ” is positive. Including the appropriate integrating factor forces the integrand in Equation B.11 to approach zero near $\pm\infty$. After the calculation is complete, the parameter α is set to 0.

$$I(k) = \int_{-\infty}^{\infty} dx \frac{e^{ikx}}{2\pi} = \int_0^{\infty} dx \frac{e^{ikx}}{2\pi} + \int_{-\infty}^0 dx \frac{e^{ikx}}{2\pi} = \int_0^{\infty} dx \frac{e^{-\alpha x + ikx}}{2\pi} + \int_{-\infty}^0 dx \frac{e^{\alpha x + ikx}}{2\pi}$$

Notice that integrating factors are included in the integrals. Carrying out the integrals provides

$$I(k) = \int_{-\infty}^{\infty} dx \frac{e^{ikx}}{2\pi} = -\frac{1}{2\pi(-\alpha + ik)} + \frac{1}{2\pi(\alpha + ik)} = \frac{1}{2\pi} \frac{2\alpha}{(k - i\alpha)(k + i\alpha)} \quad (\text{B.12})$$

Notice that if $k = 0$ then as $\alpha \rightarrow 0$ the integral becomes infinite $I(k) \rightarrow \infty$. On the other hand, if $k \neq 0$ then as $\alpha \rightarrow 0$ the integral becomes zero $I(k) \rightarrow 0$. This behavior matches that for a Dirac delta function $\delta(k - 0)$.

To evaluate the integral of $I(k)$, namely $\int_{-\infty}^{\infty} dk I(k)$, a contour integration can be performed in Equation B.12. The contour can be closed in either the lower half plane or the upper half plane. A closed contour in the upper half plane encloses a pole at $k = i\alpha$. The basic formula for residues can be used

$$\oint dz \frac{f(z)}{z - z_0} = 2\pi i \sum \text{residues} = 2\pi i f(z_0)$$

to find

$$\oint dk I(k) = \oint dk \frac{1}{2\pi} \frac{2\alpha}{(k - i\alpha)(k + i\alpha)} = 2\pi i \left\{ \frac{1}{2\pi} \frac{2\alpha}{(k + i\alpha)} \right\}_{k=i\alpha} = 1$$

Appendix C: Fourier Transform from the Fourier Series

The Fourier series can be used to represent periodic functions that are piecewise continuous. As you probably realize, the analysis of linear and optical systems requires Fourier transforms. The Fourier transform provides a representation of “nonperiodic” functions in terms of e^{ikx} . This section shows how the Fourier series leads to the Fourier transform by starting with a function with period $2L$ and then allowing $L \rightarrow \infty$. If a function $f(x)$ has period $2L$ then its Fourier series expansion can be written as

$$f(x) = \sum_{n=-\infty}^{\infty} F(n) \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right) \quad (\text{C.1})$$

where $F(n)$ is usually considered the transformed function. A known function $f(x)$ produces the components $F(n)$ (i.e., the components of the vector “ f ” when it is projected into the Fourier series basis set).

$$F(n) = \left\langle \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right) \middle| f \right\rangle = \frac{1}{\sqrt{2L}} \int_{-L}^L dx \exp\left(\frac{-in\pi x}{L}\right) f(x) \quad (\text{C.2})$$

where recall that complex conjugates are taken of functions in the left slot of the inner product bracket.

The Fourier transform pair can be found from Equation C.1 by making the following replacements

$$\frac{n\pi}{L}x = k_n y \quad \text{where } k_n = n\sqrt{\frac{\pi}{L}} \quad \text{and} \quad y = \sqrt{\frac{\pi}{L}}x$$

and setting

$$\Delta k = k_{n+1} - k_n = (n+1)\sqrt{\frac{\pi}{L}} - n\sqrt{\frac{\pi}{L}} = \sqrt{\frac{\pi}{L}}$$

or, better yet writing

$$\frac{\Delta k}{\sqrt{2\pi}} = \frac{1}{\sqrt{2L}}$$

Equation C.1 becomes

$$f(x) = \sum_{n=-\infty}^{\infty} F(k_n) \frac{1}{\sqrt{2\pi}} \exp(ik_n y) \Delta k \quad (\text{C.3})$$

Next, let the period of $f(x)$ become large as $L \rightarrow \infty$. This requires

$$\Delta k = \sqrt{\frac{\pi}{L}} \rightarrow 0 \quad \text{as } L \rightarrow \infty$$

and Equation C.3 becomes an integral

$$f(x) = \int_{-\infty}^{\infty} dk F(k) \frac{e^{iky}}{\sqrt{2\pi}} \quad (\text{C.4})$$

The inverse transform comes from Equation C.2 which is

$$F(n) = \left\langle \frac{1}{\sqrt{2L}} \exp\left(i \frac{n\pi x}{L}\right) \middle| f \right\rangle = \frac{1}{\sqrt{2L}} \int_{-L}^L dx \exp\left(-i \frac{n\pi x}{L}\right) f(x)$$

Making the same substitutions with $y = \sqrt{\frac{\pi}{L}}x$

$$F(k) = \frac{1}{\sqrt{2L}} \int_{-\sqrt{\pi L}}^{\sqrt{\pi L}} \left(\sqrt{\frac{L}{\pi}} dy \right) e^{-iky} f(y)$$

or, letting $L \rightarrow \infty$, we have

$$F(k) = \int_{-\infty}^{\infty} dy \frac{e^{-iky}}{\sqrt{2\pi}} f(y)$$

We discuss the basis set in the chapter on linear algebra. People write $f(y)$ as the function and $f(k)$ as the Fourier transform. Note that the same symbol “ f ” is used for both $f(y)$ and $f(k)$, where “ y ” is the real spatial coordinate and “ k ” is the real Fourier transform coordinate. As discussed previously, $f(y)$ and $f(k)$ are different representations of the same thing namely a function f .

Example C.1

Find the Fourier transform of

$$f(x) = \begin{cases} 1 & x \in [-L, L] \\ 0 & \text{elsewhere} \end{cases}$$

which represents an optical aperture (Figure C.1).

The Fourier transform is

$$f(k) = \int_{-\infty}^{\infty} dx f(x) \frac{e^{-ikx}}{\sqrt{2\pi}} = \frac{1}{ik\sqrt{2\pi}} [e^{ikL} - e^{-ikL}] = \sqrt{\frac{2}{\pi}} \frac{\sin kL}{k} \quad (\text{C.5})$$

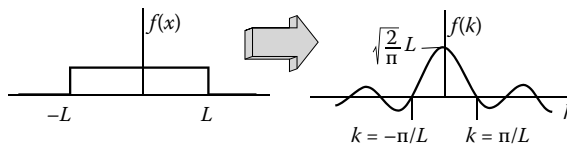


FIGURE C.1 The Fourier transform of a square aperture.

Notice that as the width of the aperture increases $L \rightarrow \infty$, the width of $f(k)$ decreases but its height increases. In fact, other chapters show that one representation of the Dirac delta function is

$$\delta(x) = \lim_{\alpha \rightarrow \infty} \frac{\sin(\alpha x)}{\pi x}$$

Then Equation C.5 gives

$$\lim_{L \rightarrow \infty} f(k) = \sqrt{2\pi} \lim_{L \rightarrow \infty} \frac{\sin(kL)}{\pi k} = \sqrt{2\pi} \delta(k)$$

So very wide optical apertures give spatial Fourier transforms $f(k)$ that approximate a Dirac delta function (i.e., very narrow).

Appendix D: Brief Review of Probability

The present appendix reviews selected topics from probability and statistics. Most of the examples focus on optics and noise processes. We first introduce the probability density, cumulative probability, and the average.

D.1 PROBABILITY DENSITY

The probability density function ρ measures the probability per unit “something” such as per unit length or per unit volume. If $\rho(x)dx$ is the probability of finding a particle in the infinitesimal interval dx centered at the position x , then the probability of finding the particle in the interval $[a, b]$ is given by

$$P(a \leq x \leq b) = \int_a^b dx \rho(x) \quad (D.1)$$

The integral presumes the random variable x is continuous. Discrete random variables reduce the integral in Equation D.1 to a summation. As is typical for classical probability theory, the integral of the density function ρ must be 1

$$\int_{-\infty}^{\infty} dx \rho(x) = 1 \quad (D.2)$$

The fact that the integral over all space equals unity is a reflection of the fact that a particle, for example, must be found somewhere in space (i.e., the total probability equals one for finding the particle somewhere). For volume rather than length, the probability density is $\rho(x, y, z)$. The probability of finding a particle in a volume V of space is then

$$P(a \leq x \leq b, c \leq y \leq d, e \leq z \leq f) = \int_a^b \int_c^d \int_e^f dx dy dz \rho(x, y, z) = \int_V dV \rho \quad (D.3)$$

The average of a real-valued random variable “ x ” can be symbolized several ways

$$\bar{x} = \langle x \rangle = E[x] \quad (D.4)$$

where $E(\)$ is the expectation operator from probability theory. These averages are calculated as usual

$$\langle f(x) \rangle = \int_{-\infty}^{\infty} dx f(x) \rho(x) \quad (D.5)$$

The probability density must be known prior to calculating the average. For quantum theory, the probability density originates in the wave function and therefore, one must know the wave function (i.e., the state of the particle) prior to calculating the average.

The variance of a real-valued random variable x can be written

$$\sigma^2 = \langle (x - \bar{x})^2 \rangle \quad (\text{D.6})$$

where σ is the standard deviation. The term $(x - \bar{x})$ measures the deviation between x and its average. The average of all of the terms $(x - \bar{x})$ gives zero since, by definition of average, x is larger than $\bar{x}$ as often as it is smaller. Taking the square $(x - \bar{x})^2$ makes the term always positive and it still tends to measure the deviation between x and $\bar{x}$. We are not interested in a point-by-point difference $(x - \bar{x})^2$ but instead, we want the expected behavior over all the possible values. Therefore, the variance is defined with the average in Equation D.6.

For a complex-valued random variable z , the average is given similar to Equation D.5. The average can have real and imaginary parts. The variance must be real (as a measure of total deviation) and is given by

$$\sigma_z^2 = \langle (z - \bar{z})^*(z - \bar{z}) \rangle = \langle |z - \bar{z}|^2 \rangle \quad (\text{D.7})$$

The probability density leads to a probability for discrete random variables rather than those having a continuous range as appropriate for the probability density. We convert the integral for the average of an arbitrary function $\langle f \rangle = \int_0^L dx f(x)\rho(x)$ into a discrete summation. First divide the region of integration $(0, L)$ into very small intervals δx so that $L = N \delta x$. Let x_i be a point in interval $\#i$. Assume the interval δx centered on the interval $\#i$ small enough that a measurement of x produces value x_i with probability Π_i . The probability Π_i must be $\Pi_i = \rho(x)\delta x$. Notice the units of Π_i and ρ differ. We can now write

$$\langle f \rangle = \int_0^L dx f(x)\rho(x) \cong \sum_{i=1}^N f(x_i)\Pi_i \quad (\text{D.8})$$

This suggests writing the discrete form of the probability density as

$$\rho(x) = \sum_{i=1}^N \Pi_i \delta(x - x_i) \quad (\text{D.9})$$

D.2 PROCESSES

Making a series of measurements of a quantity Y produces a set of discrete points $\{y\}$. Each measurement takes place at a separate time t_i . For example, we might measure slight fluctuations in optical power $P(t)$ from a laser as a function of time (Figure D.1). Each sequence of points (i.e., each possible graph like Figure D.1) produces a realization of a random process. For a given value of the parameter t , the quantity $Y(t)$ represents a random variable. The collection of random variables $\{Y\}$, with one such Y for each t , constitutes the random process. The set $\{y_i; i = 1, 2, \dots\}$ provides a representation of the random process. These sets might be so dense as to approximate a continuous set.

Consider an example for the power $P(t)$ from a laser or light emitting diode. Measurements at time $\{t_i\}$ produce results $\{P_i\}$ that can be plotted as points on a graph. The time t serves as an “index” for the points (t_1, P_1) , (t_2, P_2) , and so on. Each sequence of points $P(t)$ (i.e., each possible graph like Figure D.1) represents a realization of the random process. Let t_1 be a specific time.

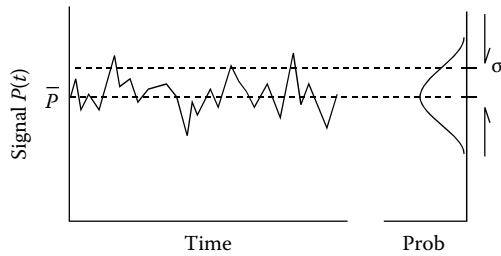


FIGURE D.1 A signal as a function of time. A large amount of noise is superimposed on the average signal.

The power $P(t_1)$, a random variable, can assume any number of values. That is, for a given fixed time t_1 , the value of P can assume a range of values. For example, P might be in the range $[-1,1]$ or it might assume a set of discrete values in that range. Therefore, for every time t_1, t_2 , and so on, there exists a random variable $P_1 = P(t_1)$, and another random variable $P_2 = P(t_2)$, and so on. The collection of all random variables forms the random process $P(t)$. Sometimes people refer to quantities such as $P(t)$ as a time-dependent random variable rather than a process.

A probability density ρ describes the distribution of possible values at a given value of the parameter. The probability density $\rho(y_1, t_1)$ refers to a single random variable $Y_1 = Y(t_1)$ indexed by a particular value of the parameter t_1 . The quantity $\rho(y_1, t_1)$ represents the probability (density) for finding the random variable Y_1 has the value y_1 at the specific time t_1 . The joint probability density $\rho(y_1, t_1, y_2, t_2)$ gives the probability of finding the value of $Y_1 = y_1$ at t_1 and $Y_2 = y_2$ at t_2 . Sometimes people refer to $\rho(y_1, t_1, y_2, t_2)$ as the “two-time probability density.” Notice that the two-time probability refers to two separate values of the parameter for the same process. The multitime probability density provides more information than does the single-time probability density. As will be seen momentarily, the multitime probabilities contain information on correlation.

D.3 ENSEMBLES

An ensemble consists of the collection of “all possible” realizations of the random process. For example, consider an experiment to measure the optical power $P(t)$ at times $t_1, t_2, \dots$ from a semiconductor laser. Suppose the experiment starts at 2 PM on December 3, ends at 3 PM, and produces the results given by plot #1 in Figure D.2. Now suppose the experimenter goes back in

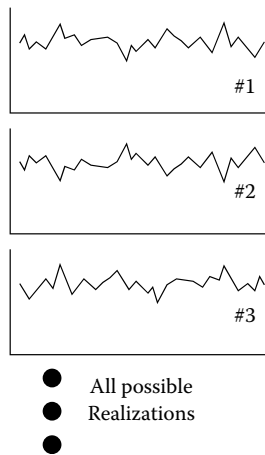


FIGURE D.2 The ensemble consists of all possible realizations.

time to 2 PM on December 3 and repeats the experiment. Realization #2 in Figure D.2 shows this data set. In fact, suppose that the experimenter goes back an infinite number of times and collects all possible realizations. That collection represents the ensemble. Of course, we only imagine going back in time and obtaining the ensemble; we cannot really collect the information. Sometimes we focus on a single time t such as t_1 . An ensemble might consist of all possible values $P(t_1)$. The average power $\bar{P}(t) = \langle P(t) \rangle$ can be found by averaging all of the possible realizations of the process at time t_1 . Using the density function, the average becomes

$$\bar{P}_1 = \bar{P}(t_1) = \int dP_1 P_1 \rho(P_1, t_1)$$

That is, the average is found by a point-by-point average over the infinite number of possible points at time t_1 .

D.4 STATIONARY AND ERGODIC PROCESSES

A process receives the designation of “stationary” when its characteristics do not change with time. For example, the average and the standard deviation do not depend on time. The time-dependence of the probability distribution determines the stationary character of a process. Consider again the power from a laser. The single-time probability distribution $\rho(P_1, t_1)$ should not depend on time for a stationary process. However, a multitime probability distribution $\rho(P_1, t_1; P_2, t_2; P_3, t_3; \dots)$ describing for example, the power P_i in an optical beam at time “ t_i ,” depends only on a difference in time.

$$\rho(P_1, t_1; P_2, t_2; P_3, t_3 \dots) = \rho(P_1, 0; P_2, t_2 - t_1; P_3, t_3 - t_1 \dots) \quad (\text{D.10})$$

Some stationary processes have the designation of “ergodic” when the average such as $\bar{P} = \langle P(t) \rangle$ can be found by either (1) the ensemble average or by (2) a time-average. The two averages must produce identical results for the process to be ergodic. The time-average has the usual definition

$$\langle P \rangle_T = \frac{1}{N} \sum_{i=1}^N P(t_i) \quad \langle P(t) \rangle = \frac{1}{T} \int_0^T dt P(t) \quad (\text{D.11a})$$

while the ensemble average uses only a single time t_i and calculates

$$\bar{P}_i = \bar{P}(t_i) = \sum P_i \rho(P_i, t_i) \quad \langle P_i \rangle = \int dP_i P_i \rho(P_i, t_i) \quad (\text{D.11b})$$

The distinction will become clear in the following examples. Strictly speaking, a process can only be ergodic if every realization contains exactly the same statistical information as the ensemble. In this case, the realizations do not all need to start at the same time.

Example D.1: Nonstationary Process

Figure D.3 shows two examples of nonstationary processes. The first one shows that the standard deviation of the noise $a(t)$ decreases with time. The second one shows that the average value of $b(t)$ decreases with time.

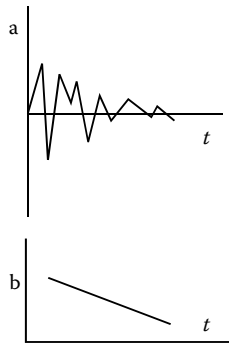


FIGURE D.3 A nonstationary processes.



FIGURE D.4 Two realizations of a nonergodic process.

Example D.2: Nonergodic Process

Figure D.4 shows a nonergodic process because the standard deviation differs for two different realizations (perhaps taken at widely different times).

D.5 CORRELATION

This section discusses the meaning of correlation for a single random variable and cross correlation for two random variables. Two random variables X and Y are correlated if the values of one are “linked” (to some extent) with the values of the other. Probability and statistics courses define the covariance. We freely interchange the names correlation and covariance. The correlation (or perhaps more properly the covariance) of two random variables X and Y is defined by

$$\Gamma_{XY} = \text{cov}(X, Y) = \langle (X - \overline{X})(Y - \overline{Y})^* \rangle \tag{D.12}$$

The complex conjugate only applies to complex-valued random variables; the correlation function generally has real and imaginary parts. Sometimes Γ_{XY} is interpreted as an element of a matrix

(the covariance matrix); the elements are Γ_{XX} , Γ_{XY} , Γ_{YX} , Γ_{YY} . The “correlation coefficient” is defined as

$$\frac{\langle (X - \bar{X})(Y - \bar{Y})^* \rangle}{\sigma_X \sigma_Y} \quad (\text{D.13})$$

where σ_X and σ_Y are the standard deviations for X and Y , respectively. The complex conjugate only applies to complex-valued random variables. Both the correlation function and the correlation coefficient measure the linkage between two random variables. However, the correlation coefficient removes arbitrary scaling factors. As an important note, if $X=Y$ then the correlation function $\Gamma_X = \Gamma_{X,Y}|_{X=Y}$ reduces to the usual variance according to

$$\Gamma_{XX} = \langle (X - \bar{X})(Y - \bar{Y})^* \rangle = \langle |X - \bar{X}|^2 \rangle = \sigma_X^2 \quad (\text{D.14})$$

Figure D.5, as an example, shows two sets of measured values exhibiting positive and negative correlation, and a third exhibiting negligible correlation. The values x_i and y_i have positive cross correlation for the set marked “pos” because, as the values of one increase, so do the values of the other. For example, the set of points might represent the $x-y$ position of an ant as it follows a scent across a tabletop. The subscript “ i ” represents the time (in seconds) on a clock. For the eight points, the cross correlation between x and y is

$$\langle (x(t) - \bar{x})(y(t) - \bar{y}) \rangle = \frac{1}{8} \sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y})$$

The cross correlation is positive for the “pos” case in Figure D.5.

Next consider the autocorrelation function defined by

$$\Gamma_X = \langle X(t)X^*(t + \tau) \rangle = \Gamma_X(t, t + \tau) \quad (\text{D.15})$$

Equation D.14 looks similar to the cross correlation function in Equation D.12. In some sense, the term $X^*(t + \tau)$ acts like a new random variable $Y(t)$. This brings us back to interpreting t and $t + \tau$ as indices in a sequence of measured values; the symbol τ is then similar to an offset. The autocorrelation function measures the similarity between two subsets of a single string of numbers.

The following set of examples lead the reader to the meaning of the correlation and autocorrelation of the Langevin noise sources.

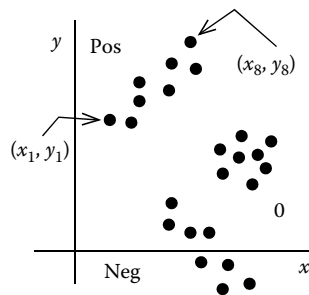


FIGURE D.5 Three types of cross correlation.

Example D.3: Correlation (for Illustration Purposes)

Consider a discrete process with realization given by

$$\underbrace{x_0, x_1, \dots, x_i}_{\leftarrow n \rightarrow} \dots$$

that is, $x(t_0) = x_0$, and so on. The correlation between the set $x_0, x_1, \dots$ and the set $x_i, x_{i+1}, \dots$ must be given by

$$\Gamma = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_{i+n} - \bar{x})$$

where a string of N numbers is taken for each subset. This is the same as the autocorrelation. Notice that if the offset $n=0$ then the autocorrelation becomes the variance

$$\Gamma = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_{i+n} - \bar{x}) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x}) = \sigma_x^2$$

We have not been careful to properly define estimators, which would require N to be replaced by $N-1$.

Example D.4: Autocorrelation

Suppose a coin has sides labeled with $+1$ and -1 . Suppose 22 tosses of the coin yields the following string of numbers.

$$x_0 = -1, +1, -1, -1, +1, +1, -1, +1, -1, +1, -1, +1, \\ -1, +1, +1, +1, -1, -1, +1, +1, +1, -1 = x_{22}$$

Consider two small subsets with $N=7$ elements. Suppose the first subset starts at x_0 and the second one starts at x_{12} .

$$x_0 = -1, +1, -1, -1, +1, +1, -1 = x_6 \quad x_{12} = -1, +1, +1, +1, -1, -1, +1 = x_{18}$$

The correlation between these two sets (assuming $\bar{x} \cong 0$ for convenience) is therefore $-3/7$. This is the autocorrelation because the two subsets are from the same initial string. For the coin toss, the 7-number sets could produce a correlation value anywhere between -1 and $+1$ (with 0 as the expected outcome so long as the sets are different). For this case, the offset is $n=12$.

Example D.5: The Kronecker-Delta Correlation

For the previous example, what is the autocorrelation for $n=0$? The answer is 1. It is not too hard to imagine a situation where the correlation is 0 for $n \neq 0$. The correlation as a function of “ n ” is then

$$\Gamma_x(n) = \sigma^2 \delta_{n,0} = \begin{cases} \sigma^2 & n = 0 \\ 0 & n \neq 0 \end{cases}$$

where $\delta_{a,b}$ is the Kronecker-delta function which is 1 when $a=b$ and 0 otherwise.

Appendix E: Review of Integrating Factors

This appendix contains a quick review of integrating factors as used for solving first order differential equations. Suppose we want to solve the equation

$$\dot{y} - ay = f(t) \quad (\text{E.1})$$

where $y = y(t)$ and the dot indicates the first derivative with respect to time. Suppose we multiply through by a function $\mu(t)$, the integrating factor

$$\mu\dot{y} - a\mu y = \mu f(t) \quad (\text{E.2})$$

with the particular property that the left-hand side is an exact derivative

$$\frac{d}{dt}(\mu y) = \mu\dot{y} - a\mu y \quad (\text{E.3})$$

Then we could write Equation E.2 as

$$\frac{d}{dt}(\mu y) = \mu f(t)$$

If the forcing function $f(t)$ starts at $t = 0$, we can integrate both sides of the equation with respect to time to obtain

$$\mu(t)y(t) = \mu(0)y(0) + \int_0^t d\tau \mu(\tau) f(\tau)$$

or

$$y(t) = \frac{\mu(0)y(0)}{\mu(t)} + \frac{1}{\mu(t)} \int_0^t d\tau \mu(\tau) f(\tau) \quad (\text{E.4})$$

Once we know the integrating factor $\mu(t)$ then we also know the form of the solution even when the exact form of the forcing function has not been specified. This is the property that makes the integrating factor useful for our purposes.

How do we find the integrating factor? Use Equation E.3 and expand the derivative

$$\frac{d}{dt}(\mu y) = \mu\dot{y} + \dot{\mu}y \quad (\text{E.5})$$

Combining Equations E.3 and E.5 we find

$$\mu\dot{y} + \dot{\mu}y = \mu\dot{y} - a\mu y$$

to arrive at

$$\dot{\mu} = -a\mu$$

By separating variables, this simple first order differential equation has the solution

$$\mu(t) = e^{-at}$$

Notice that constants of integration are unimportant for integrating factors—they cancel out of the final equation.

Appendix F: Group Velocity

The “group” velocity describes a type of average speed of a wave packet. A wave packet consists of many sinusoidal waves with each having a specific wavelength and frequency. That is, a wave packet consists of the superposition of multiple plane waves. “Phase” velocity describes the speed of a single sinusoidal wave with a single frequency. The phase velocity of the plane wave

$$\psi(x, t) = A_k e^{ikx - i\omega t} \quad (\text{F.1})$$

can be found by watching the motion of a single point of the wave. Focus on the point initially at $x = 0$ at $t = 0$. Setting the phase to zero

$$kx - \omega t = 0 \quad (\text{F.2})$$

provides the phase velocity

$$v_p = \frac{x}{t} = \frac{\omega}{k}$$

The group velocity describes the average speed of “wave packets” traveling in a dispersive medium. Plane waves with different frequencies travel with different phase velocities in a dispersive medium. For optics, this means that the index of refraction depends on wavelength. Wave packets can represent photons, electrons, holes, and phonons (and so on). These wave packets can perhaps be most conveniently pictured as traveling Gaussian waves $f(z, t)$ as indicated in Figure F.1 although they can have any arbitrary form. As we will shortly discuss, these Gaussian waves are “envelope” wave functions.

The Fourier transform of the wave packet appears in Figure F.2 that shows the amplitude $\varphi(k)$ of the various spectral components plotted against the wave vector. For an optics example, the wave packet and its Fourier transform might describe a pulse of light. Suppose the center wavelength corresponds to green and the smaller amplitudes on either side of the center correspond to red and blue (see Figure F.3). Obviously, the average wave vector $k = 2\pi/\lambda$ denoted by $\bar{k}$ cannot be anywhere near zero! Also, some pulses have narrow Fourier transforms $\phi(k)$ (unlike the one shown in Figure F.3).

For a nondispersive medium, the wave packet shown in Figure F.1 does not spread because all of the constituent components travel at the same speed. On the other hand, a dispersive medium (such as glass) requires the components to travel at different speeds. This means that the wave packet will spread out with time. A dispersive medium does not require the various components making up the pulse to interact with each other. Two spectral components can interact with each other in a nonlinear medium. For example, a blue component might get larger at the expense of two nearby infrared components.

One issue concerns the motion of a packet as compared with the motion of a plane wave. This is especially important for dispersive media where $\omega = \omega(k)$ or equivalently, $E = E(k)$ (and definitely applies to massive particles in free space for the quantum theory). For optics, the relations are especially easy to picture. Consider the speed of the wave. If we write the phase velocity of a given plane wave as $v = \frac{\omega(k)}{k}$, we see that different colors travel at different speeds (this is dispersion). For example, blue light interacts more with a piece of glass than does red light; therefore blue light runs slower (some materials are the reverse of this behavior). It is also blue light that is most deflected from its straight-line path by a glass prism (the index of refraction is larger for blue). As an example,

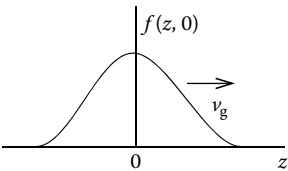


FIGURE F.1 A wave packet moving to the right with group velocity v_g .

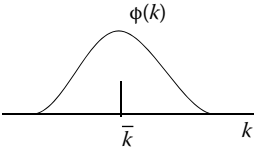


FIGURE F.2 The Fourier transform of the wave packet f .

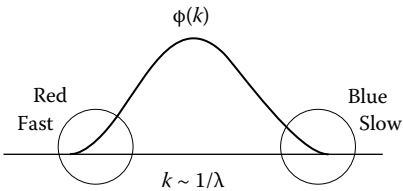


FIGURE F.3 Various colors of light travel faster or slower than the average. Note that “ k ” refers to the carrier wave vector.

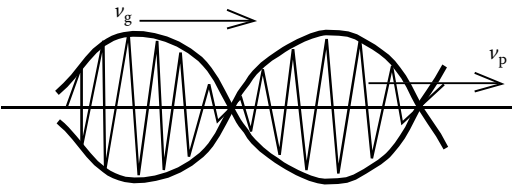


FIGURE F.4 Envelope and phase velocity.

consider Figure F.3 showing that certain colors of light travel faster than the average while others travel slower. We might expect the width of the Gaussian to change as some of the waves run slower than an average while others run faster. The issue becomes one of describing the motion of the wave packet (the envelope) in spite of the fact that the various components travel at different speeds. The phase velocity is not the correct measure. Usually, people describe the wave packet as consisting of a slowly varying envelope function superimposed on the fast moving carrier waves. The function in Figure F.1 provides one example of the envelope, and Figure F.4 provides another for two superimposed sine waves with nearly identical frequencies and wave vectors (discussed in the next paragraph). The envelope function is very long compared with the small wavelength carrier. The figure shows the group velocity v_g describes the speed of the envelop.

F.1 SIMPLE ILLUSTRATION OF GROUP VELOCITY

We can easily understand how the envelope can travel slower (much slower) than the plane waves by considering a simple example of adding two traveling sine waves together. We will work the same example in two ways that both lead to the same conclusion. First, assume that κ, κ' and ω, ω'

are wave vectors and angular frequencies and that they are very close together in value. Assume two sine waves travel parallel to each other.

$$\begin{aligned} y &= A \sin(\kappa x - \omega t) + A \sin(\kappa' x - \omega' t) \\ &= 2A \cos\left(\frac{\kappa - \kappa'}{2}x - \frac{\omega - \omega'}{2}t\right) \sin\left(\frac{\kappa + \kappa'}{2}x - \frac{\omega + \omega'}{2}t\right) \end{aligned} \quad (\text{F.3})$$

These last equations show that the summation of the two sine waves can be viewed as another sine wave with modulated amplitude. We can identify the carrier as

$$\sin\left(\frac{\kappa + \kappa'}{2}x - \frac{\omega + \omega'}{2}t\right) \quad (\text{F.4})$$

having approximate wave vector and frequency of

$$\frac{\kappa + \kappa'}{2} \sim \kappa \quad \text{and} \quad \frac{\omega + \omega'}{2} \sim \omega$$

(since $\kappa \cong \kappa'$ and $\omega \cong \omega'$). The envelope (modulation) function must be

$$\cos\left(\frac{\kappa - \kappa'}{2}x - \frac{\omega - \omega'}{2}t\right) \quad (\text{F.5})$$

The envelope function has a very long wavelength encompassing many cycles of the sine term since

$$\kappa - \kappa' \ll \kappa \quad \rightarrow \quad \lambda_{\text{env}} = \frac{2\pi}{(\kappa - \kappa')/2} \gg \lambda = \frac{2\pi}{\kappa}$$

As far as Fourier series and transforms are concerned, the results seems a little unfamiliar because we are adding two high-frequency waves whereas we normally add two low frequency waves (with equal speed) to get a square wave, etc. Anyway, to continue, the speed of the carrier wave is approximately $v_p = \frac{\omega}{k}$ and the speed of the envelope is

$$v_{\text{env}} \cong \frac{\omega - \omega'}{\kappa - \kappa'} = \frac{\Delta\omega}{\Delta k} \cong \frac{d\omega}{dk} \quad (\text{F.6})$$

Notice we only required two waves (at high frequency) with slightly different phase velocities $v_p = \omega/k$. So the wave packet motion is really the motion of the beat wave.

There is another way to see this result that perhaps better illustrates the role of the different speeds of the two individual waves. Figure F.4 shows the sum of two waves

$$y = y_1 - y_2 = A \sin(\kappa x - \omega t) - A \sin(\kappa' x - \omega' t) \quad (\text{F.7})$$

near $x=0$ and $t=0$. The minus sign for the second term is chosen so that the envelop function crosses zero near $x=0$ for convenience. The point where the envelope crosses through zero depends on the relative positions of the two waves y_1 and y_2 . If one wave moves faster than the other one then

the zero point of $y_1 - y_2$ must move. Near the origin $x=0$ and $t=0$ both $y_1 - y_2$ and the envelope crosses zero. To find the group velocity, consider x and t to be very small but not necessarily zero. Focus on the zero-point crossing by setting the sum of the two waves $y_1 - y_2$ to zero to find

$$0 = y = y_1 - y_2 = A \sin(\kappa x - \omega t) - A \sin(\kappa' x - \omega' t) \cong A[(\kappa x - \omega t) - (\kappa' x - \omega' t)]$$

from the lowest order Taylor approximation. Solving for $v_g = x/t$ provides

$$v_g = \frac{x}{t} = \frac{\omega - \omega'}{k - k'} = \frac{\Delta\omega}{\Delta k} \cong \frac{d\omega}{dk} \quad (\text{F.8})$$

similar to the previous result. Figure F.5 illustrates the results. The top portion of the figure shows the superposition of two sine waves at $t=0$. The wave vectors and angular frequencies are $k_1 = 1$, $k_2 = 1.03$, $\omega_1 = 10$, and $\omega_2 = 10.1$ which gives two slightly different phase velocities nearly equal to 10. The slight difference in the wave vectors yields a group velocity three times smaller than the phase velocity. Focus on the point in the top portion of Figure F.5 where the envelope passes through zero. The bottom portion shows a close-up view for three different times: $t=0$, $t=0.03$,

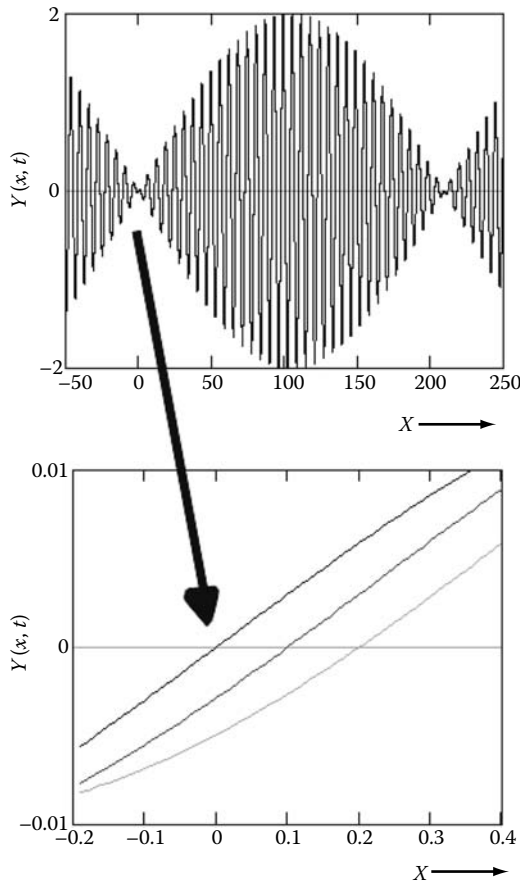


FIGURE F.5 Focus on the point where the envelop crosses zero. The velocity with which it moves to the right is the same as the group velocity.

and $t = 0.06$. Notice how the zero-point crossing moves to the right; this motion corresponds to the envelope (wave packet) moving toward the right (top portion). You can measure directly from the lower portion of the figure or calculate $v_g = (\omega_2 - \omega_1)/(k_2 - k_1)$ to find a group velocity of $v_g = 3.3$.

F.2 GROUP VELOCITY OF THE ELECTRON IN FREE SPACE

The above considerations apply equally well to the wave motion of electrons. This is especially true for free space since the free-space dispersion relation is

$$E = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m} \quad (\text{F.9})$$

Using $\omega = E/\hbar$, we see that the phase velocity $v_p = \omega/k$ depends on k according to

$$v_p = \frac{\hbar k}{2m} \quad (\text{F.10})$$

(note the extra factor of 2). The reason for the k -dependence of the phase velocity in Equation F.10 is that $\hbar k$ is related to the particle momentum (however, infinitely long plane waves do not intuitively represent particles very well). The point of Equation F.10 is that the phase velocity of the electron depends on the wave vector (i.e., wavelength) even for a free particle. The free photon propagating through free space behaves completely different. The speed of light in free space is independent of the wave vector since the speed of light $c = \omega/k$ is constant for all EM waves. The previous section shows that the group velocity for a dispersion relation such as F.9 must be

$$v_g = \frac{\partial \omega}{\partial k} = \frac{\partial}{\partial k} \frac{\hbar k^2}{2m} = \frac{\hbar k}{m} \quad (\text{F.11})$$

F.3 GROUP VELOCITY AND THE FOURIER INTEGRAL

Now is a good time to talk about the mathematics for group velocity. Suppose $f(x, t)$ is a wave packet made up of a discrete set of spectral components—this is a good illustration of converting summations to integrals.

$$f(z, t) = \sum_j c_j e^{i(k_j z - \omega_j t)} \quad (\text{F.12})$$

For each j , there is a k , so relabel the sum as

$$f(z, t) = \sum_k c_k e^{i(kz - \omega_k t)} \quad (\text{F.13})$$

We are considering a one-dimensional problem in k -space. Assume that the sums over an extremely large number of k -values. In fact, let $\rho(k)dk$ be the number of k -values in the length dk . The summation $\sum_k \dots$ can be changed to the integral

$$f(z, t) = \int_{-\infty}^{\infty} dk c(k) \rho(k) e^{i\{kz - \omega(k)t\}} = \int_{-\infty}^{\infty} dk c(k) \rho(k) e^{i\{kz - \omega(k)t\}}$$

where ρ is the density of states (# k -values per unit length of k). Next defining the Fourier amplitude for $f(x, 0)$ as $\varphi(k) = \sqrt{2\pi} c(k)\rho(k)$ we find the expansion

$$f(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk \varphi(k) e^{i\{kz - \omega(k)t\}} \quad (\text{F.14})$$

The wave packet f and its Fourier transform $\varphi(k)$ appear in Figures F.2 and F.3. We could have started with Equation F.14 directly, but sometimes it is nice to see how the individual modes make up the wave packet.

An average wave vector $\bar{k}$ and angular frequency $\bar{\omega}$ characterize the wave packet (as in Figure F.2). For a wave packet with a very narrow spread in frequency and wave vector, we can write a Taylor expansion for the angular frequency (keeping only two terms)

$$\omega(k) \cong \omega(\bar{k}) + \left. \frac{\partial \omega}{\partial k} \right|_{\bar{k}} (k - \bar{k}) \equiv \bar{\omega} + \bar{\omega}' \cdot (k - \bar{k}) \quad (\text{F.15})$$

Substituting this last result into $f(z, t)$ in Equation F.14, we find

$$f(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk \varphi(k) e^{ikz} e^{i(k-\bar{k})z} e^{it\{\bar{\omega} + \bar{\omega}' \cdot (k-\bar{k})\}} = e^{i\bar{k}z - i\bar{\omega}t} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk \varphi(k) e^{i(k-\bar{k})z} e^{-it\bar{\omega}'(k-\bar{k})}$$

Defining a new variable that shows the deviation between the wave vector and its average as $k' = k - \bar{k}$ we find

$$f(z, t) = \underbrace{e^{i\bar{k}z - i\bar{\omega}t}}_{\text{phase-factor}} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk' \varphi(k' + \bar{k}) e^{ik'z} e^{-it\bar{\omega}'k'}}_{\text{Envelope}} = e^{i\bar{k}z - i\bar{\omega}t} f(z - t\bar{\omega}', 0) \quad (\text{F.16})$$

The leading phase factor is unimportant for our purposes. Equation F.16 defines $\varphi(k' + \bar{k}) = \varphi(k')$ to replace the original function $f(z, t)$ by the envelope function

$$f(z - \bar{\omega}'t, 0) = f\left(z - t \frac{\partial \bar{\omega}}{\partial k}, 0\right) \quad (\text{F.17})$$

where

$$f\left(z - t \frac{\partial \bar{\omega}}{\partial k}, 0\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk' \varphi(k') e^{ik'\{z - t\bar{\omega}'\}}$$

To interpret Equation F.17, if the wave packet had the value f_0 at point $z_0 = 0$ at time $t = 0$, then (to lowest approximation) it has the same value at the point $z = z_0 + \bar{\omega}'t$ at time t . On average, the wave packet moves with speed (group velocity)

$$v_{\text{group}} = \frac{\partial \omega}{\partial k} \quad (\text{F.18})$$

As a note, if f above is the electric field (for EM) or the probability amplitude (for QM), then the power or the probability becomes

$$f^*f = \left[e^{i\bar{k}z - i\bar{\omega}t} f(z - t\bar{\omega}', 0) \right]^* \left[e^{i\bar{k}z - i\bar{\omega}t} f(z - t\bar{\omega}', 0) \right] = |f(z - t\bar{\omega}', 0)|^2$$

For electromagnetics and quantum theory, it is the modulus-squared that has physical significance and the phase factor $e^{i(\bar{k}z - t\bar{\omega})}$ drops out. Equation F.17 shows that the wave packet does not change shape as it moves to the right with the group speed

$$v_g = \frac{\partial \omega}{\partial k}$$

All of the manipulations used for the Fourier transform also hold for the periodic discrete case.

F.4 THE GROUP VELOCITY FOR A PLANE WAVE

Consider a single frequency component

$$\varphi(k) = \delta(k - k_0)$$

where

$$\delta(k - k_0) = \begin{cases} \infty & k = k_0 \\ 0 & k \neq k_0 \end{cases} \quad \text{such that} \quad \int_{-\infty}^{\infty} dk \delta(k - k_0) = 1$$

(refer to the Dirac delta function). Equation F.14

$$f(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk \varphi(k) \exp(ikz - i\omega_k t)$$

reduces to

$$f(z, t) = \frac{\exp(ik_0 z - i\omega_{k_0} t)}{\sqrt{2\pi}}$$

so that the phase velocity must be identical with the group velocity.

Appendix G: Note on Combinatorials

This appendix reviews the concepts behind permutations and combinations from probability theory.

G.1 PERMUTATIONS

Suppose we have N distinguishable objects. For example, consider N balls labeled by an integer between 1 and N . Do not allow repeated numbers. Suppose we have exactly N buckets and we place only one ball into a given bucket. The number $N!$ gives the number of possible different arrangements. To see this, consider the following construction. You can place any of the N balls into the first bucket. After you choose 1, there remains $N - 1$ balls. For the second bucket you can choose any of $N - 1$ balls. After you choose 1, there remains $N - 2$ and so on. Therefore, the number of permutations must be

$$N(N - 1)(N - 2) \dots (2)(1) = N! \quad (\text{G.1})$$

G.2 COMBINATIONS OF TWO DIFFERENT TYPES

Suppose we have N buckets and n identical objects with $n < N$. How many different ways can we arrange the n objects in the N buckets. Assume only one ball per bucket. We can write the answer as

$$\binom{N}{n} = \frac{N!}{n!(N - n)!} \quad (\text{G.2})$$

The problem of demonstrating this result can be simplified (even though it might seem more complicated). Suppose the n objects are red balls. Suppose we make one arrangement. For example the first n buckets have balls and the last $N - n$ buckets have none. Instead of having empty buckets we could say that the $N - n$ buckets are filled with green balls for example. The two situations must be equivalent because the extra green balls can be arranged in any manner without affecting the way the original red balls were placed (Figure G.1). In fact, this reasoning can be extended to N objects made up of n_1 alike, n_2 alike, and n_3 alike and so on.

To show the formula, assume that contrary to the assumption of the problem, that integers 1 to N appear on all balls; every ball has its own number and none of the numbers repeat. We might as well assume that integers up to n label red balls and integers $n + 1$ to N label the remaining $N - n$ green balls. The number of ways to arrange the N balls must be given by $N!$ according to the last section. However, this number of permutations assumes that we can distinguish between all of the balls. We can distinguish them if we focus on the numbers. However, we know that the n red balls cannot be distinguished and the $N - n$ green balls cannot be distinguished.

So, suppose that we place down one combination of balls into the buckets. Suppose for simplicity that the first n buckets have red balls and the last $N - n$ buckets have green balls. We do not want to distinguish between the red balls. So, when we place the red balls into the buckets, we chose a first one out of n balls, we then chose a second one out of $n - 1$ balls, and so on. That means, there are $n!$ permutations of the balls that cannot be distinguished. Therefore, out of the $N!$ permutations of all

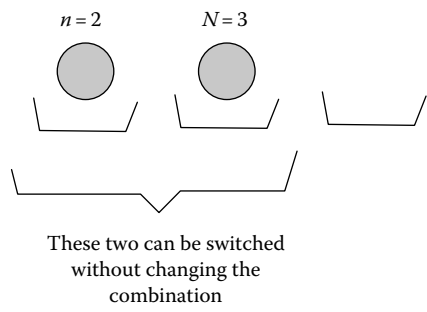


FIGURE G.1 An example.

the balls, we must factor out the permutations of $n!$ that cannot be distinguished on the basis of color. So now the number of different combinations of red must be $N!/n!$.

We must still finish with the green balls that occupy the remaining $N - n$ buckets. The green balls can also be permuted without affecting the original combination of n red balls in the first n buckets and the $N - n$ green balls in the remaining buckets. The green balls can be permuted in $(N - n)!$ ways. Therefore, the original number of permutations based on the integers painted on the sides of the balls over counts the number of combinations by $(N - n)!$ based on color alone. Therefore, the number of combinations of n -alike objects in N buckets must be given by $N!/n!/(N - n)!$ That is

$$\binom{N}{n} = \frac{N!}{n!(N - n)!} \tag{G.3}$$

Actually we can more easily compute this combination by considering N balls (rather than n) of which n are red and $N - n$ are green.

G.3 COMBINATION OF $n_1, n_2, \dots, n_m$ OBJECTS

The number of ways to place n_1 alike objects (red balls) in N buckets while placing n_2 (green balls) alike objects in the N buckets, while placing the n_3 objects (black balls), etc. is given by

$$\frac{N!}{n_1!n_2! \dots n_m!}$$

This is easy to see if we again paint numbers 1 to N on the sides without regard to color of the ball. We might as well assume that the red balls have integers 1 to n_1 and so on. There are $N!$ total ways to arrange the N balls if we keep track of the numbers (the numbers make them ALL distinguishable). For any given arrangement, say red balls all in the first n_1 buckets, etc., there are $n_1!$ permutation of red balls, $n_2!$ permutations of green balls, etc. that do not change the original placement of balls (so long as the same colors stay in the given buckets). Therefore, we must divide out the over-counted permutations to get

$$\frac{N!}{n_1!n_2! \dots n_m!}$$

Appendix H: Lagrange Multipliers

The method of Lagrange multipliers makes it possible to minimize (find the extremum) of a function $f(\xi, \eta)$ when it must be consistent with a constraint $g(\xi, \eta) = \text{constant}$. The constraint defines a path in the plain for the parameters ξ, η and also therefore, a path on the surface of the function f (when viewed as a surface in a higher dimensional space than that of the parameters). One looks for a point along the path in parameter space so as to minimize the function f .

H.1 A SLOPE APPROACH TO LAGRANGE MULTIPLIERS

Suppose one wants to minimize or maximize a function $f(\xi, \eta)$ subject to the constraint that $g(\xi, \eta) = c$. For example, imagine a paraboloid $f = \xi^2 + \eta^2$ (as shown in Figure H.1) above the $\xi - \eta$ plane. One wants to know what values of (ξ, η) make the function f as close as possible to the curve $g(\xi, \eta) = c$ which is confined to the $\xi - \eta$ plane. Here “close as possible” means to find the minimum of f above the path $g(\xi, \eta) = c$ such as shown by $P1$ in Figure H.1. Figure H.1 shows the level curves of the function f . A level curve is the contour for which $f = \text{constant}$. For the paraboloid, the level curves are circles. Figure H.2 shows the planar view with the level curves shown in the plane—setting $g(\xi, \eta) = c$ makes ξ, η dependent and thereby forms a path of the form $\eta = \eta(\xi)$. The coordinates of the minimum (ξ_0, η_0) must be positioned on both the path (equation of constraint) and on one of the contours of the function f .

Elementary calculus minimizes the function f by taking a derivative and setting it to zero.

$$0 = df = f_\xi d\xi + f_\eta d\eta \quad (\text{H.1})$$

where for example, $f_\xi = \partial f / \partial \xi$. If $d\xi$ and $d\eta$ were independent variations then we would conclude that $f_\xi = 0$ and $f_\eta = 0$ which would be the apex of the paraboloid at $(\xi, \eta) = (0, 0)$. However, the two variations must be consistent with the constraint g and therefore cannot be independent. We must include the constraint that requires the coordinates (ξ_0, η_0) satisfy $g(\xi, \eta) = c$. The variations $d\xi$ and $d\eta$ must be interrelated through the differential of $g(\xi, \eta) = c$ as in

$$0 = dg = g_\xi d\xi + g_\eta d\eta \quad (\text{H.2})$$

Equations H.1 and H.2 provide equations for the slope $d\eta/d\xi$ of the constraint curve and the level curve (i.e., the slope of the tangent to these curves) at the point (ξ_0, η_0) . Combining Equations H.1 and H.2 produces

$$\frac{d\eta}{d\xi} = -\frac{f_\xi}{f_\eta} = -\frac{g_\xi}{g_\eta} \quad (\text{H.3})$$

As next discussed and demonstrated in Section H.3, the last two terms in the previous equations can be rearranged and set to a common value λ that depends on the constant c , or equivalently the point (ξ_0, η_0)

$$\frac{f_\xi}{g_\xi} = \frac{f_\eta}{g_\eta} = \lambda \quad (\text{H.4})$$

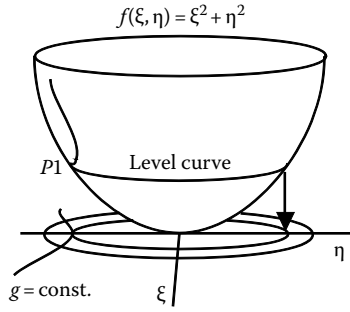


FIGURE H.1 The minimum of function f along the curve g in the q - h plane occurs at point $P1$.

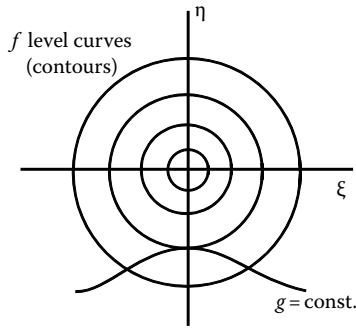


FIGURE H.2 Level curves and the path obtained from the constant $g = \text{const.}$

This result uses the geometric notion of finding a minimum along a path. It does not necessarily give the global minimum of the function but gives a local minimum similar to the distinction between the point $(0, 0, 0)$ and the point $P1$ in Figure H.1.

For the sake of argument, suppose that we had started with the function “ h ” defined by

$$h(\xi, \eta) = f(\xi, \eta) - \lambda g(\xi, \eta)$$

with λ some constant for this case, the function g has been redefined as $g - c \rightarrow g$, which is equivalent to the constraint $g = 0$. Assume that we do not include a constraint for “ h .” The minimum of “ h ” can be found by using the differential

$$0 = dh = (f_\xi - \lambda g_\xi)d\xi + (f_\eta - \lambda g_\eta)d\eta$$

This time, considering the variations $d\xi, d\eta$ to be independent, we require

$$(f_\xi - \lambda g_\xi) = 0 \quad (f_\eta - \lambda g_\eta) = 0$$

Rearranging terms produces Equation H.4 again.

$$\frac{f_\xi}{g_\xi} = \frac{f_\eta}{g_\eta} = \lambda$$

In this case, we have found the global minimum of the function h since we do not have an equation of constraint. However, this procedure gives the same result as the geometrical one because of the way we included the constant λ .

This last procedure provides the Lagrange multiplier method of incorporating constraints. We want to minimize the function $f=f(\xi, \eta)$ subject to the constraint $g(\xi, \eta)=c$ (where c can be zero). Essentially we define the new function h . However, people usually state the procedure as follows. Set the differential of the function f to zero to find

$$0 = f_{\xi}d\xi + f_{\eta}d\eta \quad (\text{H.5a})$$

We treat the variation as independent if we take into account the variation of the constant $g(\xi, \eta)=c$

$$0 = g_{\xi}d\xi + g_{\eta}d\eta \quad (\text{H.5b})$$

Multiply this last equation by the Lagrange multiplier λ and add Equations H.5a and b to find

$$0 = f_{\xi}d\xi + f_{\eta}d\eta + \lambda(g_{\xi}d\xi + g_{\eta}d\eta) = (f_{\xi} + \lambda g_{\xi})d\xi + (f_{\eta} + \lambda g_{\eta})d\eta$$

We therefore find the same result as in Equation H.4

$$\frac{f_{\xi}}{g_{\xi}} = \frac{f_{\eta}}{g_{\eta}} = \lambda$$

An example appears in Section H.4 on how to use these results. The next example perhaps helps to clarify the idea of independent variations.

Example H.1

Give a simple matrix plausibility explanation on why the coefficients A, B in

$$0 = A d\xi + B d\eta$$

Must be zero when $d\xi, d\eta$ are independent.

SOLUTION

If $d\xi, d\eta$ are independent then consider two different sets of values for $d\xi, d\eta$ as for example

$$\begin{aligned} 0.01A + 0.02B &= 0 \\ 0.03A + 0.04B &= 0 \end{aligned} \quad \text{or} \quad \begin{pmatrix} 0.01 & 0.02 \\ 0.03 & 0.04 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Then because the 2×2 matrix can be inverted, the coefficients A, B must be zero.

H.2 THE MULTIDIMENSIONAL RESULT

In general, for a function $f=f(\xi_1, \dots, \xi_n)$ and the constraint $g=g(\xi_1, \dots, \xi_n)$ we want coordinates $(\xi_1, \dots, \xi_n)$ such that $0 = df = \sum_s f_{\xi_s} d\xi_s$ but the variations $d\xi_s$ are not independent. The $d\xi_s$ can be taken as independent so long as we include the constraint by “adding or subtracting” the term $0 = \lambda dg = \sum_s \lambda g_{\xi_s} d\xi_s$. Now we have

$$0 = df - \lambda dg = \sum_s (f_{\xi_s} - \lambda g_{\xi_s}) d\xi_s$$

Now choose the value of λ such that each term is zero

$$f_{\xi_s} - \lambda g_{\xi_s} = 0$$

Now λ has a specific value.

H.3 THE USE OF GRADIENTS AND THE LAGRANGE MULTIPLIERS

This section provides a quick not on the use of the gradient to find the relation for the Lagrange multiplier such as in $f_{\xi} - \lambda g_{\xi} = 0$. Consider the function $f(\xi, \eta)$ and the constraint $g(\xi, \eta)$, which is not set to a constant. The gradient points in the direction of greatest increase of a function. At a point where the level curves of f and g are tangent, the gradients will be either parallel or antiparallel. In such a case, one can write $\nabla_{\xi, \eta} f(\xi, \eta) = \lambda \nabla_{\xi, \eta} g(\xi, \eta)$, where λ is a constant of proportionality and $\nabla_{\xi, \eta} = \tilde{\xi} \partial_{\xi} + \tilde{\eta} \partial_{\eta}$, $\partial_{\xi} = \partial / \partial \xi$, etc. Setting the components equal to each other produces the desired results

$$(f_{\xi} - \lambda g_{\xi}) = 0 \quad (f_{\eta} - \lambda g_{\eta}) = 0$$

H.4 A SIMPLE EXAMPLE FOR THE LAGRANGE MULTIPLIERS

For $f(x, y) = x^2 + y^2$ and $g(x, y) = y - x = c$, find the point that minimizes f consistent with the constraint as shown in Figure H.3.

Setting the differential of the function

$$h(x, y) = f(x, y) - \lambda g(x, y)$$

to zero provides

$$0 = df(x, y) - \lambda dg(x, y) = (f_x - \lambda g_x)dx + (f_y - \lambda g_y)dy$$

Independent variations dx and dy produce

$$(f_x - \lambda g_x) = 0 \quad (f_y - \lambda g_y) = 0$$

We want the points $x = x_0$ and $y = y_0$ that make these last equations hold. Substituting for the partial derivatives, we find

$$2x_0 + \lambda = 0 \quad \text{and} \quad 2y_0 - \lambda = 0$$

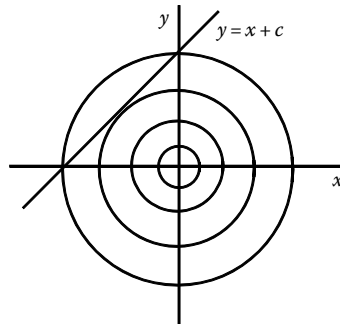


FIGURE H.3

These last two equations minimize the function f ; however, we can go further. These last two equations simultaneously hold and requires $x_0 = -y_0$; notice that only the slope of the equation $x_0 = -y_0$ is determined. We can find λ , x_0 , y_0 by using the constraint “ $g = c$.”

$$c = g(x_0, y_0) = y_0 - x_0 = -2x_0$$

So finally, the point of minimum f consistent with the constraint is

$$x_0 = -\frac{c}{2} \quad y_0 = \frac{c}{2} \quad \lambda = c$$

Appendix I: Comments on System Return to Equilibrium

A system initially disturbed from equilibrium proceeds through a transient period whereby it returns to equilibrium (so long as the disturbing agent is removed). This appendix introduces the decay of excess carriers back to equilibrium. Excess carriers can be injected into the semiconductor by optical absorption of light, electrical injection, or thermal heating. When any of these agents that increase the number of electrons is eliminated, the carrier population must return to the equilibrium values. There is a nice thin book by Rose on photoconductivity and allied problems that addresses some of these issues.

This appendix shows that the relaxation (i.e., decay) of excess carriers in the conduction (CB) and valence band (vb) follow a simple exponential decay with a time constant.

The equilibrium between the CB and vb (similar for other levels in the semiconductor) is maintained by upward and downward transitions. Electrons in the vb can absorb enough thermal energy (i.e., very energetic phonons) to promote them into the CB. On the other hand, the electrons in the CB can decay back into the vb by recombining with the holes in the vb. Thermal equilibrium obtains when the upward and downward transition rates match and when the populations agree with the Fermi-Dirac distributions. Suppose an external agent increases the number of carriers above the equilibrium value. In such a case, once the agent is eliminated, the excess carrier population returns to equilibrium when the downward transition rate is larger than the upward rate.

The transition rates can be written as

$$K \times (\text{number of candidate carriers}) \times (\text{number of available states}) \times \text{probability} \quad (\text{I.1})$$

First consider the upward transitions for an intrinsic semiconductor (no doping). The probability of upward transition from the vb to the CB is proportional to the Boltzmann factor

$$e^{-\frac{E_c - E_v}{kT}} = e^{-\frac{E_g}{kT}}$$

where E_c and E_v are the minimum and the maximum energies of the CB and vbs, respectively, and $E_g = E_c - E_v$ is the bandgap energy. The candidate carriers are the electrons in the CB. Assume that there are N_v and N_c states (“effective” density of states) in the vb and cbs. The number of carriers that can transfer to the CB must be $N_v - p$. The number of available states in the CB must be $N_c - n$. However, n and p are usually small and so $N_v - p \cong N_v$ and $N_c - n \cong N_c$. Therefore, the upward transition rate must be

$$R_{\text{up}} = KN_v N_c e^{-\frac{E_g}{kT}} \quad (\text{I.2})$$

Next consider the downward rate. The probability is taken as 1. The number of candidate carriers residing in the CB is n . The number of available states in the vb is p since those are empty. The downward transition rate is therefore

$$R_{\text{down}} = Knp \quad (\text{I.3})$$

Equating the two rates provides the condition for equilibrium

$$np = N_v N_c e^{-\frac{E_g}{kT}} \quad (\text{I.4})$$

Notice that this last result is exactly the law of mass action discussed in Section 8.6. This last expression must hold for doped semiconductors since we have counted empty states and available electrons. The result does not reference the level of doping.

Finally let us consider minority carrier recombination, which is related to the diffusion of carriers in the diode structure. Consider an n-type semiconductor. Assume the equilibrium carrier densities are n and p and the excess carrier densities are n_e and p_e . The total carrier densities are

$$n_{\text{total}} = n + n_e \quad \text{and} \quad p_{\text{total}} = p + p_e \quad (\text{I.5})$$

Consider upward transitions. The candidate carriers are the electrons in the vb which is approximately N_v . The number of available states in the CB equals the number of empty CB states, which is approximately N_c . Therefore, Equation I.1 provides the upward transition rate

$$R_{\text{up}} = K N_v N_c e^{-\frac{E_g}{kT}} = Knp \quad (\text{I.6})$$

by the law of mass action.

For downward transitions, the number of available carrier is n_{total} . However, the total number of carriers in the n-type semiconductor is approximately equal to the equilibrium value

$$n_{\text{total}} = n + n_e \cong n \quad (\text{I.7})$$

since we assume small disturbances in the carrier number. The number of available states equals the number of holes in the CB, namely

$$p_{\text{total}} = p + p_e \cong p_e \quad (\text{I.8})$$

since by the law of mass action for n-type semiconductors, the number of equilibrium holes is very small

$$n \cong N_d \gg n_i \rightarrow p = \frac{n_i^2}{n} = n_i \frac{n_i}{N_d} \ll n_i \quad (\text{I.9})$$

where n_i is the intrinsic number of electrons (or holes) as described in Section 8.6. Therefore, the downward rate is

$$R_{\text{down}} = K n_{\text{tot}} p_{\text{tot}} \cong K n p_e \quad (\text{I.10})$$

The net transition rate is

$$R_{\text{down}} - R_{\text{up}} = K p_{\text{tot}} n_{\text{tot}} - Knp = K[(p + p_e)(n + n_e) - np] = K[np_e + pn_e] \cong Knp_e \quad (\text{I.11})$$

On the other hand, the net downward transition rate decreases the number of holes, which is p_e . Therefore, we have

$$\frac{dp_e}{dt} = -(Kn) p_e \quad (\text{I.12})$$

Keep in mind that n is a constant in time since it represents the equilibrium number of electrons. The solution to Equation I.12 is

$$p_e = Ce^{-\frac{t}{1/(Kn)}} \quad (\text{I.13})$$

The time constant for the rate of decay can be surmised from the last equation

$$\tau_p = \frac{1}{Kn} \quad (\text{I.14})$$

Notice larger numbers of majority carrier n decrease the lifetime of the excess carrier p_e .

The average distance that a carrier can diffuse before recombining is

$$L = \sqrt{D\tau_p} \quad (\text{I.15})$$

where D is the hole diffusion constant.

Appendix J: Bose–Einstein Distribution

The Bose–Einstein (BE) distribution produces the phonon statistics as described in Section 6.11. The BE distribution can be found by maximizing the number of combinations

$$W = \prod_i \frac{(n_i + g_i - 1)!}{n_i!(g_i - 1)!}$$

Following the same procedure as for Sections 8.3 and 8.5, the Bose–Einstein distribution can be written as

$$f_{\text{BE}}(E) = \frac{\bar{n}_i}{g_i} = \frac{1}{z^{-1} \exp[\beta E] - 1}$$

where “ z ” denotes the fugacity which can be found by requiring the distribution to satisfy the usual constraints. The result is

$$f_{\text{BE}}(E) = \frac{1}{\exp\left(\frac{\hbar\omega}{k_b T}\right) - 1} = \frac{1}{e^{\frac{E}{k_b T}} - 1}$$

This gives the average number $\langle n \rangle$ of phonons in a mode characterized by energy $E = n\hbar\omega$.

$$\langle n \rangle = \frac{1}{e^{\frac{E}{k_b T}} - 1}$$

Nondestructible particles (not phonons) retain the chemical potential in the exponential.

Appendix K: Density Operator and the Boltzmann Distribution

We can define the density operator $\hat{\rho}_r$ is defined through a Boltzmann distribution

$$\hat{\rho}_r = \frac{1}{Z} \exp\left(-\frac{\hat{H}_r}{k_B T}\right)$$

where Z denotes the normalization (partition function)

$$Z = \text{Tr}_r \left\{ \exp\left(-\frac{\hat{H}_r}{k_B T}\right) \right\}$$

Consider the average of an operator $\hat{O}$

$$\langle \hat{O} \rangle = \text{Tr}(\hat{\rho}_r \hat{O}) = \sum_n \langle n | \hat{\rho}_r \hat{O} | n \rangle = \sum_{n,m} \langle n | \hat{\rho}_r | m \rangle \langle m | \hat{O} | n \rangle$$

where the closure relation for the “energy” basis set $\{|n\rangle\}$ has been inserted between the two operators. The energy eigenstates are chosen for the basis since the density operator is diagonal in that basis set. First, evaluate the matrix elements of the density operator.

$$\langle n | \hat{\rho}_r | m \rangle = \frac{1}{Z} \langle n | \exp\left(-\frac{\hat{H}_r}{k_B T}\right) | m \rangle = \frac{1}{Z} \langle n | m \rangle \exp\left(-\frac{E_m}{k_B T}\right)$$

where the factor $1/Z$ can be removed but the result was found by evaluating the expectation of the exponential term and where the last term obtains by operating with the Hamilton on the ket $|m\rangle$. Using the orthogonality of the basis provides

$$\langle n | \hat{\rho}_r | m \rangle = \frac{\delta_{nm}}{Z} \exp\left(-\frac{E_n}{k_B T}\right)$$

and the average of an operator becomes

$$\langle \hat{O} \rangle = \text{Tr}(\hat{\rho}_r \hat{O}) = \sum_n \frac{1}{Z} \exp\left(-\frac{E_n}{k_B T}\right) O_{nn}$$

Notice that this last expression only requires the diagonal matrix elements O_{nn} of the operator $\hat{O}$. The partition function can be similarly evaluated. The expectation value of the operator $\hat{O}$ shows that the density operator for the reservoir gives rise to the Boltzmann probability distribution. The energy levels E_n are expected to be populated according to the thermal distribution.

Appendix L: Coordinate Representations of Schrödinger Wave Equation

This appendix illustrates how the Schrödinger wave equation such as for the harmonic oscillator

$$\left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2} kx^2\right) \Psi(x, t) = i\hbar \frac{\partial}{\partial t} \Psi(x, t) \quad (\text{L.1})$$

can be found from operator-vector form of the equation

$$\left(\frac{\hat{p}^2}{2m} + \frac{1}{2} k\hat{x}^2\right) |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle \quad (\text{L.2})$$

We use the harmonic oscillator as an example with the understanding that other Hamiltonians can be similarly treated.

We begin with Equation L.2 by operating on both sides using the x -coordinate projection operator $\langle x|$ to get

$$\langle x| \left(\frac{\hat{p}^2}{2m} + \frac{1}{2} k\hat{x}^2\right) |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} \langle x|\Psi(t)\rangle$$

where the x -coordinate operator moves past the time derivative. On the left-hand side, insert the unit operator

$$1 = \int |x'\rangle \, dx' \langle x'|$$

between the Hamiltonian operator and the ket $|\Psi(t)\rangle$. We obtain

$$\langle x| \left(\frac{\hat{p}^2}{2m} + \frac{1}{2} k\hat{x}^2\right) \left(\int |x'\rangle \, dx' \langle x'|\right) |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} \langle x|\Psi(t)\rangle$$

The x -terms can be moved under the integral since they do not depend on x' .

$$\int dx' \langle x| \left(\frac{\hat{p}^2}{2m} + \frac{1}{2} k\hat{x}^2\right) |x'\rangle \langle x'| |\Psi(t)\rangle = i\hbar \frac{\partial}{\partial t} \langle x|\Psi(t)\rangle \quad (\text{L.3})$$

The momentum and position operators are diagonal in “ x ” so that

$$\langle x|\hat{p}^2|x'\rangle = \langle x|x'\rangle [\hat{p}(x')]^2 = \delta(x' - x) \left[\frac{\hbar}{i} \frac{\partial}{\partial x'}\right]^2 \quad \text{and} \quad \langle x|\hat{x}^2|x'\rangle = \delta(x' - x) [x']^2$$

since $\hat{x}|x'\rangle = x'|x'\rangle$. Therefore, Equation L.3 becomes

$$\int dx' \delta(x' - x) \left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x'^2} + \frac{1}{2} k x'^2 \right) \langle x' | \Psi(t) \rangle = i\hbar \frac{\partial}{\partial t} \langle x | \Psi(t) \rangle$$

Integrating over the delta function yields

$$\left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2} k x^2 \right) \langle x | \Psi(t) \rangle = i\hbar \frac{\partial}{\partial t} \langle x | \Psi(t) \rangle$$

and, using $\langle x | \Psi(t) \rangle = \Psi(x, t)$, gives the desired results

$$\left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + \frac{1}{2} k x^2 \right) \Psi(x, t) = i\hbar \frac{\partial}{\partial t} \Psi(x, t)$$

Index

A

ABED, *see* Aharanov–Bohm effect device
Absorption, transition probability, 370
Acceptor–donor distributions, 740
Acoustic phonons, oscillation frequency, 501
Acoustic polarizations, monatomic crystal, 507
Adjoint operator, action, 43
Aharanov–Bohm effect device, 2, 27–28
AlGaAs system, 273
Amorphous materials, bandgap, 16
Amorphous silicon bonds, 5
Amplitude
 current-density, 569–570
 for electron in plane wave state, 562–563
Angular frequency, three-level atom, 439
Angular momentum
 component of, 303
 conservation of, 296, 301–303
 definition of, 296–297
 eigenvalues and eigenvectors, 303–305
 fermions, 398
 Hilbert space, 323
 lengths, 300
 magnitude, 297, 303
 multiple systems
 addition, 323–325
 Clebsch–Gordon coefficients, 326–327
 nature, role of, 296
 Newton’s relations, 296
 nonzero, 300
 operators, 298–299
 origin of, 297–298
 pictures for, 299–301
 quantum theory, 297
 rotating point particle, 297
 rotation operator, 301
 spherical harmonics, 305–309
Angular momentum operator, 188
 spherical coordinates, 306
Angular momentum vector, 298
Annihilation and creation operators, 537–538
Antilinear isomorphism, 37
Antisymmetric tensor, 297, 299
Atom absorbing energy, 13
Atomic cluster, 467
Atomic collision/scattering processes, 395
Atomic Hamiltonian, 362
Atomic resonant frequency, 369
Atoms
 binding energy, 8
 electron collision, 13
 energy levels, 7, 439
 equilibrium positions, 216, 491
 potential energy, 7
 s–p hybrid bonds, 8
 transverse wave motion, 492

Automobile, kinetic energy, 32
Avogadro’s number, 699

B

Band-bending effect, 14
 between parallel plates, 610
Band diagrams
 3-D, 602
 and dispersion curves, 478–479
 E–*k*, 609
 energy of electrons, 11
 FBZ, 478–479
 GaAs, 10
 optical transitions, 13
Band-edge diagrams, 14–15, 730
 band bending, 610
 conduction and valence, 14
 E–*k* diagram to produce, 609–610
 for heterostructure with single
 quantum well, 611
 optoelectronic components, 15, 610–611
 PN diode, thermal equilibrium, 20
 single quantum well, 15
Bandgap calculation, 588
Bandgap states
 defects, 15–16
 nonequilibrium statistics, 19–20
 pn junction, 18–19
Bands
 dispersion relation, 620–622
 indirect, 589
 intuitive origin, 9–11
 states, 600
 zoomed-in view of, 590
Band theory, 1
Bandwidth and periodic potential, 616–617
Bar magnets
 direction, 310
 electrons, 310
Basis sets
 for angular momentum, 642
 for degenerate band theory, 643
Basis vectors
 definition of, 248
 linear combination, 105, 249
 string, classical wave, 248
Bell’s theorem, 259, 440, 442
Bennett’s original quantum Turing machine, 433
Bias current, 737
Bipolar junction transistors (BJTs), 1, 24
Bloch plane waves, 718
Bloch wave functions, 584–585
 of energy eigenfunctions, 590
 normalization factor, 631
 orthonormality relation, 594–596, 632–633

proof of, 592–594
 tight binding approximation, 619
 Body-centered cubic (BCC) lattice
 conventional unit cells, 470
 “primitive” vectors, 469
 Bohr magneton, 310, 319
 Boltzmann approximation, 731
 Boltzmann constant, 700
 Boltzmann distribution, 699, 704
 and boson-like particles, 712–717
 canonical distribution, 705
 and degenerate states, 711–712
 density operator, 811
 ensemble, derivation, 708–709
 Fermi–Dirac distribution, 704, 717, 739
 independent, distinguishable subsystems,
 717–718
 states and probability, 704–707
 Taylor expansion, 707
 temperature effects, 707
 thermal equilibrium, 704, 716
 thermal reservoir, derivation, 707–708
 Boltzmann factor, 522, 805
 Boltzmann particle, 725
 Boltzmann probability distribution, 704, 711,
 717, 811
 Bonding orbitals, 463
 Bonding order, 3
 Bonding, periodic table, 5–6
 Bose–Einstein probability distribution, 519, 809
 calculation methods, 523–524
 statistical moments
 m th statistical moment, 524–525
 variance, 525–526
 Boson creation/annihilation operators, 402
 Boson-like particles, 714
 thermal equilibrium, 716
 Boson-like properties, 714
 Boson operators, 414
 Boson particles, 723
 Bosons wave functions, 398
 Boundary conditions, 275, 279
 and interface, 565
 and phonon modes, 503
 Bounded operator, 172–173, 175
 Bragg diffraction and group velocity, 587
 Bras, definition, 38
 Bravais lattice, symmetry operations, 479–480
 Brillouin zone, 519
 Broglie wavelengths, 298
 Brownian motion, 697
 Built-in voltage, 739

C

CAD, *see* Computer-aided design
 CAIBE, *see* Chemically assisted ion beam etchers
 Cal-Tech QED-based computer, 440
 Campbell–Baker–Hausdorff theorem, 142–143
 Canonical ensemble, 702–704
 Carrier thermalization, 722
 Cartesian product, 79
 CB, *see* Conduction band
 Charge density, 588

Chemically assisted ion beam etchers, 747, 754
 Classical field theory
 classical Lagrangian, 224
 Hamiltonian, 224
 Classical Hamiltonian, 265, 286, 409
 Classical Lagrangian, 409
 Classical mechanics, 201
 Classical probability theory, 374
 Classical Turing machine, 433
 Clebsch–Gordon coefficients, 183, 325–327
 Coherent phenomena, 563
 Collisions and drift mobility, 553–555
 Column matrices, 110
 Commutator operator, 140, 161
 Computer-aided design, 747
 Condon–Shortley phase factor, 308
 Conduction band, 10, 729, 735
 electrons drop, 15
 n-type dopant states, 15
 states, 735
 Conductivity
 and mobility, 553
 and resistivity, 552
 Configuration space, 204
 Conservation of momentum, 215
 Conventional matrix notation, 135
 Copenhagen interpretation, 247, 258
 Correlation function, 783–784
 Coulomb interaction, 405
 Creation–annihilation operators, 179–180,
 537–538
 Crystal defect, *see* Defect
 Crystalline material structure, 1
 Crystal plane, intersecting axes, 468
 Crystals
 array of atoms, 461
 atomic basis of, 467
 atoms, identical clusters of, 4
 conservation of momentum in, 518–519
 Crystal symmetries
 and rotations, 481–483
 space and point groups, 479–480
 Current density
 amplitudes of, 569–570
 definition, 551–552
 incident and reflected, 568
 in terms of drift mobility, 553
 surface and charge, 555–556
 Current flow process, 555–556
 Current transfer function, 577
 Cyclic coordinates, 215

D

D’Alembert’s principle, 205
 Dangling bonds, 5, 16
 3-D band structure
 for 3-D crystals, 602–604
 effective mass, 604–608
 1-D crystals
 density of states for, 659
 dispersion relation for, 604
 P-DOS for, 511–512
 wave motion in, 508

- 2-D crystals
 - density of states for, 659–660
 - dispersion curve for, 605
 - of k -space for vectors, 512
 - P-DOS for, 512, 514–515
 - wave motion in, 508–509
- 3-D crystals
 - density of states for, 658, 661–662
 - dispersion curve for, 605
 - infinitely deep well in, 673–676
 - k -state density for, 513
 - in long-wavelength limit, 516–517
- Debye model, 528–529
- Debye temperature, 529
- Defect, 484
- Degeneracy factor, 735
- Degenerate bands, $\vec{k} \cdot \vec{p}$ band theory for
 - basis sets, 642–643
 - Bloch eigenstates, 641
 - effective mass for band, 647
 - eigenvalue equation for periodic functions, 638–639
 - eigenvalues, 646–647
 - Hamiltonian for Kane model, 640–641
 - matrix of Hamiltonian, 643–646
 - wave functions, 648–649
- Degenerate band theory, basis states for, 643
- Density of k -states, 657–659
- Density of states
 - definition of, 375
 - Dirac delta functions, 376
 - vs. energy, 377
- Density operator
 - basis expansion, 386–389
 - coherence, loss of, 394–395
 - off-diagonal elements of, 388–389
 - quantum mechanical averages, 390–391
 - wave function, 382–385
- 3-D Euclidean space, 122
- 2-D Hilbert space, 261
- Diagonal matrices
 - change-of-basis operator, 169–170
 - eigenvectors and eigenvalues
 - diagonalize, 165–166
 - motivation, 162–163
- Diamond structures, 471–472
- 3-D diatomic crystals, 502
- Diatomic linear crystal
 - acoustic branches for 3-D motion, 508
 - phonons in, dispersion curves for, 498–500
- Dielectric constant, 729, 742
- Dim, definition, 133
- Dirac delta forcing function, 422
- Dirac delta functions, 72, 76–77, 83, 185, 410, 424, 763–764, 795, 814
 - basis vectors, representations of, 64–65
 - convergence factors, 773–774
 - coordinate space basis vectors, 61
 - cosine basis functions, 75–76
 - definition of, 48, 764
 - Fourier series basis functions, 77–78
 - Fourier transform, 767, 774, 776–777
 - Kronecker delta, 73–74
 - normalization, 60
 - presentation of, 764
 - representation of, 76, 768–770
 - set of, 61
 - sine basis functions, 77
 - theorems, 770
- Dirac notation, 51, 70, 144, 192, 313
 - euclidean and function spaces, 107
 - matrices, 106
- Direct product spaces, 133
 - continuous basis sets, 85
 - discrete basis sets, 84–85
 - Fourier series, 82–83
 - matrices of, 134
 - conventional matrix notation, 135–137
 - matrix notation, 135
 - matrix representation, 137–138
 - operators and matrices, 133–134
 - overview of, 79–81
 - review of, 133–134
 - single electron, 318–319
 - two euclidean vectors, 82
- Dispersion curve
 - bandgaps in, 586
 - for 3-D motion along 1-D crystal., 495
 - for free electron and nearly free electron, 585
 - functional form of, 500
 - for monatomic crystal
 - with 1-D motion, 492
 - in FBZ, 493
 - phonon modes for, 502
 - for phonons in cubic lattice, 498–500
 - for quantum mechanical free electron, 583–584
- Dispersion relation
 - and bands, 620–622
 - with transverse and longitudinal modes, 495
- 2-D lattice with primitive vectors, 465
- 1-D monatomic crystal
 - extended band structure for LA phonons in, 517
 - periodic potential for electron in, 582
 - phonon modes in
 - with 1-D motion, 502–503
 - fixed-endpoint boundary conditions, 503–505
- 3-D monatomic crystal, phonon states in, 529
- 2-D monatomic crystal, P-DOS for, 512, 514–515
- Dopant atoms, 8–9
- Dopant ionization statistics
 - derivation of, 735–736
 - dopant Fermi function, 734–735
- Drift/diffusion currents, 18
- Drift mobility, 552
 - collisions and, 553–555
 - current density in terms of, 553
- Drude model, 551
- Dual space, 152
- Dulong–Petit model, 528
- Dyadic notation, 192
- Dyhdral angle, 5
- Dynamical system, operators and groups, 99–102
 - basis vectors, transformations of, 100–101
 - isomorphism, 101
 - linear operators, 100
 - matrix representation, 103–104
 - permutation group, 103–104

- E**
- ECR, *see* Electron cyclotron resonance
- EDOS, *see* Energy density of states
- Effective mass, 596
- in band theory, 634–637
 - 3-D, 606
 - degenerate bands, 647–649
 - dependence on wave vector, 606
 - for electron and hole, 597
 - equation
 - diagonal matrix elements of V_E , 629–630
 - envelope approximation, 628–629
 - single-band, 625–628
 - thesis, 623–625
 - for three-dimensional band structure, 604–608
- Effective reflectance, 577
- Ehrenfest's theorem, 334–336
- Eigenequation for periodic Bloch states, 641–642
- Eigenvector
 - basis sets, 253
 - equation, 249, 265
 - of Hamiltonian, 486
 - vector collapses, 261
- Einstein convention, 228
- Einstein–Podolsky–Rosen paradox, 259
- Einstein repeated summation convention, 86, 190, 299, 324
- Einstein's special relativity, 201
- E - k band diagrams, 609–610
- E - k dispersion relation, 597
- Electrical conduction, 16
- Electrical contacts and quantum tunneling, 580–581
- Electrical injection, 805
- Electric dipole, 3
- Electric field, 19, 204, 552–553, 732–733
 - applied to crystal, 600
 - for diodes, 730
 - and electron distribution, 601
- Electromagnetically induced transition, schematic
 - illustration of, 374
- Electromagnetic fields, 23, 201, 224, 238, 251, 265
- Electromagnetic systems, 226
- Electromagnetic theory, 246
- Electromagnetic waves, 288, 365, 379
 - frequency of, 373
 - probability amplitude, 367
 - transition, 380
 - selection rules, 368
 - upward/downward transitions, 366
- Electron band diagram, 478–479
- Electron beam, 562
- Electron current, 598
- Electron cyclotron resonance, 747
- Electron cyclotron resonant etcher, 755–757
 - block diagram, 749
 - reflectometer, 756
 - wafer surface, 750
- Electron energy, 25
- Electronic components, 432
- Electronic transitions, 9
- Electronic waveguide
 - block diagram for, 573
 - scattering-matrix equation for, 574
 - transfer-matrix equation for, 574
- Electron lithography, 23
- Electron lodges, 374
- Electron magnetic dipole, energy of, 319
- Electron parameterization, 286
- Electron-resonant device
 - resonance conditions, 575–578
 - transfer matrix, 574–575
- Electrons drop, 15
- Electrons flow, entropy, 724
- Electron traps, 375
- Electron wave, 255
- Electron wavelength, 2, 276
- Electrostatic forces, 9, 202
- Electrostatic potential, periodic, 466
- Endpoint boundary conditions, 654
- Energy bandgap, 1
- Energy bands, Kronig–Penney model, 614–616
- Energy density of states, 649
 - and boundary conditions, relation between, 653–654
 - calculation using periodic boundary conditions, 667–668
 - for computing summations, 668–669
 - for 1-D crystals, 659
 - for 2-D crystals, 659–660
 - for 3-D crystals, 658, 661–662
 - definition of, 650–653
 - density of k -states, 657–659
 - for infinitely deep well, 676–677
 - k -space and E -space, 662–663
 - probability and, 669–670
 - for quantum well
 - in 1-D crystal, 681–682
 - in 2-D crystal, 682–683
 - in 3-D crystal, 680–681
 - and subbands, 684–685
 - for quantum wire, 685–689
 - for reduced dimensional structures, 677
 - envelope function approximation, 678–680
 - tensor effective mass and, 663–665
- Energy eigenfunctions, 270
 - Bloch's form of, 590
 - "standing wave", 591
- Energy eigenvalues, 270
- Energy surfaces, 606
- Ensembles, realizations of, 781
- Entropy, derivative of, 736
- Envelope function, 585
 - for infinitely deep well, 591
- Envelope wave function in plane waves, 627–628
- Equation of continuity, 551
 - for charged quantum particle, 557–558
 - differential form of, 556–557
 - integral form of, 556
- Ergodic processes, 782
- E -space density of states, 662–663
- Euclidean basis vectors, 297–298
- Euclidean inner product, 234
- Euclidean vector spaces, 32–34, 50, 52, 54, 62, 65, 116, 143, 190, 236, 428
 - adjoint operator, 42–43
 - basis and completeness, 39–40
 - closure relation, 40–41
 - commutivity of, 43
 - components of, 46

- Dirac notation, 37–38
 - discrete/continuous basis, 63
 - dual space, 41
 - inner product, 35
 - and norm, 44–45
 - kets, bras, and brackets, 38–39
 - Kronecker delta function, 49, 74–75
 - Euler–Lagrange equations, 230
 - Eulers’ equations, 770
 - Extended states, localized and, 649–650
- F**
- Fabricate devices, 747
 - Fabry–Perot cavity, 248, 698
 - Face-centered cubic crystal, 4
 - FBZ, *see* First Brillouin zone
 - FCC cell, 470
 - FCC crystal, *see* Face-centered cubic crystal
 - FEBC, *see* Fixed-endpoint boundary conditions
 - Fermi–Dirac distributions, 19, 695, 718, 738, 805
 - Boltzmann distribution, 720
 - carrier distribution, 721
 - carriers, density of, 720–721
 - density of states, 720
 - derivation of
 - Bose–Einstein distributions, 725–729
 - Maxwell, 724–725
 - Pauli exclusion principle, 722–724
 - electrons, 704, 720–722
 - electrons occupying states, probability of, 718–720
 - holes, 721–722
 - Fermi energy, 704
 - Fermi function, 728, 730
 - electron occupying, 720
 - temperatures, 719
 - Fermi levels
 - bandgap, 18
 - electron, 19
 - materials, 734
 - Fermion amplitude operators, 412
 - Fermion particles, 722
 - interchanging effect, 400
 - Fermions, wave functions of, 398
 - Fermi’s golden rule, 2, 353, 365
 - computational tool, 373
 - energy density-of-states, 374–377
 - equations, 377–381
 - phonon/photon absorption, 373
 - phonon/photon emission, 373
 - phonon/photon scattering, 373
 - semiconductor gain, demonstration, 374
 - time-dependent perturbation theory, 373
 - FETs, *see* Field-effect transistors
 - Feynman computer, 435–436
 - Feynman path integral, 205, 245, 428, 430, 435
 - classical limit, 430–431
 - derivation of, 428–430
 - Schrödinger equation, 431–432
 - stress, 422
 - Feynman processor, 435, 437
 - FIB, *see* Focused ion beam
 - Field-effect transistors, 1
 - n-channel, 25
 - semiconductor devices, 24
 - Finitely deep well
 - energy levels for, 689
 - lowest energy level for, 279
 - representation of wave function, 671
 - First Brillouin zone, 585
 - boundaries, motion of atoms at, 494
 - edges of, 494, 586
 - for k -space, 586
 - for materials with zinc blende crystal structure, 603
 - total number of states in, 599
 - wave vectors within, 599
 - Fixed-endpoint boundary conditions, 654
 - density of k -states from, 667
 - and phonon modes, 503
 - Fluctuation-dissipation theorem, 696
 - Fock ket, 403
 - Fock states, 408, 413, 421
 - basis vector, 407
 - bosons, 406
 - fermions, 404
 - origin of
 - bosons, 406
 - fermion, 408
 - Hilbert space, 405
 - wave functions, 404
 - orthonormality condition, 403
 - Focused ion beam, 747
 - Forward biasing, I–V characteristics, 17
 - Fourier amplitude, 794
 - Fourier basis sets
 - Fourier cosine series, 67–68
 - Fourier series, 66, 69–71
 - Fourier sine series, 68–69
 - Fourier transform, 71–73
 - Fourier series, 66, 251, 791
 - basis vectors for, 531
 - expansion, 50, 274, 775
 - and general lattice translations, 475–476
 - importance of reciprocal lattice for, 474–475
 - types of, 74
 - Fourier transform coordinate space, 59
 - Fourier transforms, 66
 - 2-D Fourier transform, 85
 - nonperiodic functions, 775
 - Free electron dispersion curve, 584
 - Free electron model, 581
 - classical, 551
 - for 1-D case, 583
 - dispersion curve, 583–584, 586
 - nearly (*see* Nearly free electron model)
 - normalization, wave function, 584
 - Schrödinger wave equation, 582
 - Functions
 - basis set of, 50
 - coordinate basis set, 47–49
 - coordinate space representation, 45–47
 - inner product, 49
 - representations, 49–50
 - rotation, 188
 - vector space representation, 45–46

Function space

- continuous basis sets, 59–61
- discrete basis sets
 - closure relation, 53–54
 - coordinate space, 61–64
 - Dirac delta function, 64–65
 - Hilbert space, 50–53
 - norms and inner products, 54–55
 - weight functions, 55–58

GGaAs, *see* Gallium arsenide

GaAs–AlGaAs interface, wave reflectivity and transmissivity, 563–564

Gallium arsenide, 17

- AlGaAs, 730
- band diagram, 10, 604, 639
- laser devices, 751
- light-heavy-hole band, 12
- p-type, 17
- quantum well, 15
- strains, 12
- surface metal adhesion, 758
- valence band, 10–11
- optical circuits, fabrication steps
 - CAD phase design, 751–752
 - cleaning, 752–753
 - e-beam evaporator, 751
 - growing grass, 755
 - Ohmic p-type contact, 757–758
 - oxygen ion implant, 758
 - photolithography, 753–754
 - polyimide insulating layers, 759
 - thermal evaporator, 750
 - top-metal contact pad, 759–760
 - wafer cleaving, 752–753
 - wafer thickness, 760

Gauss' law, 733

Graham–Schmidt orthonormalization procedure, 65–66

Grand canonical ensemble, 704

Gravitational force, 440

Green function, 422, 424

Group velocity

- from dispersion relation, 597
- electron, free space, 793
- Fourier integral, 793–795
- illustration of, 790–793
- for monatomic chain, 495
- plane wave, 795
- wave packet, 789–790

H

Hamilton formulation, phase space coordinates, 204

Hamiltonian

- classical, 535–536
- closed system, 354
- for continuous system, 542–543
- eigenfunctions of, 270, 486
- eigenvectors and eigenvalues, 345, 592
- for Kane's model, 640–641
- matrix of, 643–646
- for one-dimensional wave motion, 542–543

phonon field quantization and

- conjugate variables, 536
- creation and annihilation operators, 537–538

quantum phonon, 536

rotational invariance of, 303

symmetries of, 485–486

Hamiltonian density, 227, 231, 542–543

Hamilton's canonical equations, 211–213

Hamilton's principle, 205, 207

Harmonic oscillator, 183, 245, 289, 813

atom, 289

classical and quantum, 285–288

energy eigenfunctions, 294–296

energy eigenvalues, 294

energy eigenvectors, 290

Hamiltonian, 285, 290

ladder operators, 288–290

Hamiltonian, 290–292

linear restoring force, 285

motion of, 286

quantum mechanical solutions, 287

raising and lowering operators, 289

raising–lowering operators, 290

raising properties, 292–293

square well, 289

Harmonic oscillator solutions, 287

Heavy-hole band, 11–12, 665

Heisenberg coordinate, 428

Heisenberg operators, 338

Heisenberg representation, 331, 337, 424

Heisenberg uncertainty, 176–179, 261, 266, 269

Heisenberg wave function, 337

Helium, 5

HEMT, *see* High electron mobility transistor

Hermite polynomials, 341

Hermitian conjugate, 42, 124

Hermitian interaction energy, 366

Hermitian operators, 32–33, 99, 114–115, 151, 161, 170, 246, 253, 255–256, 259, 262, 264, 269, 330

adjoint, self-adjoint, 151–155

bounded Hermitian operators, 172–176

eigenvectors and eigenvalues of, 162

basic theorems, 158–161

direct product space, 162

orthogonal eigenvectors, 160

orthonormal eigenvectors, 224

theorems, 170–172

unitary operators, 156–158

Hermitian property, 159

Heteropolymer-based computer, 439

Heterostructure materials, 564

components of, 567

modifications for, 567–568

HH band, *see* Heavy-hole band

High electron mobility transistor, 25–26

Hilbert space, 56, 102, 108–109, 144, 245, 248, 250, 254, 284, 311, 324, 384, 411, 538

basis vectors, 169

definition of, 34–36

Dirac delta function, 185

1-D translation and lie group, 186

Hermitian linear operator, 171

of linear operators, 130

- linear transformation, 110
 - mapping physical space, 312
 - N-dimensional, 105
 - notion of, 124
 - vs. physical space, 312–314
 - quantum bits, 433
 - rotation operator, 317
 - rotations, 150, 317–318
 - Taylor series expansion, 437
 - time-dependent components, 277
 - type of, 50
 - vector spaces, 31, 34
 - wave functions, 36, 331, 356
 - Hilbert space number, 405
 - Hilbert space vector, 318
 - physical spin vector, 314
 - symmetrical fashion, 314
 - Hilbert space wave function, 312
 - Hole current, 602
 - Hook's law, 217
 - Human-made quantum wells, 271
 - Hybrid orbital, 463–464
- I**
- Image reversal process, 753
 - Impulse function, 48
 - Indirect bandgaps
 - dispersion relation for, 604
 - semiconductor, 12
 - Indirect bands, 11–12, 589
 - Infinite discrete set, completeness for, 175
 - Infinitely deep well
 - basis functions, 256
 - in 3-D crystal, 673–676
 - density of states calculation, 676–677
 - EM wave in, 360
 - envelope function approximation for, 672–673
 - Ionization energy, 729
 - Ion-trap computer, 439
 - I–V characteristics, 730
- J**
- JJ, *see* Josephson junction
 - Josephson junction, 28
- K**
- Kinetic energy, 421
 - $\vec{k} \cdot \vec{p}$ band theory, 632
 - for degenerate bands (*see* Degenerate bands, $\vec{k} \cdot \vec{p}$ band theory for)
 - for nondegenerate bands, 637–638
 - for periodic Bloch function, 633–634
 - Kronecker-delta correlation, 785
 - Kronecker delta functions, 39, 59, 62, 74, 81, 85, 251, 345–346, 358
 - Kronecker matrix product, 136
 - Kronig–Penney model, 1, 271, 586
 - energy bands, 614–616
 - goal of, 611
 - Sturm–Liouville problem, solving, 612–614
 - k -space
 - density of states in, 512–513, 662–663
 - isotropy in, 515
- L**
- Lagrange density, 225–226, 228, 542
 - Lagrange formulation, 202–203
 - Lagrange multipliers method, 710, 716, 728, 736, 799
 - constraints, 801
 - gradients, use of, 802
 - Hamiltonian density, 409
 - slope approach, 799–801
 - Lagrange's equations, 204, 206–207, 210, 217, 229
 - conjugate momentum, definition, 210
 - 1-D wave motion, 227–229
 - equations of motion, 216–217
 - Hamiltonian density, 225–227
 - normal coordinates
 - discrete array, 216
 - normal modes, 222–224
 - transformation, 217–221
 - Schrödinger equation
 - Hamiltonian density, 231–232
 - Schrödinger wave equation, 230–231
 - variational principle, 207–209
 - Lagrangian
 - for continuous system, 542
 - density, 226
 - formulation, 230
 - for line of atoms
 - kinetic energy, 533–534
 - momentum, 535
 - reasons for developing, 532
 - Lagrangian-derived Hamilton, 409
 - LaHospital's rule, 771
 - Langevin force, 697
 - Langevin function, 697
 - Langevin noise, 696
 - LA phonons
 - dispersion curve for, 500
 - in 1-D monatomic crystal, extended band structure for, 517
 - group velocity for, 500
 - Laplace's equations, 305–306
 - Laser beam, 749
 - Lattice
 - definitions, 464–465
 - dispersion relation for phonons in, 492
 - translating, 465–466
 - types
 - BCC, 469–470
 - diamond and zinc blende structures, 471–472
 - diamond-like structures, 472
 - FCC, 470
 - Wigner–Seitz primitive cell, 470–471
 - Law of least action, 204
 - Law of mass action, 732
 - LED, *see* Light emitting diode
 - Legendre equation, 307–308
 - Legendre polynomials, 307
 - Levi-Cevita symbol, 118
 - Light emitting diode, 2
 - Light-hole (LH) valence bands, 11–12, 665

Light, optical absorption of, 805
 Linear isomorphism, 37
 Linear monatomic crystal, *see* Monatomic linear crystal
 Linear operators, 139
 Line defects, 484
 Liouville equation, 695, 698
 Liquid crystals, 3
 Liquids, atomic/molecular order, 3
 Localized and extended states, 649–650
 Longitudinal motion, 221, 495
 Longitudinal vibration
 springs, masses, 218
 Lorentz invariance, 409
 Lorentz transformation, 88, 233, 235–236,
 238–239
 equations, 238
 Minkowski space, 234
 LO waves, dispersion curve for, 501

M

Macroscopic classical systems, 202
 Massless rods, 202
 Matrix operations
 Hermitian conjugate, 123–124
 isomorphism, 117
 operator, determinant of, 118
 operator, inverse of, 120
 operators, composition of, 116–117
 operator, trace of, 122
 transpose operation, 123–124
 Matrix representation
 for averages, 115–116
 definition of, 105–106
 Dirac notation, 107–109
 function spaces, 113
 matrix equation, 110
 operating, arbitrary vector, 109
 operator, 106–107
 expectation values, 114–115
 map vectors, 104
 Maxwell-Boltzmann distribution, 724–725
 MBE, *see* Molecular beam epitaxy
 MEMs, *see* Microelectrical machines
 Metal evaporator system, 750
 Metal organic chemical vapor deposition, 747
 Microelectrical machines, 4
 Micro-optical-electric machines, 21
 Microstates, 711
 Miller indices
 for plane, 468
 rules for, 469
 Minkowski space, 31, 86, 233–234, 237
 coordinates and pseudo-inner product, 86
 derivatives, 87–88
 pseudo-orthogonal vector notation, 86
 tensor notation, 86–87
 Mobile holes, 18
 Mobility and conductivity, 552–553
 MOCVD, *see* Metal organic chemical vapor deposition
 MOEMs, *see* Micro-optical-electric machines
 Molecular beam epitaxy, 2, 273, 747–748
 Monatomic crystal
 acoustic polarizations for, 507–508

 1-D, (*see* 1-D monatomic crystal)
 nearly free electron model for, 584–585
 Monatomic linear crystal; *see also* Three-dimensional
 monatomic crystals
 acoustic branches for 3-D motion in, 508
 allowed states for, 502
 dispersion curve for
 with 1-D motion, 492
 FBZ and, 493
 dispersion relation $v(k)$ for, 491
 1-D motion, 507
 2-D motion, 507
 3-D motion, 507
 equations of motion
 harmonic motion, 491
 transverse wave motion, 492
 normal modes for
 collective motions of atoms, 487, 489
 coordinates for, 489
 examples, 490
 longitudinal vibration of masses, 488
 motion of single atoms, 487
 shape of, 490
 phonon group velocity for, 494–495
 phonon states in FBZ for, 529
 Monatomic phonon system, 527
 Multielement electronic device, 570
 Multiparticle system
 microstates, 709
 quantum mechanics, 397

N

Nanometer-scale devices
 resonant-tunnel device, 26
 resonant-tunneling transistor, 26–27
 Aharanov-Bohm effect device, 27–28
 Josephson junction, 28
 quantum cellular automation, 27
 quantum interference device, 28
 single-electron transistor, 27
 Nano-optoelectronic components, 22
 Nanophotonic components, 22
 N-contact metal, application of, 758–759
 Nearly free electron model
 for monatomic crystal, 584–585
 time-independent Schrödinger equation, 584
 Noncommuting Hermitian operators, 267
 Nondegenerate bands
 and effective mass, 634–635
 $\vec{k} \cdot \vec{p}$ band theory for, 637–638
 periodic portion of Bloch wave function, 636–637
 Nonergodic process, realizations of, 783
 Non-Hermitian operators, 253
 Nonstationary process, 783
 Nonzero wave vector, 13
 Normal modes
 collective motions of atoms, 487, 489
 coordinates for, 489
 examples, 490
 longitudinal vibration of masses, 488
 motion of single atoms, 487
 shape of, 490
 Norm, definition of, 44

NOT gate, 437
 N-P homojunction, 730
 N-type dopants
 density of, 729
 representation of, 729
 Nuclear magnetic resonance computer, 440

O

Ohmic contacts, 2, 730
 Operator algebra, commutators, 138–143
 linear operators, 139
 theorems, 141–143
 Operator expansion theorem, 141–142
 Operator, inverse of, 120
 Operator maps basis vectors, 108
 Operator space
 concepts of, 124–125
 inner product
 Hilbert space, 129–131
 proof of, 131–132
 linear operators, 124
 basis expansion, 126–127
 matrices, basis vectors, 132
 Operator, trace of, 122
 Optical phonons, oscillation frequency for, 501
 Optical transitions, 13
 absorption, 370–371
 band diagram, 13
 EM interaction potential, 365–367
 emission, 371–372
 probability amplitude, integral for, 367–369
 results, 372–373
 rotating wave approximation, 369–370
 Optics polarization, 249
 Optics theory, 246
 Optoelectronic devices, 2
 device trends, 21–22
 fabrication challenges, 23
 monolithic integration, 21
 small optical signals, 22–23
 Orbital
 electron energy and, 461
 hybrid (*see* Hybrid orbital)
 spherical harmonic corresponding to, 462
 Orthogonal operators, 143
 Orthonormal expansion, 71
 Orthonormality
 for Bloch wave functions, 595–596
 Hilbert space of envelope functions, 631
 Output waves, 567
 Overlapping bands
 density of states for, 665–666
 for reduced dimensional structures, 665
 Overlapping 3-D bands, 666

P

Pauli exclusion principle, 414–416, 722–723, 725
 Pauli operators, 313
 Pauli spin matrices, 315–317, 320
 Pauli spin operators, 315, 443
 Pauli spin vector, 319
 P-DOS, *see* Phonon density of states

PECVD, *see* Plasma-enhanced chemical vapor deposition
 Periodic Bloch states, eigenequation for, 641–642
 Periodic boundary conditions
 allowed modes satisfying, 655–656
 basis states for Fourier series with, 531–535
 and 3-D cubic systems, 657
 density of states calculation using, 667–668
 k -space for, 658
 longest wavelength satisfying, 505
 and normalization, 586
 P-DOS in k -plane allowed by, 513
 and phonon modes, 505–507
 Periodic functions, periodicity of, 475
 Periodic potential and bandwidth, 616–617
 Periodic structure with input and diffracted waves, 477
 Periodic table, 6
 Phase space coordinates, 530
 Phase velocity, 789–790
 Phonon, 13
 applications of, 526
 Bose–Einstein probability distribution for (*see* Bose–Einstein probability distribution)
 conduction and optical processes, 486
 and continuous media, 539
 Hamiltonian density, 542–543
 wave equation and speed, 540–541
 in diatomic linear crystal, 498–500
 LA phonons, 501
 emission, 13
 Fock state, 538–539
 group velocity for monatomic crystal, 494–495
 modes for
 amplitude and, 509–510
 for 2-D and 3-D crystals, 508–509
 for 2-D and 3-D waves, 507–508
 for monatomic linear crystal, 502–505
 periodic boundary conditions, 505–507
 momentum, 517
 and crystal momentum, relation between, 518–519
 in free space, 518
 properties, 486
 quantum field theory, 528
 states in acoustic branch, 511
 total energy, 527
 Phonon density of states
 calculation of, 515–516
 for 1-D crystal, 511–512
 for 2-D crystal, 512, 514–515
 for 3-D crystal, 516–517
 definition, 510, 512
 in FBZ, 511
 in k -plane, 512–513
 for ω -space, 517
 Phonon system
 energy levels in thermal equilibrium, 522
 in microstate, probability of finding, 522–523
 Phosphorus nucleus, 15
 Photodetector, beams interfere, 757
 Photolithography process, 754
 Photon
 electron absorption, 353
 polarization of, 259
 Photonic computer, 440
 Photonic crystals, 381

- PIN heterostructure, 15
 - PIN photodetector, 2
 - Plane group, 480
 - Plasma-enhanced chemical vapor deposition, 5, 747–749
 - pn junction, 16–17
 - current–voltage characteristics, 738
 - at equilibrium (*see* pn junction at equilibrium)
 - forward biasing, 738
 - junction technology, 17–18
 - pn junction at equilibrium
 - built-in voltage, 739–741
 - concepts of, 736–739
 - junction fields, 741–743
 - Point defects, 484
 - Point group, 480
 - Poisson brackets, 211, 213, 215, 245
 - basic properties, 214–215
 - definition of, 213–214
 - motion/conserved quantities, constants of, 215–216
 - Poisson noise limit, 22
 - Poisson's equations, 305, 742–743
 - Polycrystalline materials, 4
 - Polycrystalline silicon, 4
 - Polyimide insulating layers, 758–759
 - p-orbitals, 6
 - angular momentum states, 462
 - lobes of, 463
 - Position operator, rotation of, 189–190
 - Primitive reciprocal lattice vectors, 473
 - Primitive unit cell, 467–468
 - Primitive vector for direct lattice, 474
 - Principal quantum number, 5
 - Probability amplitude, 91, 250, 269
 - sinusoidal waves, 272
 - wavelength of, 276
 - Probability-current density, 280
 - Probability density function, 779–780
 - Probability theory, 251, 797
 - classical concepts, 90
 - combinations, 797–798
 - permutations, 797
 - Probability vs. frequency, plot of, 372
 - Propagator
 - alternate formulation, 424
 - conditional probability, 422
 - conservative system, 423–424
 - free-particle, 426–427
 - Green function, 422–423
 - path integral, 425–426
 - p-type dopant, 729
 - p-type semiconductor, 742
 - Pulley system, 209, 213
 - Pythagoras relation, 235
 - Pythagorean's theorem, 86
- Q**
- QCA, *see* Quantum cellular automation
 - QD, *see* Quantum dot
 - QED-based computer, block diagram of, 440
 - QED, *see* Quantum electrodynamics
 - Quadratic potential, 286
 - Quantum cellular automation inverter, 27
 - Quantum computing
 - block diagrams, 434–435
 - Feynman computer, 436–438
 - memory register with multiple spins, 435–436
 - original Turing machine, 432–434
 - physical realizations, 439–440
 - Quantum dot, 26
 - Quantum electrodynamics, 22, 366, 409, 428
 - Quantum electromagnetic fields, 411
 - Quantum gate, classical view of, 434
 - Quantum Hamiltonian, 265
 - Quantum interference device, 28
 - Quantum mechanical
 - angular momentum, 330
 - Heisenberg equation, 338–339
 - Heisenberg representation, 337
 - Newton's second law, 339
 - interaction representation, 340–341
 - momentum operator, 334
 - multiparticle systems
 - angular momentum, 398
 - fermion and bosons, 398–399
 - Fock states, 403–404
 - Hamiltonian, eigenvectors of, 401–403
 - Hilbert space, 397
 - permutation operator, 399–401
 - presentations, 330
 - Schrödinger representation
 - Ehrenfest's theorem, 335–336
 - operator, rate of change, 334–335
 - Quantum mechanical analysis, 11, 303, 386, 388, 390–391, 700
 - Quantum mechanical Hamiltonians, 265, 270, 288
 - Quantum mechanical model for electron spin, 309
 - Quantum mechanical probability, 384
 - Quantum mechanical system, 433
 - transformation of, 156
 - Quantum mechanical wave functions, 255, 314
 - Quantum mechanics, 250
 - angular momentum, origin of, 297–298
 - fundamental operators
 - commutation relations and Heisenberg uncertainty relations, 266–267
 - Heisenberg uncertainty relations, derivation of, 267–269
 - Hermitian operators, 263
 - momentum operator, 264
 - program, 269–271
 - Schrödinger wave equation, 264–266
 - linear algebra, relations
 - atomic systems, 245
 - averages, 252–253
 - basis states, superposition of, 249–250
 - collapse, interpretations of, 257–259
 - eigenstates, 247–248
 - Heisenberg uncertainty relation, 259–262
 - Hermitian operators, 246–247
 - Hilbert space, 246
 - observables, complete sets of, 262–263
 - probability interpretation, 250–252
 - wave function, collapse of, 255–257
 - wave function, motion of, 254–255
 - Quantum operator, 253
 - Quantum optics, 409

Quantum phonon Hamiltonian, 536
 Quantum teleportation, 443–445
 Bell's theorem, 442–443
 EPR paradox, 441–442
 local vs. nonlocal, 440–441
 Quantum teleportation setup, 444
 Quantum theory, 1, 245, 250, 264
 basic principles of, 425
 commutators, 139
 creation–annihilation operation, 179
 Hermitian operators for, 170
 lowering–raising operation, 179
 Schrödinger presentation of, 246
 vector length, 36–37
 Quantum tunneling
 and electrical contacts, 580–581
 through barrier, 579–580
 Quantum Turing machine, 433
 Quantum wave function, 255
 Quantum well, density of states
 in 1-D crystal, 681–682
 in 2-D crystal, 682–683
 in 3-D crystal, 680–681
 subbands, 684–685
 Quantum well lasers, 245, 273
 Quantum wells, 269, 671, 705
 Quasi-Fermi levels, 20, 722
 QUIT, *see* Quantum interference device

R

Raising-lowering operators, 179
 direct product space, 182–183
 matrix and basis-vector representations, 180–182
 Random vector variable, 92
 Reactive ion etchers, 747
 block diagram of, 749
 plasma-enhanced chemical vapor deposition, 748
 Reciprocal lattice vectors
 application to
 electron and phonon bands, 478–479
 X-ray diffraction, 476–477
 atomic spacing, 478
 for Fourier series, 474–475
 primitive, 473–474
 Recombination process, 556
 Rectangular matrices, 110
 Reflectance, 568–569, 578
 Reflection of unit vectors, 481
 Reflectivity, 565
 Reservoir
 comment, 698–699
 definition of, 695–696
 energy distribution, 697
 fluctuation-dissipation theorem, 697–698
 optical emitter, 698
 particle exchange, 704
 thermodynamics, role, 695
 Resonant-tunneling diode, 26
 electron transport, band-edge diagram, 27
 Resonant tunneling effect, 2
 Resonant-tunneling transistor, 26
 RIEs, *see* Reactive ion etchers
 Rivest-Shamir-Adleman codes, 2

Rotating electron, 309
 Rotation matrices, 483
 Rotation operator, 187–189
 Rotations, 481
 and angle, relation between, 483
 consistent with translational symmetry, 482
 between primitive vectors, 482
 RSA codes, *see* Rivest-Shamir-Adleman codes
 RTD, *see* Resonant-tunneling diode
 RTT, *see* Resonant-tunneling transistor

S

Scalar multiplication, 33
 Scattering matrix
 amplitudes for, 576
 amplitudes in terms of traveling waves, 562–563
 and electronic element, 562
 for electronic waveguide, 574
 for interface, 572
 for simple interface, 567
 Scattering theory
 phase information, 562
 propagating waves, 560–561
 wave reflection, 561
 Schottky diode, 2
 Schrödinger's quantum mechanics, 334–335
 Schrödinger wave equation, 231, 245, 273–274, 280, 286,
 320, 332, 341, 355, 422, 813
 boundary conditions, 288
 electron wave functions satisfying, 484–485
 energy basis set, 408
 evolution of, 285
 finitely deep square well, 279–284
 finitely deep wells, 271
 free electron model, 582
 harmonic oscillator, 288
 for heterostructure, 630
 infinitely deep well, 271, 273–277
 partial differential equation, 254
 plot of, 283
 quantum theory, 431
 quantum wells, discussion of, 272–273
 for single electron in periodic potential, 589
 solution to, 485
 time-dependent, 274, 383
 time-independent, 241
 total wave function satisfying, 588
 Schrödinger wave function, 363
 Second quantization
 amplitude/field operators, interpretation of, 414–415
 annihilation operators, 410–412
 Boson creation, origin of, 418–422
 Fermion–Boson occupation, 415–416
 field commutators, 409–410
 Fock states, 412–414
 Hamiltonian/dynamical variables, 408
 operators, 416–418
 Semiclassical theory, 366
 Semiconductor crystal, 725
 Semiconductor devices, 9
 bipolar junction transistors, 24–25
 field-effect transistor, 24–26
 Semiconductor diode, 17

- Semiconductor GaAs lasers fabrication process, 751
 - Semiconductor materials, 5
 - Semi-insulating substrates, 758–759
 - SET, *see* Single-electron transistor
 - Signal-to-noise ratio, 21
 - Single-band effective mass equation
 - electrons in conduction band, 627
 - envelope wave function in plane waves, 627–628
 - Hamiltonians, 625–626
 - in terms of Bloch wave functions, 626
 - Single-electron transistor, 2, 26
 - Single harmonic oscillator
 - energy eigenfunction, 295
 - Hamiltonian, 291–292
 - ladder operators, 294
 - momentum/position operators, 295
 - Single-particle Hilbert spaces, 397
 - Sinusoid-like waves
 - coherence, multiple paths, 431
 - Si–Si bond, 10
 - Si–Si bonding electrons, 16
 - Slater determinant, 399, 403, 723
 - SL system, *see* Sturm–Liouville system
 - SM, *see* Scalar multiplication
 - SNR, *see* Signal-to-noise ratio
 - Solar cells, 730
 - Solid rod
 - longitudinal vibrations of, 496–497
 - Young's modulus of, 497–498
 - s-orbital, 6
 - angular momentum, 462
 - spherical harmonic corresponding to, 462
 - Space group, 480
 - Space-time coordinate system, 236
 - Space-time position, 441
 - Spatial-coordinate representation, 264
 - Specific heat
 - actual gasses, 527
 - Debye model for, 528–530
 - definition, 526
 - Einstein model for, 528
 - ideal gas, 527
 - Spectroscopic notation, 301
 - Spherical coordinates, 58
 - Spherical harmonics
 - eigenvectors, 305–309
 - list of, 308
 - sp^3 hybridization, 463
 - Spin angular momentum
 - magnitude of, 311
 - wave function, 313
 - Spin direction, 310
 - Spin Hamiltonian, 319–320
 - Spin-on process, 747
 - Spinors, 309–319
 - Spin vectors, 311, 319
 - Spring constant, 487
 - Springs
 - “amount of stretch from equilibrium” for, 488
 - longitudinal vibration of masses coupled by, 487
 - normal modes for transverse oscillations on, 489
 - Stacked electronic elements, 570
 - transfer-matrix equation for, 571
 - Standing waves and charge distribution, 588
 - Static electric field, 733
 - Stationary solutions, 276
 - Statistical mechanics, statistical ensembles
 - canonical ensemble, 700–704
 - entropy and states, 699–700
 - grand canonical ensemble, 704
 - microcanonical ensemble, 699–700
 - Step potential, 564
 - Sterling's approximation, 716, 736
 - Structure of space-time
 - Lorentz transformation, 236–238
 - Minkowski space, 233–236
 - space-time warping, 232–233
 - Sturm–Liouville problem, 99, 151, 158, 165, 275–276, 288, 306–307, 410
 - boundary conditions, 341
 - boundary conditions for, 560
 - solving, 612–614
 - Superposition of plane waves, 596–597
 - Surface defects, 484
 - SWE, *see* Schrödinger wave equation
 - Switching distinguishable particles, 714
 - Symmetry operations
 - on Bravais lattice, 479–480
 - and quantum mechanics, 484–486
- ## T
- Taylor approximation, 792
 - TBA, *see* Tight binding approximation
 - TCE, *see* Trichloroethane
 - Tensor effective mass
 - 3-D band diagrams and, 602–611
 - and density of states, 663–665
 - Tensor product space, 79; *see also* Direct product spaces
 - Tetrahedral bonding and diamond structure, 472
 - Thermal energy, 805
 - Thermal equilibrium, 521–522, 703, 705, 737, 805
 - Thermal heating, 805
 - Thermalization, 13
 - Thermal reservoir, 702
 - in thermal contact with matter, 519
 - thermal equilibrium, 521–522
 - Three-dimensional monatomic crystals, 496
 - Tight binding approximation
 - Bloch wave function, 619
 - single atom having single electron Hamiltonian, 617–619
 - wave function for electron in, 618–619
 - Time-dependent perturbation theory, 245, 341, 361–362
 - interaction representation, 362–365
 - optoelectronics, 352
 - physical concept, 353–355
 - Schrödinger picture, 355–359
 - Time-dependent Schrödinger equation, 270
 - Time-independent perturbation theory, 245, 341–352, 360–361
 - meaning of, 341–342
 - nondegenerate perturbation theory, 342
 - unitary operator for, 349–352
 - Time-independent Schrödinger equation, 270, 280, 402, 485
 - Bloch wave functions, 584–585

plane wave solutions, 581
 solutions to, 589, 591
 for step potential, 564
 Sturm–Liouville problem, 271
 Transfer matrix, 570
 amplitudes for, 576
 for electron-resonant device
 resonance conditions, 575–578
 wave reflections, 574–575
 for interface, 573
 vs. scattering matrices, 571–572
 for stacked electronic elements, 571
 Transfer-matrix equation
 for electronic waveguide, 574
 physical output variables in, 571–572
 Transistors, 23–24
 Translation eigenvalues
 as complex numbers, 593–594
 and primitive vectors, 593
 product of, 592–593
 traveling wave form of, 594
 Translation of function, 466
 Translation operators, 183
 definition, 466
 Dirac delta function, 185–186
 exponential form of, 183–184
 for lattice, 465
 position-coordinate ket, 185
 position operator, 184–185
 three dimensions, 186
 Translation vectors, 482
 Transmissivity, 566
 Transmittance, 568–569
 Transverse motion of atoms, 495
 Transverse wave motion, 227
 Trichloroethane, 752
 Triode vacuum tube, 24
 Tunneling, *see* Quantum tunneling
 Turing machine, 434
 Two-level atom, 126

U

UHV, *see* Ultrahigh vacuum
 Ultrahigh vacuum, 747
 Umklapp phonon process, 519
 Unitary operators
 basis vectors, mapping of, 146
 group, matrix representation of, 150–151
 orthogonal rotation matrices, 143–144
 similarity transformations, 148–149
 trace and determinant, 148
 unitary transformation, 146–147
 visualizing unitary transformation, 147
 Unitary transformation, 146
 Unit cells
 conventional, 468
 primitive, 467–468

V

Vacuum tubes, 23
 Valence band, 13, 603, 735, 805
 VB, *see* Valence band

Vector components and probability, 88
 applications of, 90
 contrast, random vectors, 92
 discrete and continuous Hilbert spaces,
 91–92
 starters, 2-D space for, 88–90
 Vector space, 104
 antilinear isomorphism, 37
 basis functions, 69
 conceptual diagram, 125
 definition of, 33
 Euclidean/function spaces, 127
 functions, composition of, 117
 linear isomorphism, 37
 linear operator, 107
 quantum theory, linear algebra,
 31–32

W

Wafer
 atoms, 748
 isotropic and anisotropic etches, 748
 mask pattern, photolithography transfers, 753
 photolithography, CAD designs, 753
 structure, 752
 Wave function, 258, 284, 335
 absorption, 370
 antisymmetry of, 408
 collapse of, 255–257
 components of, 382, 390
 delta-function type, 399
 due to potential barrier, 580
 emission, 371
 frequency, 263
 Hilbert space, 255
 incident optical waves, 423
 infinitely deep well, 383
 linear algebra, 436
 motion of, 254–255
 nonzero standard deviation, 262
 normalization of, 594–596
 probability density of, 259
 quantum field theory, 409
 quantum mechanical object, 335
 s-orbital, 6
 spherical harmonics, 305
 time, 336
 two basis sets, 261
 wavelength, 263
 Wave motion
 in 1-D crystals, 508
 in 2-D crystals, 508–509
 electrons, 793
 quantization of, 509–510
 Wave packet, 790
 Wave tunneling through barrier, 579
 Wave vectors
 angular frequencies, 792
 electron gun, 392
 in FBZ, 518–519
 function of, 11
 magnitude and frequency, 515
 and reciprocal lattice vectors, 493

Weight functions, 55–58
Wentzel-Kramers-Brillouin (WKB) approximation, 580
Weyl ordering, 428
Wigner–Seitz primitive cell, 470–471

X

X-ray diffraction and reciprocal lattice vectors, 476–477

Z

Zinc blende
 diamond and structures, 471–472
 FBZ, 603
Zone diagram, reduced and extended, 598